Functions and sequences in analysis and applications

Habilitation thesis

Faculty of Mathematics and Physics
of Charles University in Prague

Dr. rer. nat. habil. Jan Vybíral, Ph.D.
born in Hranice, Czech Republic

In Prague, October 23, 2015

Preface

This cumulative habilitation thesis presents the work done in 14 research articles and one survey chapter. The summary has two parts. The first one introduces the mathematical background of the subject and contains a historical survey of decomposition techniques in the frame of function spaces and an overview of the techniques of sparse recovery. After that, in the second part, the results of the above mentioned papers are discussed. Although I tried to comment also on the proofs of the results and put them into the historical perspective given before, I would like to point the reader to the original papers for full proofs and further references.

Acknowledgment

I would like to acknowledge the support from the DFG under the grant Hi 584/2-2, from FWF project Y 432-N15 START-Preis "Sparse Approximation and Optimization in High Dimensions", from the DFG Research center Matheon in Berlin, to ERC CZ grant LL1203 of the Czech Ministry of Education and to the private Neuron Fund for Support of Science, who all supported me during my research.

The work presented here was done mainly at the Friedrich-Schiller University Jena, at RICAM in Linz, at Technical University Berlin and at to a smaller extent also at Charles University in Prague. I would like to thank the colleagues in these places for their hospitality and friendly environment.

I would like to thank my collaborators, namely H. Boche, R. Calderbank, C. Draxl, M. Fornasier, L. M. Ghiringhelli, J. Haškovec, A. Hinrichs, H. Kempka, A. Kolleck, G. Kutyniok, S. V. Levchenko, S. Mayer, M. Scheffler, K. Schnass, C. Schneider, and T. Ullrich for many inspiring discussions and for the work they contributed to the articles collected in this thesis.

Prag, October 2015 Jan Vybíral

Contents

Included publications

This habilitation thesis is based on the results obtained in a joint work with a number of coauthors in the following publications.

[P1] J. Vybíral, A new proof of Jawerth-Franke embedding, Rev. Mat. Complut. 21 (2008), 75–82.

[P2] J. Vybíral, Widths of embeddings in function spaces, J. Compl. 24 (2008), 545–570.

[P3] J. Vybíral, Sobolev and Jawerth embeddings for spaces with variable smoothness and integrability, Ann. Acad. Sci. Fenn. Math. 34:2 (2009), 529–544.

[P4] C. Schneider and J. Vybíral, Non-smooth atomic decompositions, traces on Lipschitz domains, and pointwise multipliers in function spaces, J. Funct. Anal. 264 (5) (2013),1197–1237

[P5] H. Kempka and J. Vybíral, Spaces of variable smoothness and integrability: Characterizations by local means and ball means of differences, J. Fourier Anal. Appl. 18 (4) (2012), 852–891.

[P6] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, A Survey of Compressed Sensing, First chapter in Compressed Sensing and its Applications, Birkäuser, Springer, 2015

[P7] A. Hinrichs and J. Vybíral, Johnson-Lindenstrauss lemma for circulant matrices. Random Struct. Algor. 39(3) (2011), 391–398

[P8] J. Vybíral, A variant of the Johnson-Lindenstrauss lemma for circulant matrices, J. Funct. Anal. 260(4) (2011), 1096–1105

[P9] J. Vybíral, Average best m-term approximation, Constr. Approx. 36 (1) (2012), 83–115

[P10] M. Fornasier, J. Haškovec, and J. Vybíral, Particle systems and kinetic equations modeling interacting agents in high dimension, SIAM: Multiscale Modeling and Simulation, 9(4)(2011), 1727–1764

[P11] M. Fornasier, K. Schnass, and J. Vybíral, Learning functions of few arbitrary linear parameters in high dimensions, Found. Comput. Math. 12 (2) (2012), 229–262

[P12] A. Kolleck and J. Vybíral, On some aspects of approximation of ridge functions, J. Appr. Theory 194 (2015), 35–61

[P13] S. Mayer, T. Ullrich, and J. Vybíral, Entropy and sampling numbers of classes of ridge functions, Constr. Appr. 42 (2) (2015), 231–264

[P14] A. Kolleck and J. Vybíral, Non-asymptotic analysis of ℓ_1-Support Vector Machines, submitted

[P15] L. M. Ghiringhelli, J. Vybíral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big data of materials science - Critical role of the descriptor, Phys. Rev. Lett. 114, 105503 (2015)

Part I

Introduction

The main subject of this habilitation thesis is to follow the historical path from decomposition techniques in function spaces to sparse decompositions and sparse recovery, which finally resulted into the novel area of compressed sensing. We start with a brief historical overview of function spaces and their decomposition properties, which we use also to introduce some basic notation. As we are not able to cover all the topics of the theory of function spaces in this short survey, we refer to [2, 3, 65, 75, 76, 102, 88, 107] for much more details and further references. Our selection of the topics is mainly governed by our interest in decomposition techniques. In the second part, we sketch the basic aspects of the area of compressed sensing. The material in these two parts is by no means new and is essentially taken over from [117] and [14].

Decomposition techniques

The very first traces of the study of function spaces may be found already in the second half of eighteen century. This period was devoted to the study of classical spaces of continuous and continuously differentiable functions. A new era of function spaces started with the pioneering work of Sobolev [99, 100, 101] (with some forerunners [53, 94]). The theory of distributions became an essential tool, which allowed to achieve new results (e.g. embedding theorems) applicable in the study of partial differential equations.

In later years, the area became an object of a vastly growing interest. More and more function spaces were defined with the help of explicit norms. In the parallel, the advantages of the techniques of Fourier analysis (like Littlewood-Paley theory) became evident. In this connection, the Hardy spaces $H_p(\Delta)$ (cf. Section 1.2) played a crucial role.

During the 60's and 70's of the last century, the well structured scales of Besov and Triebel-Lizorkin spaces, cf. Definition 1.1, emerged from the variety of function spaces available so far. They exhibit several advantages. Many classical spaces may be identified as Besov or Triebel-Lizorkin spaces for a special choice of parameters. Furthermore, their definition is given in terms of distributions and Fourier analysis and these spaces have “good” properties from the Fourier-analytic point of view, cf. [108, Section 2.2.3]. Also the spaces with fractional (or even negative) smoothness could be incorporated easily into these two scales. On the other hand, the definition of Besov and Triebel-Lizorkin spaces involves a certain smooth dyadic decomposition of unity, which makes it look much more complicated than that of Sobolev spaces.

Further essential breakthrough was achieved in the work of Frazier and Jawerth [51] and [52] (with an important forerunner being [28]). It was discovered that spaces of functions and distributions may be characterized in terms of their decomposition properties. They considered the decomposition formula $f = \sum_Q \langle f, \varphi_Q \rangle \psi_Q$ for all $f \in S'(\mathbb{R}^d)$, where Q runs over all dyadic cubes of $\mathbb{R}^d$ and φ_Q and ψ_Q are shifts of dilations of special functions φ and ψ.

A similar approach was then followed in all other decomposition techniques, which appeared afterwards. They all say, roughly speaking, that a function (or a distribution) f belongs to a certain function space (say $B^s_{p,q}(\mathbb{R}^d)$) if, and only if, it may be written in a form

$$f = \sum_{j,m} \lambda_{j,m} a_{j,m}, \tag{0.1}$$

where $\lambda_{j,m}$ are (real or complex) scalars and $a_{j,m}$ are certain special building blocks. Furthermore, the (quasi-)norm of f in the given function space is in some sense equivalent to the (quasi-)norm of the sequence $\lambda = (\lambda_{j,m})_{j,m}$ in an appropriate sequence space (i.e. $b^s_{p,q}$ in the case of Besov spaces).

Of course, the formula (0.1) gives arise to many questions, like the uniqueness of the decomposition or the linearity of the dependence of λ on f. For example, in the decomposition of Frazier and Jawerth the mapping $f \to \{\langle f, \varphi_Q\rangle\}_Q$ is linear, but it is not an isomorphism between the given function space and the corresponding sequence space.

But three properties of the building blocks $a_{j,m}$ appearing already in [51] and [52] are common to most of all the known decomposition techniques. Those are *smoothness*, *vanishing moment conditions* and *localization.*

• Quite naturally, the basic building blocks are supposed to exhibit at least the same degree of smoothness as the functions (or distributions) in the function space under consideration. Due to the very weak convergence of (0.1) (which is usually assumed to converge in $S'(\mathbb{R}^d)$), the smoothness of the building blocks is not limited from above. As the classical Haar wavelets are not even continuous, the question of minimal smoothness required in (0.1) has also been studied, cf. [110].

• The necessity of the moment conditions becomes clear when dealing with singular distributions. Therefore, the number of moment conditions needed grows with s (the smoothness of the space) decreasing, cf. Theorem 1.8. Let us point out that one possible way how to achieve (even an infinite number of) vanishing moments is to work with a function, whose Fourier transform has its support bounded away from zero.

• Finally, the localization of the building blocks is also necessary. One may observe that for $p > 1$ overlapping building blocks would allow to consider decompositions of f with arbitrarily small norm of the sequence of coefficients $\lambda = (\lambda_{j,m})_{j,m}$. This corresponds to no localization conditions needed in the decomposition theorem of $H_p(\mathbb{R}^d)$, $0 < p \leq 1$ of Coifman [28], cf. Theorem 1.4.

During last two decades, various different decomposition techniques appeared. They are usually named after the building blocks used, so that we speak about *atomic*, *molecular*, *quarkonial* or *wavelet decomposition.* Furthermore, these decompositions were adapted to a number of different function spaces (anisotropic spaces, spaces with dominating mixed smoothness, spaces of Morrey and Campanato type, ...). Last, but not least, the methods were adapted to spaces on domains.

We want to point out, how the theory of decomposition techniques is helping to deal with problems in the theory of function spaces. It turns out (and it has been like that since the work of Frazier and Jawerth) that many classical problems may be much more easily formulated and handled in the language of sequence spaces. We shall deal here mainly with Sobolev and trace embeddings of function spaces and their properties.

Sparse recovery

The huge interest in these techniques was driven by the large number of applications based on or making a use of them, i.e. signal processing in many disciplines (like medicine or geology), algorithm design, data compression or numerical analysis to name at least a few of them. Actually, the theory of decompositions developed into a subject on its own under the term of “frame theory”. The corresponding tools became more and more important with another driving force of applied science - the growing dimensionality of the problems we deal with nowadays. The necessity of processing larger and larger data sets (which can be often interpreted as larger and larger decompositions of continuous objects) lead to the development of special techniques. The most important tools in this area make a heavy use of the following observations: Although the dimensionality of the underlying problem grows rapidly with our ability to measure more and more data, its intrinsic dimension stays low. The highdimensional data sets are therefore well structured - and the most simple structural assumption on a vector in $\mathbb{R}^n$ is that most of its coordinates are zero, or at least very small. This observation is nowadays a basis for many algorithms in electric engineering, including the well known JPEG2000 format.

The real breakthrough in this field came with the advent of theory of *compressed sensing* of Donoho [41] and Candés, Romberg, and Tao [17, 19], cf. also [18]. In its most simple form, this theory proves that a sparse vector $x \in \mathbb{R}^n$ can be recovered effectively (i.e. in the polynomial time) from a small number m of carefully chosen linear and non-adaptive measurements $\langle a_i, x\rangle, i = 1, \ldots, m$, where m grows only linearly in the number of non-zero components of x and logarithmically in the dimension n. Furthermore, the recovery is stable with respect to noise and to small defects of sparsity, cf. [16, 20]. And last, but not least, the recovery is provided by the very well known LASSO algorithm of Tibshirani [105]. The methods used in this area combine powerful techniques of concentration of measure [67], geometry of Banach spaces [68], optimization theory and linear programming [54]. Following our survey chapter [14], we give more details on compressed sensing in Section 2.

The plan of this survey is as follows. In Section 1, we present a historically oriented overview of decomposition techniques in function spaces, Section 2 introduces the basic concepts of sparse recovery and compressed sensing. Finally, Section 3 discusses the results of the papers, which are part of this cumulative thesis. As mentioned already above, the material in the Sections 1 and 2 is essentially taken over from [117] and [14].

1 Decomposition techniques in function spaces

1.1 Definitions and basic notation

In this section we give the necessary notation and the definitions of the function spaces considered in this work.

We denote by $\mathbb{R}$ the set of all real numbers and by $\mathbb{R}^d$ the d-dimensional Euclidean space. Furthermore, $\mathbb{N}$ stands for the set of all natural numbers, $\mathbb{Z}$ for the set of all integers and $\mathbb{C}$ for the set of all complex numbers.

We denote by $S(\mathbb{R}^d)$ the Schwartz space of all complex-valued rapidly decreasing infinitely differentiable functions equipped with the usual topology and its dual by $S'(\mathbb{R}^d)$.

The Fourier transform of $\varphi \in S(\mathbb{R}^d)$ is given by

$$\mathcal{F}\varphi(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(x) e^{-i\xi\cdot x} dx, \quad \xi \in \mathbb{R}^d$$

with ins inverse denoted by

$$\mathcal{F}^{-1}\varphi(\xi) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(x) e^{i\xi\cdot x} dx, \quad \xi \in \mathbb{R}^d.$$

Both $\mathcal{F}$ and $\mathcal{F}^{-1}$ are extended to $S'(\mathbb{R}^d)$ by duality. We often write $\hat{\varphi}$ as a shortcut for $\mathcal{F}\varphi$ and $\varphi^\vee$ for $\mathcal{F}^{-1}\varphi$.

Although we are mainly interested in function spaces of Besov and Triebel-Lizorkin type (as defined in Section 1.1.2), we first collect the definitions of (some of) the classical function spaces.

1.1.1 Classical spaces

(i) The space of all complex-valued bounded and uniformly continuous functions is denoted by $C(\mathbb{R}^d)$ and is equipped with the norm $\|f|C(\mathbb{R}^d)\| = \sup_{x\in\mathbb{R}^d}|f(x)|$.

Let $m \in \mathbb{N}$. Then we denote by $C^m(\mathbb{R}^d)$ the space of all functions on $\mathbb{R}^d$, such that $D^\alpha f \in C(\mathbb{R}^d)$ for all multiindices α with $|\alpha| \le m$. The norm is then given by $\|f|C^m(\mathbb{R}^d)\| = \max_{|\alpha|\le m}\|D^\alpha f|C(\mathbb{R}^d)\|$.

(ii) The Lebesgue spaces $L_p(\mathbb{R}^d)$, $0 < p \le \infty$ are spaces of measurable functions, for which

$$\|f|L_p(\mathbb{R}^d)\| := \begin{cases} \left(\displaystyle\int_{\mathbb{R}^d}|f(x)|^p dx\right)^{1/p}, & \text{if } 0 < p < \infty \\ \text{ess sup}_{x\in\mathbb{R}^d}|f(x)|, & \text{if } p = \infty \end{cases}$$

is finite. Sometimes, we write only $\|f\|_p$ instead of $\|f|L_p(\mathbb{R}^d)\|$ for short.

(iii) Let $1 \le p \le \infty$ and $k \in \mathbb{N}_0$. Then the *Sobolev space* $W_p^k(\mathbb{R}^d)$ is defined by

$$W_p^k(\mathbb{R}^d) = \{f \in S'(\mathbb{R}^d) : D^\alpha f \in L_p(\mathbb{R}^d) \text{ if } |\alpha| \le k\}.$$

Here, the derivatives are interpreted in the distributional sense. One of the cornerstones of the theory of Sobolev spaces is the embedding property (usually called *Sobolev embedding*)

$$W_{p_0}^{k_0}(\mathbb{R}^d) \hookrightarrow W_{p_1}^{k_1}(\mathbb{R}^d) \tag{1.1}$$

if $0 \le k_1 \le k_0$ are non-negative integers, $1 \le p_0 \le p_1 < \infty$ and

$$k_0 - \frac{d}{p_0} = k_1 - \frac{d}{p_1}. \tag{1.2}$$

When considering the spaces on domains, then (under conditions which we shall discuss in detail later) (1.1) becomes even compact.

(iv) An essential effort was devoted to the extension of the theory of function spaces also to spaces with fractional (or even negative) smoothness. One of the reasons for that is hidden already in (1.2) - for given p_0, p_1 and k_0, the optimal k_1 may be a fractional real number. The classical way is represented by *Hölder spaces* $C^s(\mathbb{R}^d)$. Let $s > 0$ be not an integer. Then we define

$$C^s(\mathbb{R}^d) = \Biggl\{ f \in C^{[s]}(\mathbb{R}^d) : \tag{1.3}$$
$$\|f|C^s(\mathbb{R}^d)\| := \|f|C^{[s]}(\mathbb{R}^d)\| + \sum_{|\alpha|=[s]}\sup_{x\neq y}\frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x-y|^{\{s\}}} < \infty \Biggr\}.$$

Here, $s = [s] + \{s\}$ with $0 \le \{s\} < 1$ is a decomposition of s into its integer and fractional part.

The closely related *Zygmund spaces* $\mathcal{C}^s(\mathbb{R}^d)$ are obtained by replacing the first order by second order differences in (1.3). The definition of the (classical) *Besov spaces* reflects a similar idea. It works with the decomposition of the smoothness parameter $s = [s]^- + \{s\}^+$, where $0 < \{s\}^+ \leq 1$. Let $s > 0$ and $1 \leq p, q < \infty$. Then

$$\Lambda^s_{p,q}(\mathbb{R}^d) = \Bigg\{ f \in W^{[s]^-}(\mathbb{R}^d) : \|f|\Lambda^s_{p,q}(\mathbb{R}^d)\| := \|f|W^{[s]^-}(\mathbb{R}^d)\| \tag{1.4}$$

$$+ \sum_{|\alpha|=[s]^-} \left(\int_{\mathbb{R}^d} |h|^{-\{s\}^+ q} \|\Delta^2_h D^\alpha f\|^q_p \frac{dh}{|h|^d} \right)^{1/q} < \infty \Bigg\}, \tag{1.5}$$

where $\Delta^2_h g$ are the usual second order differences of g. If $q = \infty$, only notational changes are necessary. Let us refer to [108, Section 2.2] for other spaces (i.e. *Slobodeckij spaces* and *Bessel potential spaces*) with fractional smoothness.

1.1.2 Besov and Triebel-Lizorkin spaces

We give a Fourier-analytic definition of Besov and Triebel-Lizorkin spaces, which relies on the so-called *smooth dyadic resolution of unity.* Let $\varphi \in S(\mathbb{R}^d)$ with

$$\varphi(x) = 1 \quad \text{if} \quad |x| \leq 1 \quad \text{and} \quad \varphi(x) = 0 \quad \text{if} \quad |x| \geq \frac{3}{2}. \tag{1.6}$$

We put $\varphi_0 = \varphi$ and $\varphi_j(x) = \varphi(2^{-j}x) - \varphi(2^{-j+1}x)$ for $j \in \mathbb{N}$ and $x \in \mathbb{R}^d$. This leads to the identity

$$\sum_{j=0}^\infty \varphi_j(x) = 1, \qquad x \in \mathbb{R}^d.$$

Definition 1.1. (i) Let $s \in \mathbb{R}$ and $0 < p, q \leq \infty$. Then $B^s_{pq}(\mathbb{R}^d)$ is the collection of all $f \in S'(\mathbb{R}^d)$ such that

$$\|f|B^s_{pq}(\mathbb{R}^d)\| = \left(\sum_{j=0}^\infty 2^{jsq} \|(\varphi_j \widehat{f})^\vee | L_p(\mathbb{R}^d)\|^q \right)^{1/q} \tag{1.7}$$

is finite (with the usual modification for $q = \infty$).

(ii) Let $s \in \mathbb{R}, 0 < p < \infty$ and $0 < q \leq \infty$. Then $F^s_{pq}(\mathbb{R}^d)$ is the collection of all $f \in S'(\mathbb{R}^d)$ such that

$$\|f|F^s_{pq}(\mathbb{R}^d)\| = \left\| \left(\sum_{j=0}^\infty 2^{jsq} |(\varphi_j \widehat{f})^\vee(\cdot)|^q \right)^{1/q} | L_p(\mathbb{R}^d) \right\| \tag{1.8}$$

is finite (with the usual modification for $q = \infty$).

Remark 1.2. (i) The spaces $B^s_{pq}(\mathbb{R}^d)$ and $F^s_{pq}(\mathbb{R}^d)$ are independent on the choice of the function φ as soon as it satisfies (1.6). Unfortunately, if $p = \infty$ in the F-case (which was excluded in Definition 1.1), then this is no longer true and a different approach is necessary. We shall not go into details and refer to the recent monograph [120].

(ii) Let $s \in \mathbb{R}$, $0 < p < \infty$ and $0 < q \leq \infty$. Then the embedding

$$B^s_{p,\min(p,q)}(\mathbb{R}^d) \hookrightarrow F^s_{p,q}(\mathbb{R}^d) \hookrightarrow B^s_{p,\max(p,q)}(\mathbb{R}^d).$$

is an easy consequence of the Definition 1.1.

(iii) Let $-\infty < s_1 < s_0 < \infty, \quad 0 < p_0 < p_1 < \infty, \quad 0 < q_0 \leq q_1 \leq \infty$ with

$$s_0 - \frac{d}{p_0} = s_1 - \frac{d}{p_1}.$$

Then the classical Sobolev embedding (1.1) has its counterpart also for Besov and Triebel-Lizorkin spaces

$$B^{s_0}_{p_0,q_0}(\mathbb{R}^d) \hookrightarrow B^{s_1}_{p_1,q_1}(\mathbb{R}^d) \quad \text{and} \quad F^{s_0}_{p_0,\infty}(\mathbb{R}^d) \hookrightarrow F^{s_1}_{p_1,q_0}(\mathbb{R}^d). \tag{1.9}$$

Furthermore, the Jawerth-Franke embedding [50, 59] states that

$$F^{s_0}_{p_0,\infty}(\mathbb{R}^d) \hookrightarrow B^{s_1}_{p_1,p_0}(\mathbb{R}^d) \quad \text{and} \quad B^{s_0}_{p_0,p_1}(\mathbb{R}^d) \hookrightarrow F^{s_1}_{p_1,q_0}(\mathbb{R}^d). \tag{1.10}$$

(iv) The books [108, 88, 13] describe the stage of the theory of function spaces of Besov and Triebel-Lizorkin type as it stood in the late 1970's. For the more modern aspects of this theory we refer to the books of Triebel [109, 112, 113] and to [120].

(v) We use this place to introduce the symbols

$$\sigma_p = \max(1/p - 1, 0), \quad \sigma_{pq} = \max(1/p - 1, 1/q - 1, 0)$$

and

$$\sigma_p^d = d\max(1/p - 1, 0), \quad \sigma_{pq}^d = d\max(1/p - 1, 1/q - 1, 0).$$

These quantities play an important role in the theory of this spaces and shall be used frequently later on.

(vi) Definition 1.1 covers many of the classical spaces defined by derivatives and/or differences (cf. Section 1.1.1 for some examples). Especially,

$$\begin{aligned}
&B^s_{\infty,\infty}(\mathbb{R}^d) = \mathcal{C}^s(\mathbb{R}^d) \quad \text{if} \quad s > 0,\\
&B^s_{\infty,\infty}(\mathbb{R}^d) = C^s(\mathbb{R}^d) \quad \text{if} \quad s > 0, \quad s \notin \mathbb{N},\\
&B^s_{p,q}(\mathbb{R}^d) = \Lambda^s_{p,q}(\mathbb{R}^d) \quad \text{if} \quad s > 0, \quad 1 \leq p < \infty, \quad 1 \leq q \leq \infty,\\
&F^s_{p,2}(\mathbb{R}^d) = W^s_{p,2}(\mathbb{R}^d) \quad \text{if} \quad s > 0, \quad s \in \mathbb{N}, \quad 1 < p < \infty.
\end{aligned}$$

(vii) Definition 1.1 of isotropic Besov and Triebel-Lizorkin spaces has numerous modifications and extensions, which lead to specific function spaces, for example anisotropic spaces, spaces of generalized smoothness or spaces of variable smoothness and/or integrability.

1.2 Hardy spaces

The history of atomic decompositions is closely related to Hardy spaces H_p. In its original form, the Hardy space $H_p(\Delta)$ is a space of holomorphic functions on the unit disc $\Delta := \{z \in \mathbb{C} : |z| < 1\}$ satisfying

$$\|f|H_p(\Delta)\| := \sup_{0<r<1} \left(\frac{1}{2\pi} \int_0^{2\pi} |f(re^{it})|^p dt \right)^{1/p} < \infty.$$

This definition (which goes back to F. Riesz) was extended to functions of real variables by C. Fefferman and E. M. Stein in [46]. The space $H_p(\mathbb{R}^d), 0 < p \leq \infty$ is a space of $f \in S'(\mathbb{R}^d)$, such that

$$(M_\Phi f)(x) := \sup_{t>0} |(f * \Phi_t)(x)|, \quad x \in \mathbb{R}^d$$

is in $L_p(\mathbb{R}^d)$. Here $\Phi \in S(\mathbb{R}^d)$ with $\int_{\mathbb{R}^d} \Phi(x)dx = 1$ is arbitrary and $\Phi_t(x) = t^{-d}\Phi(x/t)$. Furthermore,

$$\|f|H_p(\mathbb{R}^d)\| := \|M_\Phi f|L_p(\mathbb{R}^d)\|$$

is a quasinorm on $H_p(\mathbb{R}^d)$. Different choices of Φ lead to equivalent quasinorms. If $1 < p < \infty$, then $H_p(\mathbb{R}^d)$ coincides with $L_p(\mathbb{R}^d)$. But for $0 < p \le 1$, one obtains new function spaces of distributions on $\mathbb{R}^d$.

The first atomic decomposition of $H_p(\mathbb{R}^d)$ with $d = 1$ and $0 < p \le 1$ was given in [28] and generalized to $d > 1$ in [66]. It uses the notion of p-atoms on the real line.

Definition 1.3. Let $0 < p \le 1$. A p-atom is a real-valued function b on $\mathbb{R}$ such that $\int_{-\infty}^{\infty} b(x)x^k dx = 0$, $0 \le k \le [1/p] - 1$, $k \in \mathbb{N}_0$, and the support of which is contained in an interval I for which $\sup_{x\in\mathbb{R}} |b(x)| \le |I|^{-1/p}$.

The quantity $[1/p]$ is the integer part of $1/p$. The corresponding decomposition theorem then takes the following form.

Theorem 1.4. ([28]) *A distribution f lies in $H^p(\mathbb{R})$, $0 < p \le 1$ if, and only if, it can be written in the form*

$$f = \sum_{i=0}^{\infty} \alpha_i b_i,$$

where α_i are in $\mathbb{R}$, b_i are p-atoms for $i \in \mathbb{N}$ and

$$A\|f|H^p(\mathbb{R})\|^p \le \sum_{i=0}^{\infty} |\alpha_i|^p \le B\|f|H^p(\mathbb{R})\|^p.$$

Here the constants $A, B > 0$ depend only on p.

1.3 Besov and Triebel-Lizorkin spaces

M. Frazier and B. Jawerth extended in [51, 52] the method of Coifman to a huge variety of other function spaces. They studied the decomposition formula $f = \sum_Q \langle f, \varphi_Q\rangle \psi_Q$ for $f \in S'(\mathbb{R}^d)$. Here, Q runs over all dyadic cubes of $\mathbb{R}^d$ and φ_Q and ψ_Q arise through shifting and dilating of special functions φ and ψ. These functions are smooth, rapidly decreasing and possess compactly supported Fourier transform. The mapping

$$S_\varphi : f \to (\langle f, \varphi_Q\rangle)_Q$$

is called φ-transform. Theorem 2.2 of [52] then states that S_φ maps the homogenous Triebel-Lizorkin space $\dot{F}^s_{p,q}(\mathbb{R}^d)$ into a special sequence space $\dot{f}^s_{p,q}$, which is defined through the (quasi)norm

$$\|\lambda|\dot{f}^s_{p,q}\| := \left\| \left(\sum_Q (|Q|^{-s/n-1/2}|\lambda_Q|)^q \chi_Q(\cdot) \right)^{1/q} \right\|_p,$$

where the sum runs again over all dyadic cubes of $\mathbb{R}^d$, $|Q|$ stands for the Lebesgue measure of Q and χ_Q is the characteristic function of Q.

Furthermore, the inverse φ-transform defined as

$$T_\psi : \lambda = (\lambda_Q)_Q \to \sum_Q \lambda_Q \psi_Q$$

maps $\dot{f}^s_{p,q}$ onto $\dot{F}^s_{p,q}(\mathbb{R}^d)$ and $T_\psi \circ S_\varphi$ is the identity on $\dot{F}^s_{p,q}(\mathbb{R}^d)$.

Remark 1.5.

- Frazier and Jawerth worked mainly with the homogenous function spaces and stated only in Section 12 of [52] the necessary modifications needed to deal with inhomogeneous spaces.
- Unfortunately, the φ-transform S_φ is no isomorphism between $\dot{F}^s_{p,q}(\mathbb{R}^d)$ and $\dot{f}^s_{p,q}$, i.e. S_φ does not map $\dot{F}^s_{p,q}(\mathbb{R}^d)$ *onto* $\dot{f}^s_{p,q}$. This was essentially improved using the theory of wavelets.
- The theory of [52] applies exactly to those function spaces which admit some sort of Littlewood-Paley characterization. This is in a very good agreement with the the observation of Triebel (see [108, Section 2.2.3]), who divided the function spaces into *good* and *bad* spaces according to their Fourier-analytic properties. Let us mention on this place that some prominent function spaces (like $L_1(\mathbb{R}^d)$, $L_\infty(\mathbb{R}^d)$ or $C(\mathbb{R}^d)$) are considered as *bad* function spaces from this point of view.
- The condition on vanishing moments of Coifman is incorporated in [52] through the assumption, that the support of the Fourier transform of φ and ψ stays away from zero. The new condition of [52] is that the building blocks ψ_Q are essentially localized on the dyadic cube Q (i.e. rapidly decreasing outside Q). This is reflected in all other decomposition techniques which involve both the vanishing moments condition and some kind of localization of the building blocks.

The central role in the theory of decomposition of function spaces is played by the atomic decomposition. We give the version as presented by Triebel in Section 1.5 of [112]. First, we define the corresponding building blocks. Let us observe that in contrast with Definition 1.3, the localization of the atoms is required.

Definition 1.6. (i) Let $\nu \in \mathbb{N}_0$ and $m \in \mathbb{Z}^d$. Then we denote by $Q_{\nu m}$ the closed cube in $\mathbb{R}^d$ with sides parallel to the coordinate axes, centered at $2^{-\nu}m$, and with side-length $2^{-\nu+1}$. Furthermore, $c\,Q_{\nu m}$ stands for the cube in $\mathbb{R}^d$ concentric with $Q_{\nu m}$ and with side length $c\,2^{-\nu+1}$.

(ii) Let $K \in \mathbb{N}_0$ and $c \geq 1$. A continuous function $a : \mathbb{R}^d \to \mathbb{C}$ for which there exist all derivatives $D^\alpha a$ if $|\alpha| \leq K$ is called a 1_K-atom if

$$\operatorname{supp} a \subset c\,Q_{0,m} \text{ for some } m \in \mathbb{Z}^d$$

and

$$|D^\alpha a(x)| \leq 1 \text{ for } |\alpha| \leq K. \tag{1.11}$$

(iii) Let $K \in \mathbb{N}_0, L \geq 0$, and $c \geq 1$. A continuous function $a : \mathbb{R}^d \to \mathbb{C}$ for which there exist all derivatives $D^\alpha a$ if $|\alpha| \leq K$ is called an (K, L)-atom if

$$\operatorname{supp} a \subset c\,Q_{\nu m} \text{ for some } \nu \in \mathbb{N}, m \in \mathbb{Z}^d,$$

$$|D^\alpha(x)a| \leq 2^{|\alpha|\nu} \text{ for } |\alpha| \leq K, \tag{1.12}$$

and

$$\int_{\mathbb{R}^d} x^\beta a(x) dx = 0 \text{ for } |\beta| < L.$$

Also the sequence spaces used in the frame of Besov and Triebel-Lizorkin spaces are somewhat more complicated compared to Theorem 1.4. We present a version, which reflects all the three parameters of the corresponding function spaces.

Definition 1.7. If $0 < p, q \leq \infty$, $s \in \mathbb{R}$ and

$$\lambda = \{\lambda_{\nu m} \in \mathbb{C} : \nu \in \mathbb{N}_0, m \in \mathbb{Z}^d\} \tag{1.13}$$

then we define

$$b^s_{pq} = \left\{\lambda : \|\lambda|b^s_{pq}\| = \left(\sum_{\nu=0}^{\infty} 2^{\nu(s-\frac{d}{p})q}\Big(\sum_{m\in\mathbb{Z}^d} |\lambda_{\nu m}|^p\Big)^{q/p}\right)^{1/q} < \infty\right\} \tag{1.14}$$

and

$$f^s_{pq} = \left\{\lambda : \|\lambda|f^s_{pq}\| = \left\|\left(\sum_{\nu=0}^{\infty}\sum_{m\in\mathbb{Z}^d} |2^{\nu s}\lambda_{\nu m}\chi_{\nu m}(\cdot)|^q\right)^{1/q} |L_p(\mathbb{R}^d)\right\| < \infty\right\} \tag{1.15}$$

with the usual modification for p and/or q equal to ∞. Here $\chi_{\nu m}$ stands for the characteristic function of $Q_{\nu m}$.

The atomic decomposition of Besov and Triebel-Lizorkin spaces is then given very much in the spirit of Theorem 1.4 and it goes back in a similar form to [51] and [52].

Theorem 1.8. ([112], Theorem 1.19) *(i) Let $0 < p \leq \infty$, $0 < q \leq \infty$, $s \in \mathbb{R}$. Let $K \in \mathbb{N}_0, L \geq 0$ with*

$$K > s \text{ and } L > \sigma^d_p - s$$

be fixed. Then $f \in S'(\mathbb{R}^d)$ belongs to $B^s_{p,q}(\mathbb{R}^d)$ if, and only if, it can be represented as

$$f = \sum_{\nu=0}^{\infty}\sum_{m\in\mathbb{Z}^d} \lambda_{\nu m} a_{\nu m}, \text{ unconditional convergence being in } S'(\mathbb{R}^d), \tag{1.16}$$

where for fixed $c \geq 1$, $a_{\nu m}$ are 1_K-atoms ($\nu = 0$) or (K, L)-atoms ($\nu \in \mathbb{N}$) and $\lambda \in b^s_{pq}$. Furthermore,

$$\|f|B^s_{p,q}(\mathbb{R}^d)\| \approx \inf \|\lambda|b^s_{pq}\|$$

are equivalent quasi-norms where the infimum is taken over all admissible representations (1.16).
(ii) Let $0 < p < \infty$, $0 < q \leq \infty$, $s \in \mathbb{R}$. Let $K \in \mathbb{N}_0, L \geq 0$ with

$$K > s \text{ and } L > \sigma^d_{pq} - s$$

be fixed. Then $f \in S'(\mathbb{R}^d)$ belongs to $F^s_{p,q}(\mathbb{R}^d)$ if, and only if, it can be represented by (1.16), *where for fixed $c \geq 1$, $a_{\nu m}$ are 1_K-atoms ($\nu = 0$) or (K, L)-atoms ($\nu \in \mathbb{N}$) and $\lambda \in f^s_{p,q}$. Furthermore,*

$$\|f|F^s_{p,q}(\mathbb{R}^d)\| \approx \inf \|\lambda|f^s_{pq}\|$$

are equivalent quasi-norms where the infimum is taken over all admissible representations (1.16).

Nowadays, a large variety of decomposition techniques is available in the literature. We shall present (a variant of) one of the most important one - the wavelet decomposition theorem. It removes some of the obstacles of Theorem 1.8. The first is the implicit definition of atoms - atoms are building blocks satisfying certain properties but may vary from one function to the other. The other sometimes inconvenient feature of Theorem 1.8 is the dependence of the coefficients λ in the optimal decomposition (1.16) on the distribution f. Due to some applications it would be desirable that this dependence is linear. Unfortunately, this does not follow from the theory of atomic decompositions.

We do not aim to give an overview of the vast area of wavelets. We recall only the minimum needed later on and point to [36, 77, 118] as standard references. The following theorem of Daubechies ensures the existence of compactly supported wavelets.

Theorem 1.9. **([35, 36])** *For any $k \in \mathbb{N}$ there are real-valued compactly supported functions*

$$\psi_0, \psi_1 \in C^k(\mathbb{R})$$

satisfying

$$\int_{\mathbb{R}} t^\alpha \psi_1(t)dt = 0, \quad \alpha = 0, 1, \dots, k-1,$$

such that

$$\{2^{\nu/2}\psi_{\nu m} : \nu \in \mathbb{N}_0, m \in \mathbb{Z}\}$$

with

$$\psi_{\nu m}(t) = \begin{cases} \psi_0(t-m) & \text{if } \nu = 0, m \in \mathbb{Z}, \\ 2^{-\frac{1}{2}}\psi_1(2^{\nu-1}t - m) & \text{if } \nu \in \mathbb{N}, m \in \mathbb{Z} \end{cases}$$

is an orthonormal basis in $L_2(\mathbb{R})$.

Wavelets on $\mathbb{R}^d$ may be obtained as tensor products of one-dimensional wavelets. With their help we obtain the following characterization of Besov and Triebel-Lizorkin spaces.

Theorem 1.10. **([111], Theorem 19)** *Let $0 < p, q \le \infty$, $s \in \mathbb{R}$ and $k \in \mathbb{N}$ with $k > \max(s, \sigma_p^d - s)$. Let ψ_0, ψ_1 be the Daubechies wavelets of smoothness k. Let $E = \{0,1\}^d \backslash (0, \dots, 0)$. For $e = (e_1, \dots, e_d) \in E$ let*

$$\Psi_e(x) = \prod_{j=1}^d \psi_{e_j}(x_j), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

(i) Then

$$\begin{cases} \Psi(x-m) = \prod_{j=1}^d \psi_0(x_j - m_j) & m = (m_1, \dots, m_d) \in \mathbb{Z}^d, \\ 2^{\frac{\nu-1}{2}d}\Psi_e(2^{\nu-1}x - m) & e \in E, \nu \in \mathbb{N}, m \in \mathbb{Z}^d \end{cases}$$

is an orthonormal basis in $L_2(\mathbb{R}^d)$.

(ii) Let $f \in S'(\mathbb{R}^d)$. Then $f \in B^s_{pq}(\mathbb{R}^d)$ if, and only if, it can be represented as

$$f = \sum_{m \in \mathbb{Z}^d} \lambda_m \Psi(x-m) + \sum_{\nu \in \mathbb{N}} \sum_{e \in E} \sum_{m \in \mathbb{Z}^d} \lambda^e_{\nu m} 2^{-\nu d/2} \Psi_e(2^{\nu-1}x - m), \ \textit{convergence in } S'(\mathbb{R}^d) \quad (1.17)$$

with

$$\|\lambda | \mathsf{b}^s_{pq}\| = \Big(\sum_{m \in \mathbb{Z}^d} |\lambda_m|^p\Big)^{\frac{1}{p}} + \Big(\sum_{\nu=1}^\infty 2^{\nu(s-\frac{d}{p})q} \sum_{e \in E} \Big(\sum_{m \in \mathbb{Z}^d} |\lambda^e_{\nu m}|^p\Big)^{\frac{q}{p}}\Big)^{\frac{1}{q}} < \infty$$

appropriately modified if $p = \infty$ and/or $q = \infty$. The representation in (1.17) is unique, the complex coefficients $(\lambda_m)_{m \in \mathbb{Z}^d}$ and $(\lambda^e_{\nu m})_{e \in E, \nu \in \mathbb{N}_0, m \in \mathbb{Z}^d}$ depend linearly on f and the mapping, which associates to $f \in B^s_{pq}(\mathbb{R}^d)$ the sequence of coefficients, is an isomorphic map of $B^s_{pq}(\mathbb{R}^d)$ onto b^s_{pq}.

(iii) Let $f \in S'(\mathbb{R}^d)$. Then $f \in F^s_{pq}(\mathbb{R}^d)$ if, and only if, it can be represented as

$$f = \sum_{m \in \mathbb{Z}^d} \lambda_m \Psi(x-m) + \sum_{\nu \in \mathbb{N}} \sum_{e \in E} \sum_{m \in \mathbb{Z}^d} \lambda^e_{\nu m} 2^{-\nu d/2} \Psi_e(2^{\nu-1}x - m), \ \textit{convergence in } S'(\mathbb{R}^d) \quad (1.18)$$

with

$$\|\lambda|\mathsf{f}^s_{pq}\| = \Big(\sum_{m\in\mathbb{Z}^d}|\lambda_m|^p\Big)^{\frac{1}{p}} + \left\|\left(\sum_{\nu=1}^{\infty}2^{\nu(s-\frac{d}{p})q}\sum_{e\in E}\sum_{m\in\mathbb{Z}^d}|\lambda^e_{\nu m}|^q\chi_{\nu m}(x)\right)^{1/q}\right\|_p < \infty$$

appropriately modified if $p = \infty$ and/or $q = \infty$. The representation in (1.18) *is unique, the complex coefficients $(\lambda_m)_{m\in\mathbb{Z}^d}$ and $(\lambda^e_{\nu m})_{e\in E,\nu\in\mathbb{N}_0,m\in\mathbb{Z}^d}$ depend linearly on f and the mapping, which associates to $f \in F^s_{pq}(\mathbb{R}^d)$ the sequence of coefficients, is an isomorphic map of $F^s_{pq}(\mathbb{R}^d)$ onto f^s_{pq}.*

Remark 1.11. The wavelet decomposition has several very convenient advantages. The decomposition (1.17) is unique and its coefficients depend in a linear way on f. Furthermore, it provides an isomorphism between the corresponding function and sequence spaces. On the other hand, the structure of the compactly supported wavelets from Theorem 1.9 is rather complicated. For example, it is known that the their support must grow linearly with k. In particular, there are no compactly supported infinitely differentiable wavelets.

1.4 Spaces on domains

Let Ω be a bounded domain. Then one may easily modify the definitions given in Section 1.1.1 to obtain function spaces on Ω. Unfortunately, Definition 1.1 relies essentially on the use of Fourier transform and does not allow such an easy modification. Therefore, the Besov and Triebel-Lizorkin spaces on Ω are usually defined by restriction. Let $D(\Omega) = C^\infty_0(\Omega)$ be the collection of all complex-valued infinitely-differentiable functions with compact support in Ω and let $D'(\Omega)$ be its dual - the space of all complex-valued distributions on Ω.

Let $g \in S'(\mathbb{R}^d)$. Then we denote by $g|\Omega$ its restriction to Ω:

$$(g|\Omega) \in D'(\Omega), \qquad (g|\Omega)(\psi) = g(\psi) \quad \text{for} \quad \psi \in D(\Omega).$$

Definition 1.12. Let Ω be a bounded domain in $\mathbb{R}^d$. Let $s \in \mathbb{R}$, $0 < p, q \le \infty$ with $p < \infty$ in the F-case. Let A^s_{pq} stand either for B^s_{pq} or F^s_{pq}. Then

$$A^s_{pq}(\Omega) = \{f \in D'(\Omega) : \exists g \in A^s_{pq}(\mathbb{R}^d) : g|\Omega = f\}$$

and

$$\|f|A^s_{pq}(\Omega)\| = \inf \|g|A^s_{pq}(\mathbb{R}^d)\|,$$

where the infimum is taken over all $g \in A^s_{pq}(\mathbb{R}^d)$ such that $g|\Omega = f$.

Although Definition 1.12 is an easy and convenient way how to define function spaces on domains, an intrinsic characterization of these spaces is necessary on many occasions. It turns out that under only minor regularity assumptions on Ω (i.e. Lipschitz boundary), the spaces may be characterized by differences (in a fashion similar to Section 1.1.1). As this will not be needed in the sequel, we only refer to [112, Section 1.11] for details and further references.

We shall later need the existence of a universal extension operator as it was given by Rychkov [96]. This result (with many forerunners for which we refer to references given in [96]) states, that if Ω has Lipschitz boundary then there is a common bounded linear extension operator Ext : $A^s_{p,q}(\Omega) \to A^s_{p,q}(\mathbb{R}^d)$ for all admissible s, p and q. Another important fact will be the existence of atomic and wavelet decomposition techniques adapted to function spaces on domains. We shall return to this point in Section 1.2.

2 Sparse recovery and compressed sensing

2.1 Introduction and notation

Compressed sensing is a novel method of signal processing, which was introduced in [41] and [18] and which profited from its very beginning from fruitful interplay between mathematicians, applied mathematicians, and electrical engineers. The mathematical concepts are inspired by ideas from a number of different disciplines, including numerical analysis, stochastic, combinatorics, and functional analysis. On the other hand, the applications of compressed sensing range from image processing [42], medical imaging [72], and radar technology [12] to sampling theory [80, 114], and statistical learning.

In this section we collect the basic mathematical ideas from numerical analysis, stochastic, and functional analysis used in the area of compressed sensing to give an overview of basic notions, including the Null Space Property and the Restricted Isometry Property, and the relations between them. Most of the material in this section can be proven with elementary methods from approximation theory and stochastic and we refer to [14] for details. We hope that this presentation will make the mathematical concepts of compressed sensing appealing and understandable both to applied mathematicians and electrical engineers. In this and that form, similar material appeared already in many one-semester courses around the world, including my lectures given in Berlin and Prague. Let us stress that the material presented in this section is by no means new or original, actually it is nowadays considered classical, or “common wisdom” throughout the community.

We refer also to more extensive summaries of compressed sensing [34, 48, 49] for more details and further references.

As the mathematical concepts of compressed sensing rely on the interplay of ideas from linear algebra, numerical analysis, stochastic, and functional analysis, we start with an overview of basic notions from these fields. We shall restrict ourselves to the minimum needed in the sequel.

By ℓ_p^n we denote the space $\mathbb{R}^n$ equipped with the (quasi-)norm

$$\|x\|_p = \begin{cases} \Big(\sum\limits_{j=1}^n |x_j|^p\Big)^{1/p}, & p\in(0,\infty);\\ \max\limits_{j=1,\dots,n}|x_j|, & p=\infty. \end{cases} \tag{2.1}$$

If $x\in\mathbb{R}^n$, we can always find a permutation $\sigma:\{1,\dots,n\}\to\{1,\dots,n\}$, such that the nonincreasing rearrangement $x^*\in[0,\infty)^n$ of x, defined by $x_j^*=|x_{\sigma(j)}|$ satisfies

$$x_1^*\ge x_2^*\ge\dots\ge x_n^*\ge 0.$$

If $T\subset\{1,\dots,n\}$ is a set of indices, we denote by $|T|$ the number of its elements. We shall complement this notation by denoting the size of the support of $x\in\mathbb{R}^n$ by

$$\|x\|_0 = |\operatorname{supp}(x)| = |\{j : x_j\neq 0\}|.$$

Note, that this expression is not even a quasinorm. The notation is justified by the observation, that

$$\lim_{p\to 0}\|x\|_p^p = \|x\|_0 \quad \text{for all} \quad x\in\mathbb{R}^n.$$

Let k be a natural number at most equal to n. A vector $x\in\mathbb{R}^n$ is called k-sparse, if $\|x\|_0\le k$ and the set of all k-sparse vectors is denoted by

$$\Sigma_k=\{x\in\mathbb{R}^n : \|x\|_0\le k\}.$$

Finally, if $k < n$, the best k-term approximation $\sigma_k(x)_p$ of $x \in \mathbb{R}^n$ describes, how well can x be approximated by k-sparse vectors in the ℓ_p^n-norm. This can be expressed by the formula

$$\sigma_k(x)_p = \inf_{z\in\Sigma_k} \|x - z\|_p = \begin{cases} \Big(\sum\limits_{j=k+1}^{n} (x_j^*)^p \Big)^{1/p}, & p \in (0,\infty); \\ x_{k+1}^*, & p = \infty. \end{cases} \tag{2.2}$$

Linear operators between finite-dimensional spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ can be represented with the help of matrices $A \in \mathbb{R}^{m\times n}$. The entries of A are denoted by a_{ij}, $i = 1,\dots,m$ and $j = 1,\dots,n$. The transpose of a matrix $A \in \mathbb{R}^{m\times n}$ is a matrix $A^T \in \mathbb{R}^{n\times m}$ with entries $(A^T)_{ij} = a_{ji}$. The identity matrix in $\mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$ will be denoted by I.

There is a number of ways how to discover the landscape of compressed sensing. The point of view, which we shall follow in this section, is that we are looking for sparse solutions $x \in \mathbb{R}^n$ of a system of linear equations $Ax = y$, where $y \in \mathbb{R}^m$ and the $m \times n$ matrix A are known. We shall be interested in underdetermined systems, i.e. in the case $m \le n$. Intuitively, this corresponds to solving the following optimization problem

$$\min_z \|z\|_0 \quad \text{subject to} \quad y = Az. \tag{P_0}$$

Unfortunately, it can be shown that this problem is numerically intractable if m and n are getting larger. Then we introduce the basic notions of compressed sensing, showing that for specific matrices A and measurement vectors y, one can recover the solution of (P_0) in a much more effective way.

2.2 Basis pursuit

The minimization problem (P_0) can obviously be solved by considering first all index sets $T \subset \{1,\dots,n\}$ with one element and employing the methods of linear algebra to decide if there is a solution x to the system with support included in T. If this fails for all such index sets, we continue with all index sets with two, three, and more elements. The obvious drawback is the rapidly increasing number of these index sets. Indeed, there is $\binom{n}{k}$ index sets $T \subset \{1,\dots,n\}$ with k elements and this quantity grows (in some sense) exponentially with k and n.

We shall start our tour through compressed sensing by discussing that even every other algorithm solving (P_0) suffers from this drawback. This will be formulated in the language of complexity theory as the statement, that the (P_0) problem is NP-hard. Before we come to that, we introduce the basic terms used in the sequel. We refer for example to [6] for an introduction to computational complexity.

The *P-class* ("polynomial time") consists of all decision problems that can be solved in polynomial time, i.e. with an algorithm, whose running time is bounded from above by a polynomial expression in the size of the input.

The *NP-class* ("nondeterministic polynomial time") consists of all decision problems, for which there is a polynomial-time algorithm V (called verifier), with the following property. If, given an input α, the right answer to the decision problem is "yes", then there is a proof β, such that $V(\alpha,\beta) =$ yes. Roughly speaking, when the answer to the decision problem is positive, then the proof of this statement can be verified with a polynomial-time algorithm.

Let us reformulate (P_0) as a decision problem. Namely, if the natural numbers k, m, n, $m \times n$ matrix A and $y \in \mathbb{R}^m$ are given, decide if there is a k-sparse solution x of the equation $Ax = y$.

It is easy to see that this version of (P_0) is in the NP-class. Indeed, if the answer to the problem is "yes" and a certificate $x \in \mathbb{R}^n$ is given, then it can be verified in polynomial time if x is k-sparse and $Ax = y$.

A problem is called *NP-hard* if any of its solving algorithms can be transformed in polynomial time into a solving algorithm of any other NP-problem. We shall rely on a statement from complexity theory, that the following problem is both NP and NP-hard.

Exact cover problem
Given as the input a natural number m divisible by 3 and a system $\{T_j : j = 1, \dots, n\}$ of subsets of $\{1, \dots, m\}$ with $|T_j| = 3$ for all $j = 1, \dots, n$, decide, if there is a subsystem of mutually disjoint sets $\{T_j : j \in J\}$, such that $\bigcup_{j \in J} T_j = \{1, \dots, m\}$. Such a subsystem is frequently referred to as *exact cover*.

Let us observe, that for any subsystem $\{T_j : j \in J\}$ it is easy to verify (in polynomial time) if it is an exact cover or not. So the problem is in the NP-class. The non-trivial statement from computational complexity is that this problem is also NP-hard. The exact formulation of (P_0) looks as follows.

ℓ_0-minimization problem
Given natural numbers m, n, an $m \times n$ matrix A and a vector $y \in \mathbb{R}^m$ as input, find the solution of

$$\min_z \|z\|_0 \quad \text{s.t.} \quad y = Az.$$

Theorem 2.1. *The ℓ_0-minimization problem is NP-hard.*

The ℓ_0-minimization problem is NP-hard, if all matrices A and all measurement vectors y are allowed as inputs. The theory of compressed sensing shows nevertheless, that for special matrices A and for $y = Ax$ for some sparse x, the problem can be solved efficiently.

In general, we replace the $\|z\|_0$ in (P_0) by some $\|z\|_p$ for $p > 0$. To obtain a convex problem, we need to have $p \geq 1$. To obtain sparse solutions, $p \leq 1$ is necessary, cf. Figure 1.

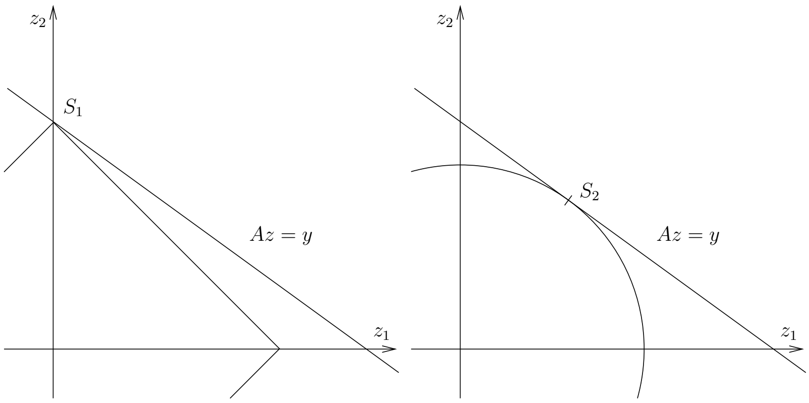


Figure 1: Solution of $S_p = \operatorname*{argmin}_{z \in \mathbb{R}^2} \|z\|_p \quad \text{s.t.} \quad y = Az$ for $p = 1$ and $p = 2$

We are therefore naturally led to discuss under which conditions the solution to (P_0) coincides with the solution of the following convex optimization problem called *basis pursuit*

$$\min_z \|z\|_1 \quad \text{s.t.} \quad y = Az, \tag{P_1}$$

which was introduced in [25]. But before we come to that, let us show, that in the real case this problem may be reformulated as a linear optimization problem, i.e. as the search for the minimizer of a linear function over a set given by linear constraints, whose number depends polynomially on the dimension. We refer to [54] for an introduction to linear programming.

Indeed, let us assume that (P_1) has a unique solution, which we denote by $x \in \mathbb{R}^n$. Then the pair (u, v) with $u = x^+$ and $v = x^-$, i.e. with

$$u_j = \begin{cases} x_j, & x_j \geq 0, \\ 0, & x_j < 0, \end{cases} \quad \text{and} \quad v_j = \begin{cases} 0, & x_j \geq 0, \\ -x_j, & x_j < 0, \end{cases}$$

is the unique solution of

$$\min_{u,v \in \mathbb{R}^n} \sum_{j=1}^{n} (u_j + v_j) \text{ s.t. } Au - Av = y \text{ and } u_j \geq 0 \text{ and } v_j \geq 0 \text{ for all } j = 1, \dots, n. \tag{2.3}$$

If namely (u', v') is another pair of vectors admissible in (2.3), then $x' = u' - v'$ satisfies $Ax' = y$ and x' is therefore admissible in (P_1). As x is the solution of (P_1), we get

$$\sum_{j=1}^{n} (u_j + v_j) = \|x\|_1 < \|x'\|_1 = \sum_{j=1}^{n} |u'_j - v'_j| \leq \sum_{j=1}^{n} (u'_j + v'_j).$$

If, on the other hand, the pair (u, v) is the unique solution of (2.3), then $x = u - v$ is the unique solution of (P_1). If namely z is another admissible vector in (P_1), then $u' = z^+$ and $v' = z^-$ are admissible in (2.3) and we obtain

$$\|x\|_1 = \sum_{j=1}^{n} |u_j - v_j| \leq \sum_{j=1}^{n} (u_j + v_j) < \sum_{j=1}^{n} (u'_j + v'_j) = \|z\|_1.$$

Very similar argument works also in the case when (P_1) has multiple solutions.

2.3 Null Space Property

If $T \subset \{1, \dots, n\}$, then we denote by $T^c = \{1, \dots, n\} \setminus T$ the complement of T in $\{1, \dots, n\}$. If furthermore $v \in \mathbb{R}^n$, then we denote by v_T either the vector in $\mathbb{R}^{|T|}$, which contains the coordinates of v on T, or the vector in $\mathbb{R}^n$, which equals v on T and is zero on T^c. It will be always clear from the context, which notation is being used.

Finally, if $A \in \mathbb{R}^{m \times n}$ is a matrix, we denote by A_T the $m \times |T|$ sub-matrix containing the columns of A indexed by T. Let us observe, that if $x \in \mathbb{R}^n$ with $T = \operatorname{supp}(x)$, that $Ax = A_T x_T$.

We start the discussion of the properties of basis pursuit by introducing the notion of Null Space Property, which first appeared in [26].

Definition 2.2. Let $A \in \mathbb{R}^{m \times n}$ and let $k \in \{1, \dots, n\}$. Then A is said to have the *Null Space Property* (NSP) of order k if

$$\|v_T\|_1 < \|v_{T^c}\|_1 \quad \text{for all } v \in \ker A \setminus \{0\} \text{ and all } T \subset \{1, \dots, n\} \text{ with } |T| \leq k. \tag{2.4}$$

Remark 2.3. (i) The condition (2.4) states that vectors from the kernel of A are well spread, i.e. not supported on a set of small size. Indeed, if $v \in \mathbb{R}^n \setminus \{0\}$ is k-sparse and $T = \operatorname{supp}(v)$, then (2.4) shows immediately, that v can not lie in the kernel of A.
(ii) If we add $\|v_{T^c}\|_1$ to both sides of (2.4), we obtain $\|v\|_1 < 2\|v_{T^c}\|_1$. If then T are the indices of the k largest coordinates of v taken in the absolute value, this inequality becomes $\|v\|_1 < 2\sigma_k(v)_1$.

Theorem 2.4. *Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$ if, and only if, A has the NSP of order k.*

Remark 2.5. Theorem 2.4 states that the solutions of (P_0) may be found by (P_1), if A has the NSP of order k and if $y \in \mathbb{R}^m$ is such that, there exists a k-sparse solution x of the equation $Ax = y$. Indeed, if in such a case, $\hat{x}$ is a solution of (P_0), then $\|\hat{x}\|_0 \le \|x\|_0 \le k$. Finally, it follows by Theorem 2.4, that $\hat{x}$ is also a solution of (P_1) and that $x = \hat{x}$.

In the language of complexity theory, if we restrict the inputs of the ℓ_0-minimization problem to matrices with the NSP of order k and to vectors y, for which there is a k-sparse solution of the equation $Ax = y$, the problem belongs to the P-class and the solving algorithm with polynomial running time is any standard algorithm solving (P_1), or the corresponding linear problem (2.3).

2.4 Restricted Isometry Property

Although the Null Space Property is equivalent to the recovery of sparse solutions of underdetermined linear systems by basis pursuit in the sense just described, it is somehow difficult to construct matrices satisfying this property. We shall therefore present a sufficient condition called Restricted Isometry Property, which was first introduced in [18], and which ensures that the Null Space Property is satisfied.

Definition 2.6. Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then the *restricted isometry constant* $\delta_k = \delta_k(A)$ of A of order k is the smallest $\delta \ge 0$, such that

$$(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2 \quad \text{for all} \quad x \in \Sigma_k. \tag{2.5}$$

Furthermore, we say that A satisfies the *Restricted Isometry Property* (RIP) of order k with the constant δ_k if $\delta_k < 1$.

Remark 2.7. The condition (2.5) states that A acts nearly isometrically when restricted to vectors from Σ_k. Of course, the smaller the constant $\delta_k(A)$ is, the closer is the matrix A to isometry on Σ_k. We will be therefore later interested in constructing matrices with small RIP constants. Finally, the inequality $\delta_1(A) \le \delta_2(A) \le \dots \le \delta_k(A)$ follows trivially.

The following theorem shows that RIP of sufficiently high order with a constant small enough is indeed a sufficient condition for NSP.

Theorem 2.8. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \le n/2$. If $\delta_{2k}(A) < 1/3$, then A has the NSP of order k.*

Combining Theorems 2.4 and 2.8, we obtain immediately the following corollary.

Corollary 2.9. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \le n/2$. If $\delta_{2k}(A) < 1/3$, then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$.*

2.5 RIP for random matrices

From what was said up to now, we know that matrices with small restricted isometry constants fulfill the null space property, and sparse solutions of underdetermined linear equations involving such matrices can be found by ℓ_1-minimization (P_1). We discuss in this chapter a class of matrices with small RIP constants. It turns out that the most simple way is to construct these matrices by taking its entries to be independent standard normal variables.

We denote until the end of this section

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \dots & \omega_{mn} \end{pmatrix}, \tag{2.6}$$

where $\omega_{ij}, i = 1, \dots, m, j = 1, \dots, n$, are i.i.d. standard normal variables. We shall show that such a matrix satisfies the RIP with reasonably small constants with high probability.

2.5.1 Concentration inequalities

If $\omega_1, \dots, \omega_m$ are (possibly dependent) standard normal random variables, then $\mathbb{E}(\omega_1^2 + \dots + \omega_m^2) = m$. If $\omega_1, \dots, \omega_m$ are even independent, then the value of $\omega_1^2 + \dots + \omega_m^2$ concentrates very strongly around m. This effect is known as *concentration of measure*, cf. [67, 68, 79].

Lemma 2.10. *Let $m \in \mathbb{N}$ and let $\omega_1, \dots, \omega_m$ be i.i.d. standard normal variables. Let $0 < \varepsilon < 1$. Then*

$$\mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \geq (1 + \varepsilon)m) \leq e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}$$

and

$$\mathbb{P}(\omega_1^2 + \dots + \omega_m^2 \leq (1 - \varepsilon)m) \leq e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}.$$

Using 2-stability of the normal distribution, Lemma 2.10 shows immediately that A defined as in (2.6) acts with high probability as isometry on one fixed $x \in \mathbb{R}^n$.

Theorem 2.11. *Let $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ and let A be as in* (2.6). *Then*

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - 1\Big| \geq t\Big) \leq 2e^{-\frac{m}{2}[t^2/2 - t^3/3]} \leq 2e^{-Cmt^2} \tag{2.7}$$

for $0 < t < 1$ with an absolute constant $C > 0$.

Remark 2.12. (i) Observe, that (2.7) may be easily rescaled to

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - \|x\|_2^2\Big| \geq t\|x\|_2^2\Big) \leq 2e^{-Cmt^2}, \tag{2.8}$$

which is true for every $x \in \mathbb{R}^n$.
(ii) A slightly different proof of (2.7) is based on the rotational invariance of the distribution underlying the random structure of matrices defined by (2.6). Therefore, it is enough to prove (2.7) only for one fixed element $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. Taking $x = e_1 = (1, 0, \dots, 0)^T$ to be the first canonical unit vector allows us to use Lemma 2.10 without the necessity of applying the 2-stability of normal distribution.

2.5.2 RIP for random Gaussian matrices

The proof of restricted isometry property of random matrices generated as in (2.6) is based on two main ingredients. The first is the concentration of measure phenomenon described in its most simple form in Lemma 2.10, and reformulated in Theorem 2.11. The second is the following entropy argument, which allows to extend Theorem 2.11 and (2.7) from one fixed $x \in \mathbb{R}^n$ to the set Σ_k of all k-sparse vectors.

Lemma 2.13. *Let $t > 0$. Then there is a set $\mathcal{N} \subset \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ with*
(i) $|\mathcal{N}| \leq (1 + 2/t)^n$ and
(ii) for every $z \in \mathbb{S}^{n-1}$, there is a $x \in \mathcal{N}$ with $\|x - z\|_2 \leq t$.

With all these tools at hand, we can now state the main theorem of this section, whose proof follows closely the arguments of [7].

Theorem 2.14. *Let* $n \geq m \geq k \geq 1$ *be natural numbers and let* $0 < \varepsilon < 1$ *and* $0 < \delta < 1$ *be real numbers with*

$$m \geq C\delta^{-2}\Big(k\ln(en/k) + \ln(2/\varepsilon)\Big), \tag{2.9}$$

where $C > 0$ *is an absolute constant. Let* A *be again defined by* (2.6)*. Then*

$$\mathbb{P}\big(\delta_k(A) \leq \delta\big) \geq 1 - \varepsilon.$$

2.5.3 Lemma of Johnson and Lindenstrauss

Concentration inequalities similar to (2.7) play an important role in several areas of mathematics. We shall present their connection to the famous result from functional analysis called Johnson-Lindenstrauss lemma, cf. [60]. The lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that the mutual distances between the points are nearly preserved. The connection between this classical result and compressed sensing was first highlighted in [7], cf. also [64].

Lemma 2.15. *Let* $0 < \varepsilon < 1$ *and let* m, N *and* n *be natural numbers with*

$$m \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1}\ln N.$$

Then for every set $\{x^1, \dots, x^N\} \subset \mathbb{R}^n$ *there exists a mapping* $f : \mathbb{R}^n \to \mathbb{R}^m$*, such that*

$$(1-\varepsilon)\|x^i - x^j\|_2^2 \leq \|f(x^i) - f(x^j)\|_2^2 \leq (1+\varepsilon)\|x^i - x^j\|_2^2, \qquad i, j \in \{1, \dots, N\}. \tag{2.10}$$

2.6 Stability and robustness

The ability to recover sparse solutions of underdetermined linear systems by quick recovery algorithms as ℓ_1-minimization is surely a very promising result. On the other hand, two additional features are obviously necessary to extend this results to real-life applications, namely

- Stability: We want to be able to recover (or at least approximate) also vectors $x \in \mathbb{R}^n$, which are not exactly sparse. Such vectors are called *compressible* and mathematically they are characterized by the assumption that their best k-term approximation decays rapidly with k. Intuitively, the faster the decay of the best k-term approximation of $x \in \mathbb{R}^n$ is, the better we should be able to approximate x.

- Robustness: Equally important, we want to recover sparse or compressible vectors from noisy measurements. The basic model here is the assumptions that the measurement vector y is given by $y = Ax + e$, where e is small (in some sense). Again, the smaller the error e is, the better we should be able to recover an approximation of x.

We shall show that the methods of compressed sensing can be extended also to this kind of scenario. There is a number of different estimates in the literature, which show that the technique of compressed sensing is stable and robust. We will present only one of them. Its proof is a modification of the proof of Theorem 2.8, and follows closely [16].

Inspired by the form of the noisy measurements just described, we will concentrate on the recovery properties of the following slight modification of (P_1). Namely, let $\eta \geq 0$, then we consider the convex optimization problem

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az - y\|_2 \leq \eta. \tag{$P_{1,\eta}$}$$

If $\eta = 0$, $(P_{1,\eta})$ reduces back to (P_1).

Theorem 2.16. *Let $\delta_{2k} < \sqrt{2} - 1$ and $\|e\|_2 \leq \eta$. Then the solution $\hat{x}$ of $(P_{1,\eta})$ satisfies*

$$\|x - \hat{x}\|_2 \leq \frac{C\sigma_k(x)_1}{\sqrt{k}} + D\eta, \tag{2.11}$$

where $C, D > 0$ are two universal positive constants.

2.7 Optimality of bounds

When recovering k-sparse vectors one obviously needs at least $m \geq k$ linear measurements. Even when the support of the unknown vector would be known, this number of measurements would be necessary to identify the value of the non-zero coordinates. Therefore, the dependence of the bound (2.9) on k can possibly only be improved in the logarithmic factor. Theorem 2.18 that even that is not possible and that this dependence is already optimal as soon as a stable recovery of k-sparse vectors is requested. The approach presented here is essentially taken over from [49].

The proof is based on the following combinatorial lemma, which plays also a fundamental role in coding theory.

Lemma 2.17. *Let $k \leq n$ be two natural numbers. Then there are N subsets $T_1, \dots, T_N$ of $\{1, \dots, n\}$, such that*

(i) $N \geq \left(\frac{n}{4k}\right)^{k/2}$,

(ii) $|T_i| = k$ *for all* $i = 1, \dots, N$ *and*

(iii) $|T_i \cap T_j| < k/2$ *for all* $i \neq j$.

The following theorem shows that any stable recovery of sparse solutions requires at least m measurements, where m is of the order $k \ln(en/k)$.

Theorem 2.18. *Let $k \leq m \leq n$ be natural numbers, let $A \in \mathbb{R}^{m\times n}$ be a measurement matrix, and let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ be an arbitrary recovery map such that for some constant $C > 0$*

$$\|x - \Delta(Ax)\|_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} \quad \text{for all} \quad x \in \mathbb{R}^n. \tag{2.12}$$

Then

$$m \geq C' k \ln(en/k) \tag{2.13}$$

with some other constant C' depending only on C.

Part II

Results of the thesis

After giving the general background in the first part, we discuss in the second part the results of the thesis. Essentially, we browse through the included publications one after another and comment on its main results. Due to the amount of the material, we shall be very brief and refer to the original publications for details.

For better readability, the results are grouped into four areas, namely

- Function spaces
- Compressed sensing and related topics
- Ridge functions
- Applications in machine learning

3 Results on function spaces

The results in this section deal with function spaces, mostly with its decomposition techniques. They were published in the following works:

[P1] J. Vybíral, A new proof of Jawerth-Franke embedding, Rev. Mat. Complut. 21 (2008), 75–82.

[P2] J. Vybíral, Widths of embeddings in function spaces, J. Compl. 24 (2008), 545–570.

[P3] J. Vybíral, Sobolev and Jawerth embeddings for spaces with variable smoothness and integrability, Ann. Acad. Sci. Fenn. Math. 34:2 (2009), 529–544.

[P4] C. Schneider and J. Vybíral, Non-smooth atomic decompositions, traces on Lipschitz domains, and pointwise multipliers in function spaces, J. Funct. Anal. 264 (5) (2013),1197–1237

[P5] H. Kempka and J. Vybíral, Spaces of variable smoothness and integrability: Characterizations by local means and ball means of differences, J. Fourier Anal. Appl. 18 (4) (2012), 852–891.

3.1 A new proof of Jawerth-Franke embedding

The classical Sobolev embedding (1.9) is in this frame of function spaces complemented by the Jawerth-Franke embedding (1.10), which describes the B to F and F to B embedding in the limiting case. The classical proofs of Jawerth and Franke [50, 59] used heavily the interpolation theory. We provided an alternative proof. Based on isomorphisms between function and sequence spaces, it is a straightforward observation that (1.10) holds if, and only if, the same is true for the sequence spaces $b^s_{p,q}$ and $f^s_{p,q}$.

The proof given in [P1] is largely self-contained, without any interpolation theory. The main ingredient is the fact that the sequence spaces $b^s_{p,q}$ and $f^s_{p,q}$ have the lattice structure. Namely, if $(\lambda_{\nu,m})_{\nu,m}$ and $(\lambda'_{\nu,m})_{\nu,m}$ are two sequences with $|\lambda_{\nu,m}| \le |\lambda'_{\nu,m}|$ for all $\nu \in \mathbb{N}_0$ and $m \in \mathbb{Z}^d$, then $\|\lambda|b^s_{p,q}\| \leq \|\lambda'|b^s_{p,q}\|$. This observation allows to use techniques like the non-increasing rearrangement of a sequence or function.

The main advantage of this technique seems to be its universality. Since its introduction in [P1], the same approach was used to provide Jawerth-Franke type embeddings for function spaces of dominating mixed smoothness [55], function spaces defined by their subatomic decompositions [98] and to spaces built upon Morrey spaces [56].

3.2 Widths of embeddings in function spaces

To describe the properties of infinite-dimensional objects (like function spaces, or operators between them), one may use several different tools. The prominent role among them is played by the theory of s-numbers as developed by Pietsch, cf. [92]. Roughly speaking, one associates to every linear operator T from one (quasi-)Banach space X into another (quasi-)Banach space Y a (non-increasing) sequence of non-negative real numbers $s_n(T)$. The properties of T are then reflected in the speed of the decay of $s_n(T)$. This approach takes it motivation from approximation theory, where it was intuitively used already in the nineteenth century. We refer to [92, 23] for further details.

Let Ω be a bounded Lipschitz domain and let $0 < p_1, p_2, q_1, q_2 \leq \infty$ and $s_1, s_2 \in \mathbb{R}$ be real numbers with

$$s_1 - s_2 > d\Big(\frac{1}{p_1} - \frac{1}{p_2}\Big)_+. \tag{3.1}$$

Then the embedding

$$\mathcal{I}d : B^{s_1}_{p_1q_1}(\Omega) \to B^{s_2}_{p_2q_2}(\Omega) \tag{3.2}$$

is compact. Using Theorem 1.10 and the existence of a universal extension operator due to Rychkov [96], the question may be reduced to the corresponding problem on the sequence space level. We obtain

$$s_n(\mathcal{I}d : B^{s_1}_{p_1q_1}(\Omega) \to B^{s_2}_{p_2q_2}(\Omega)) \approx s_n(id : \mathsf{b}^{s,\Omega}_{pq} \to \mathsf{b}^{s,\Omega}_{pq}), \tag{3.3}$$

where $\mathsf{b}^{s,\Omega}_{pq}$ is a certain variant of the spaces b^{s}_{pq} as described in Theorem 1.10 adapted to function spaces on domains.

The discretization technique was used in connection with s-numbers and embeddings of function spaces already in [73] and [71]. We refer also to [69] and [93] for the survey of the state of the art as it was in the second half of 1980's and to [70] for a more modern presentation. The main aim of the presented paper [P2] was to collect the known facts, to extend the results to the case of quasi-Banach spaces and to fill some minor gaps left up to that time. Finally, we remark that the behavior of s-numbers in connection with function spaces with dominating mixed smoothness was studied in the classical book of Temlyakov [103] and in the more recent papers [10, 11, 43, 44].

Before we discuss the results, let us define the three most important s-numbers, namely the *approximation, Kolmogorov and Gelfand numbers.*

The approximation numbers of the operator T describe, how well may this operator be approximated (in the operator norm) be finite rank operators.

Definition 3.1. Let X, Y be two quasi-Banach spaces and let $T \in \mathcal{L}(X, Y)$.

- For $n \in \mathbb{N}$, we define the nth approximation number by

$$a_n(T) = \inf\{\|T - L\| : L \in \mathcal{L}(X, Y), \mathrm{rank}(L) < n\}. \tag{3.4}$$

- For $n \in \mathbb{N}$, we define the nth Kolmogorov number by

$$d_n(T) = \inf\{\|Q_N^Y T\| : N \subset\subset Y, \dim(N) < n\}. \tag{3.5}$$

Here, Q_N^Y stands for the natural surjection of Y onto the quotient space Y/N.

- For $n \in \mathbb{N}$, we define the nth Gelfand number by

$$c_n(T) = \inf\{\|T J_M^X\| : M \subset\subset X, \operatorname{codim}(M) < n\}. \tag{3.6}$$

Here, J_M^X stands for the natural injection of M into X.

This definition goes back to Pietsch [91] and Tikhomirov [106].

Paper [P2] uses the wavelet decomposition techniques to reduce the question to the sequence space level, cf. (3.3), and the known results on these widths on the sequence space level to provide asymptotic behaviour of widths of (3.2). As the results depend typically on a number of parameters, we do not present them here and refer to [P2] for details.

3.3 Sobolev and Jawerth embeddings for spaces with variable smoothness and integrability

Paper [P3] studies the spaces of variable smoothness and integrability as introduced recently by L. Diening, P. Hästö, and S. Roudenko in [40].

The definition of these spaces is based on the Lebesgue spaces of variable integrability. The modern era of interest in these spaces dates back essentially to the paper by Kováčik and Rákosník [63].

Definition 3.2. Let $p : \mathbb{R}^d \to (0,\infty)$ be a measurable function. Then the space $L_{p(\cdot)}(\mathbb{R}^d)$ consists of all measurable functions $f : \mathbb{R}^d \to [-\infty,\infty]$ such that $\|f|L_{p(\cdot)}(\mathbb{R}^d)\| < \infty$, where

$$\|f|L_{p(\cdot)}(\mathbb{R}^d)\| = \inf\{\lambda > 0 : \int_{\mathbb{R}^d} \left(\frac{|f(x)|}{\lambda}\right)^{p(x)} dx \le 1\}$$

is the Minkowski functional of the set $\{f : \int_{\mathbb{R}^d} |f(x)|^{p(x)} dx \le 1\}$.

To ensure that $L_{p(\cdot)}(\mathbb{R}^d)$ are quasi-Banach spaces, we assume that

$$p^- := \inf_{x\in\mathbb{R}^d} p(x) > 0.$$

Furthermore, to avoid the known difficulties of the Triebel-Lizorkin scale for $p = \infty$, we require also that

$$p^+ = \sup_{x\in\mathbb{R}^d} p(x) < \infty,$$

hence we assume that

$$0 < p^- := \inf_{z\in\mathbb{R}^d} p(z) \le p(x) \le \sup_{z\in\mathbb{R}^d} p(z) =: p^+ < \infty, \quad x \in \mathbb{R}^d. \tag{3.7}$$

This allows to define Triebel-Lizorkin spaces of variable smoothness and integrability by assuming that s, p and q in Definition 1.1 are (locally integrable) functions of x.

Definition 3.3. Let $s:\mathbb{R}^d\to\mathbb{R}$, $p:\mathbb{R}^d\to(0,\infty)$ and $q:\mathbb{R}^d\to(0,\infty]$ be measurable functions. Then $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)$ is the collection of all $f\in S'(\mathbb{R}^d)$ such that

$$\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)\|=\left\|\left(\sum_{j=0}^{\infty}2^{js(\cdot)q(\cdot)}|(\varphi_j\widehat{f})^{\vee}(\cdot)|^{q(\cdot)}\right)^{1/q(\cdot)}|L_{p(\cdot)}(\mathbb{R}^d)\right\|<\infty \tag{3.8}$$

(with the usual modification for $q(x)=\infty$). Here, the sequence $(\varphi_j)_{j\in\mathbb{N}_0}$ is the decomposition of unity used in Definition 1.1.

This definition places (almost) no conditions on the functional parameters s,p and q. Unfortunately, in that case the spaces may depend on the choice of the decomposition of unity - an effect very well from the theory of $F^s_{\infty,q}$-spaces, cf. [120]. Therefore we pose some regularity restrictions (identical to those made in [40]).

Definition 3.4. Let g be a continuous function on $\mathbb{R}^d$.

(i) We say that g is *1-locally* log-*Hölder continuous*, abbreviated $g\in C^{\log}_{1-\mathrm{loc}}(\mathbb{R}^d)$, if there exists $c>0$ such that

$$|g(x)-g(y)|\le\frac{c}{\log(e+1/\|x-y\|_\infty)}\quad\text{for all}\quad x,y\in\mathbb{R}^d\quad\text{with}\quad\|x-y\|_\infty\le 1.$$

Here, $\|z\|_\infty=\max\{|z_1|,\dots,|z_d|\}$ denotes the maximum norm of $z\in\mathbb{R}^d$.

(ii) We say that g is *locally* log-*Hölder continuous*, abbreviated $g\in C^{\log}_{\mathrm{loc}}(\mathbb{R}^d)$, if there exists $c>0$ such that

$$|g(x)-g(y)|\le\frac{c}{\log(e+1/|x-y|)},\quad x,y\in\mathbb{R}^d.$$

(iii) We say that g is *globally* log-*Hölder continuous*, abbreviated $g\in C^{\log}(\mathbb{R}^d)$, if it is locally log-Hölder continuous and there exists $c>0$ and $g_\infty\in\mathbb{R}$ such that

$$|g(x)-g_\infty|\le\frac{c}{\log(e+|x|)},\quad x\in\mathbb{R}^d.$$

Definition 3.5. (Standing assumptions of [40]). Let p and q be positive functions on $\mathbb{R}^d$ such that $\frac{1}{p},\frac{1}{q}\in C^{\log}(\mathbb{R}^d)$ and let $s\in C^{\log}_{\mathrm{loc}}(\mathbb{R}^d)$ with $s(x)\ge 0$ and let $s(x)$ have a limit at infinity.

Remark 3.6. Our approach in [P3] was based on the results of [40]. Especially, to ensure that the norm (3.8) does not depend on the choice of the decomposition of unity, it was necessary to pose the standing assumptions throughout. Later on, Kempka [62] proved that (3.8) gives equivalent quasi-norms for different decompositions of unity also for a wider range of parameters.

We introduce the sequence spaces associated with the Triebel-Lizorkin spaces of variable smoothness and integrability. We shall use again the notation of the dyadic cubes as given in Definition 1.6. If

$$\gamma=\{\gamma_{jm}\in\mathbb{C}:j\in\mathbb{N}_0,m\in\mathbb{Z}^d\},$$

$-\infty<s(x)<\infty$, $0<p(x)<\infty$ and $0<q(x)\le\infty$ for all $x\in\mathbb{R}^d$, we define

$$\begin{aligned}\|\gamma|f^{s(\cdot)}_{p(\cdot),q(\cdot)}\|&=\left\|\left(\sum_{j=0}^{\infty}\sum_{m\in\mathbb{Z}^d}2^{js(\cdot)q(\cdot)}|\gamma_{jm}|^{q(\cdot)}\chi_{jm}(\cdot)\right)^{1/q(\cdot)}|L_{p(\cdot)}(\mathbb{R}^d)\right\| &(3.9)\\ &=\left\|\sum_{j=0}^{\infty}\sum_{m\in\mathbb{Z}^d}2^{js(\cdot)}|\gamma_{jm}|\chi_{jm}(\cdot)|L_{p(\cdot)}(\ell_{q(\cdot)})\right\|.\end{aligned}$$

Establishing the connection between the function spaces $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)$ and the sequence spaces $f^{s(\cdot)}_{p(\cdot),q(\cdot)}$ was the main aim of [40]. Following [51] and [52], these authors investigated the properties of the φ-transform (as discussed briefly in Section 1.3 and denoted by S_φ) and obtained the following result.

Theorem 3.7. ([40], Corollary 3.9) *Under the Standing assumptions of [40]*

$$\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)\| \approx \|S_\varphi f|f^{s(\cdot)}_{p(\cdot),q(\cdot)}\|$$

with constants independent of $f \in F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)$.

Although the technique of non-increasing rearrangement fails in many aspects in the frame of variable-exponent Lebesgue spaces, it was possible to use some ideas from [P1] and to prove the embedding theorem for the sequence spaces. If the first summability index $q(\cdot)$ should be replaced by ∞ (as one would guess from (1.9)), we have to assume that $s_0(x)$ is strictly larger than $s_1(x)$, i.e. $\inf_{x\in\mathbb{R}^d}(s_0(x)-s_1(x))>0$.

Theorem 3.8. ([P3], Theorems 3.1 and 3.2) *Let $-\infty < s_1(x) \le s_0(x) < \infty$, $0 < p_0(x) \le p_1(x) < \infty$ for all $x \in \mathbb{R}^d$ with $0 < p_0^- \le p_1^+ < \infty$. Let $s_0, \frac{1}{p_0} \in C^{\log}_{1-\mathrm{loc}}(\mathbb{R}^d)$ and*

$$s_0(x) - \frac{d}{p_0(x)} = s_1(x) - \frac{d}{p_1(x)}, \quad x \in \mathbb{R}^d.$$

(i) Let $q(x) = \infty$ for all $x \in \mathbb{R}^d$ or $0 < q^- \le q(x) < \infty$ for all $x \in \mathbb{R}^d$. Then

$$f^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}.$$

(ii) Let

$$\varepsilon := \inf_{x\in\mathbb{R}^d}(s_0(x)-s_1(x)) = d\inf_{x\in\mathbb{R}^d}\left(\frac{1}{p_0(x)}-\frac{1}{p_1(x)}\right) > 0. \tag{3.10}$$

Then, for every $0 < q \le \infty$,

$$f^{s_0(\cdot)}_{p_0(\cdot),\infty} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q}.$$

Using the theory of [40], our results can be translated immediately into embeddings of function spaces.

Theorem 3.9. ([P3], Theorem 3.4) *Let $s_0, s_1, p_0, p_1, q, q_0$ and q_1 be continuous functions satisfying the Standing assumptions of [40] with $s_0(x) \ge s_1(x)$ and $p_0(x) \le p_1(x)$ for all $x \in \mathbb{R}^d$ and*

$$s_0(x) - \frac{d}{p_0(x)} = s_1(x) - \frac{d}{p_1(x)}, \quad x \in \mathbb{R}^d.$$

(i) Then

$$F^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)}(\mathbb{R}^d) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}(\mathbb{R}^d).$$

(ii) If moreover

$$\inf_{x\in\mathbb{R}^d}(s_0(x)-s_1(x)) = d\inf_{x\in\mathbb{R}^d}\Big(\frac{1}{p_0(x)}-\frac{1}{p_1(x)}\Big) > 0,$$

then

$$F^{s_0(\cdot)}_{p_0(\cdot),q_0(\cdot)}(\mathbb{R}^d) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q_1(\cdot)}(\mathbb{R}^d).$$

The proof of Theorem 3.9 follows directly from the corresponding estimates on the sequence space level (cf. Theorem 3.8) and the properties of the φ-transform (cf. Theorem 3.7). One may observe that the conditions posed on the sequence space level are much milder than those of Theorem 3.7.

Let us remark that using the recent results of Kempka [62], one can obtain a connection between $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^d)$ and $f^{s(\cdot)}_{p(\cdot),q(\cdot)}$ for a larger set of parameters, which would then lead to an improvement of Theorem 3.9.

3.4 Non-smooth atomic decompositions, traces on Lipschitz domains, and pointwise multipliers in function spaces

There are several definitions of Besov spaces $B^s_{p,q}(\mathbb{R}^n)$ to be found in the literature. Two of the most prominent approaches are the *Fourier-analytic approach* using Fourier transforms on the one hand and the *classical approach* via higher order differences involving the modulus of smoothness on the other. These two definitions are equivalent only with certain restrictions on the parameters, in particular, they differ for $0 < p < 1$ and $0 < s \leq n(\frac{1}{p} - 1)$, but may otherwise share similar properties.

In [P4] we focused on the *classical approach*, which introduces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ as those subspaces of $L_p(\mathbb{R}^n)$ such that

$$\|f|\mathbf{B}^s_{p,q}(\mathbb{R}^n)\|_r = \|f|L_p(\mathbb{R}^n)\| + \left(\int_0^1 t^{-sq}\omega_r(f,t)_p^q \,\frac{\mathrm{d}t}{t}\right)^{1/q}$$

is finite, where $0 < p, q \leq \infty$, $s > 0$, $r \in \mathbb{N}$ with $r > s$, and $\omega_r(f,t)_p$ is the usual r-th modulus of smoothness of $f \in L_p(\mathbb{R}^n)$. Choosing different values of $r > s$ leads to the same space in the sense of equivalent quasi-norms. These spaces occur naturally in nonlinear approximation theory, especially in the case $p < 1$ where they are needed in the description of approximation classes for the classical methods such as rational approximation and approximation by splines with free knots.

We developed the so-called non-smooth atomic decompositions of these spaces, where the conditions (1.11) and (1.12) get replaced by the less restrictive $\|a(2^{-j}\cdot)|B^\sigma_p(\mathbb{R}^n)\| \leq 1$.

This allowed us to prove

Theorem 3.10. *Let $n \geq 2$, $0 < p, q \leq \infty$, $0 < s < 1$, and let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ with boundary Γ. Then the operator*

$$\mathrm{tr}: \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega) \longrightarrow \mathbf{B}^s_{p,q}(\Gamma) \tag{3.11}$$

is linear and bounded.

Theorem 3.11. *Let $n \geq 2$ and Ω be a bounded Lipschitz domain with boundary Γ. Then for $0 < s < 1$ and $0 < p, q \leq \infty$ there is a bounded (non-linear) extension operator*

$$\widetilde{Ext}: \mathbf{B}^s_{p,q}(\Gamma) \longrightarrow \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega). \tag{3.12}$$

The existence of non-smooth atomic decompositions was then further used to characterize the trace space also in the limiting cases and to derive statements about pointwise multipliers. We refer to [P4] for details.

3.5 Spaces of variable smoothness and integrability: Characterizations by local means and ball means of differences

If

$$s > \sigma_p = n\left(\frac{1}{\min(p,1)} - 1\right) \tag{3.13}$$

in the B-case and

$$s > \sigma_{p,q} = n\left(\frac{1}{\min(p,q,1)} - 1\right) \tag{3.14}$$

in the F-case, Besov and Triebel-Lizorkin spaces with constant indices may be characterized by expressions involving only the differences of the function values without any use of Fourier analysis. Paper [P5] shows that the same is true also for spaces with variable indices. Let us first give the necessary notation.

Let f be a function on $\mathbb{R}^n$ and let $h \in \mathbb{R}^n$. Then we define

$$\Delta_h^1 f(x) = f(x+h) - f(x), \quad x \in \mathbb{R}^n.$$

The higher order differences are defined inductively by

$$\Delta_h^M f(x) = \Delta_h^1(\Delta_h^{M-1} f)(x), \quad M = 2, 3, \dots$$

This definition also allows a direct formula

$$\Delta_h^M f(x) := \sum_{j=0}^{M} (-1)^j \binom{M}{j} f(x + (M-j)h). \tag{3.15}$$

By *ball means of differences* we mean the quantity

$$d_t^M f(x) = t^{-n} \int_{|h|\le t} |\Delta_h^M f(x)| dh = \int_B |\Delta_{th}^M f(x)| dh,$$

where $B = \{y \in \mathbb{R}^n : |y| < 1\}$ is the unit ball of $\mathbb{R}^n$, $t > 0$ is a real number and M is a natural number.

Let us now introduce the (quasi-)norms, which shall be the main subject of our study. We define

$$\begin{aligned} \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^* &:= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| \\ &\quad + \left\| \left(\int_0^\infty t^{-s(x)q(x)} \left(d_t^M f(x)\right)^{q(x)} \frac{dt}{t} \right)^{1/q(x)} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \end{aligned} \tag{3.16}$$

and its partially discretized counterpart

$$\begin{aligned} \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**} &:= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| \\ &\quad + \left\| \left(\sum_{k=-\infty}^{\infty} 2^{ks(x)q(x)} \left(d_{2^{-k}}^M f(x)\right)^{q(x)} \right)^{1/q(x)} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \\ &= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| + \left\| \left(2^{ks(x)} d_{2^{-k}}^M f(x)\right)_{k=-\infty}^{\infty} \middle| L_{p(\cdot)}(\ell_{q(\cdot)}) \right\|. \end{aligned} \tag{3.17}$$

The norm $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ admits a direct counterpart also for Besov spaces, namely

$$\|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**} := \|f|L_{p(\cdot)}(\mathbb{R}^n)\| + \left\| \left(2^{ks(x)} d_{2^{-k}}^M f(x)\right)_{k=-\infty}^{\infty} |\ell_{q(\cdot)}(L_{p(\cdot)}) \right\|, \tag{3.18}$$

where $\ell_{q(\cdot)}(L_{p(\cdot)})$ is the (quasi-)Banach space of sequences of functions introduced in [5].

Using the notation introduced above, we may now state the main result of [P5].

Theorem 3.12. *(i) Let $p,q\in\mathcal{P}^{\log}(\mathbb{R}^n)$ with $p^+,q^+<\infty$ and $s\in C^{\log}_{loc}(\mathbb{R}^n)$. Let $M\in\mathbb{N}$ with $M>s^+$ and let*

$$s^->\sigma_{p^-,q^-}\cdot\left[1+\frac{c_{\log}(s)}{n}\cdot\min(p^-,q^-)\right]. \tag{3.19}$$

Then

$$F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)=\{f\in L_{p(\cdot)}(\mathbb{R}^n)\cap\mathcal{S}'(\mathbb{R}^n):\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^*<\infty\}$$

and $\|\cdot|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ and $\|\cdot|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^*$ are equivalent on $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$. The same holds for $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$.

(ii) Let $p,q\in\mathcal{P}^{\log}(\mathbb{R}^n)$ and $s\in C^{\log}_{loc}(\mathbb{R}^n)$. Let $M\in\mathbb{N}$ with $M>s^+$ and let

$$s^->\sigma_{p^-}\cdot\left[1+\frac{c_{\log}(1/q)}{n}+\frac{c_{\log}(s)}{n}\cdot p^-\right]. \tag{3.20}$$

Then

$$B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)=\{f\in L_{p(\cdot)}(\mathbb{R}^n)\cap\mathcal{S}'(\mathbb{R}^n):\|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}<\infty\}$$

and $\|\cdot|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ and $\|\cdot|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ are equivalent on $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

Remark 3.13. Let us comment on the rather technical conditions (3.19) and (3.20).

- If $\min(p^-,q^-)\geq 1$, then (3.19) becomes just $s^->0$. Furthermore, if p, q and s are constant functions, then (3.19) coincides with (3.14).
- If $p^-\geq 1$, then (3.20) reduces also to $s^->0$ and in the case of constant exponents we again recover (3.13).

We refer to [P5] for the proof of this assertion. We only mention that it is based on the local mean characterization. In the isotropic case, this tool goes back to Rychkov [95], for spaces with variable indices it was developed in [P5].

4 Compressed sensing and related topics

In this part we review the results of this thesis, which are connected directly to the theory of compressed sensing. They were published in one survey chapter and four research papers:

[P6] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, A Survey of Compressed Sensing, First chapter in Compressed Sensing and its Applications, Birkäuser, Springer, 2015

[P7] A. Hinrichs and J. Vybíral, Johnson-Lindenstrauss lemma for circulant matrices. Random Struct. Algor. 39(3) (2011), 391–398

[P8] J. Vybíral, A variant of the Johnson-Lindenstrauss lemma for circulant matrices, J. Funct. Anal. 260(4) (2011), 1096–1105

[P9] J. Vybíral, Average best m-term approximation, Constr. Approx. 36 (1) (2012), 83–115

[P10] M. Fornasier, J. Haškovec, and J. Vybíral, Particle systems and kinetic equations modeling interacting agents in high dimension, SIAM: Multiscale Modeling and Simulation, 9(4)(2011), 1727–1764

4.1 A Survey of Compressed Sensing

In December 2013, Holger Boche (Technical University Munich), Robert Calderbank (Duke University), Gitta Kutyniok and Jan Vybíral (both Technical University Berlin) organized the Matheon workshop on Compressed Sensing and its Applications (CSA2013). The proceedings of this workshop with contributions from the plenary and invited speakers were then published by Birkhäuser, Springer. This chapter was the introductory one, its main aim was to present the most important aspects of the theory of compressed sensing with self-contained proofs, accessible also to non-mathematicians. This chapter was mainly based on the book [49] and the course on the subject given by the last author at TU Berlin. We followed this chapter closely in our introduction of compressed sensing in Section 2.

4.2 Johnson-Lindenstrauss lemma for circulant matrices

In papers [P7] and [P8] we studied the possibility of using circulant matrices in the random dimensionality reduction as described by the Johnson-Lindenstrauss lemma 2.15.

The original proof of Johnson and Lindenstrauss [60] uses (up to a scaling factor) an orthogonal projection onto a random k-dimensional subspace of $\mathbb{R}^d$. We refer also to [33] for a beautiful and self-contained proof. Later on, this lemma found many applications, especially in design of algorithms, where it sometimes allows to reduce the dimension of the underlying problem essentially and break the so-called "curse of dimension", cf. [57] or [58].

The evaluation of $f(x)$, where f is a projection onto a random k dimensional subspace, is a very time-consuming operation. Therefore, a significant effort was devoted to

- minimize the running time of $f(x)$,
- minimize the memory used,
- minimize the number of random bits used,
- simplify the algorithm to allow an easy implementation.

There has been an enormous effort to provide improved constructions of Johnson-Lindenstrauss mappings [1, 4, 74, 15] and references therein. Let us recall that the close connection between Johnson-Lindenstrauss lemma and the Restricted Isometry Property is nowadays well understood, cf. [7] and [64].

Papers [P7] and [P8] investigated the possibility of using structured random matrices for dimensionality reduction. Let us give the necessary definitions and the statement of the theorem proven in [P7].

Let $a = (a_0, \dots, a_{d-1})$ be independent identically distributed random variables. We denote by $M_{a,k}$ the partial circulant matrix

$$M_{a,k} = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{d-1} \\ a_{d-1} & a_0 & a_1 & \dots & a_{d-2} \\ a_{d-2} & a_{d-1} & a_0 & \dots & a_{d-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{d-k+1} & a_{d-k+2} & a_{d-k+3} & \dots & a_{d-k} \end{pmatrix}.$$

Furthermore, if $\varkappa = (\varkappa_0, \dots, \varkappa_{d-1})$ are independent Bernoulli variables, we put

$$D_\varkappa = \begin{pmatrix} \varkappa_0 & 0 & \dots & 0 \\ 0 & \varkappa_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varkappa_{d-1} \end{pmatrix}.$$

The main result of [P7] was then the following statement.

Theorem 4.1. *Let $x_1, \dots, x_n$ be arbitrary points in $\mathbb{R}^d$, let $\varepsilon \in (0, \frac{1}{2})$ and let $k = \Omega(\varepsilon^{-2} \log^3 n)$ be a natural number. Let $a = (a_0, \dots, a_{d-1})$ be independent Bernoulli variables or independent normally distributed variables. Let $M_{a,k}$ and $D_\varkappa$ be as above and put $f(x) = \frac{1}{\sqrt{k}} M_{a,k} D_\varkappa x$.*

Then with probability at least 2/3 the following holds

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|f(x_i) - f(x_j)\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2, \qquad i, j = 1, \dots, n.$$

The proof is based on decoupling the dependencies of the randomness used in the entries. Obviously, the main disadvantage of Theorem 4.1 is the high dependence of k on n. This was improved in [P8], where a similar theorem was proven with $k = \Omega(\varepsilon^{-2} \log^2 n)$. The proof techniques used in [P8] differ essentially, and are of more geometric nature.

4.3 Average best m-term approximation

The concept of best m-term approximation was defined in (2.2) and is the main prototype of non-linear approximation, cf. [104, 37]. Moreover for $0 < p \le q \le \infty$, we introduce the *best m-term approximation widths*

$$\sigma_m^{p,q} := \sup_{x:\|x\|_p \le 1} \sigma_m(x)_q.$$

The use of this concept goes back to Schmidt [97] and after the work of Oskolkov [87], it was widely used in the approximation theory, cf. [32, 38]. It is well known that

$$2^{-1/p}(m+1)^{1/q-1/p} \le \sigma_m^{p,q} \le (m+1)^{1/q-1/p}, \quad m = 0, 1, 2, \dots. \tag{4.1}$$

The proof of (4.1) is based on the simple fact that (roughly speaking) the best m-term approximation error of $x \in \ell_p$ is realized by subtracting the m largest coefficients taken in absolute value. Hence,

$$\sigma_m(x)_q = \begin{cases} \left(\sum_{j=m+1}^\infty (x_j^*)^q \right)^{1/q}, & 0 < q < \infty, \\ x_{m+1}^* = \sup_{j \ge m+1} x_j^*, & q = \infty, \end{cases}$$

where $x^* = (x_1^*, x_2^*, \dots)$ denotes the so-called *non-increasing rearrangement* [12] of the vector $(|x_1|, |x_2|, |x_3|, \dots)$.

Let us recall the proof of (4.1) in the simplest case, namely $q = \infty$. The estimate from above then follows by

$$\sigma_m(x)_\infty = \sup_{j \ge m+1} x_j^* = x_{m+1}^* \le \left((m+1)^{-1} \sum_{j=1}^{m+1} (x_j^*)^p \right)^{1/p} \le (m+1)^{-1/p} \|x\|_p. \tag{4.2}$$

The lower estimate is supplied by taking

$$x = (m+1)^{-1/p} \sum_{j=1}^{m+1} e_j, \tag{4.3}$$

where $\{e_j\}_{j=1}^\infty$ are the canonical unit vectors.

For general q, the estimate from above in (4.1) may be obtained from (4.2) and Hölder's inequality

$$\|x\|_q \le \|x\|_p^\theta \cdot \|x\|_\infty^{1-\theta}, \quad \text{where} \quad \frac{1}{q} = \frac{\theta}{p}. \tag{4.4}$$

The estimate from below follows for all q's by simple modification of (4.3).

The discussion above exhibits two effects.

(i) Best m-term approximation works particularly well, when $1/p - 1/q$ is large, i.e. if $p < 1$ and $q = \infty$.

(ii) The elements used in the estimate from below (and hence the elements, where the best m-term approximation performs at worse) enjoy a very special structure.

Therefore, there is a reasonable hope that the best m-term approximation could behave better, when considered in a certain average case. We now present the definition of the so-called *average best m-term widths*, which were the main subject of our study in [P9].

First, we observe that

$$\sigma_m((x_1,\dots,x_n))_q = \sigma_m((\varepsilon_1 x_1,\dots,\varepsilon_n x_n))_q = \sigma_m((|x_1|,\dots,|x_n|))_q$$

holds for every $x \in \mathbb{R}^n$ and $\varepsilon \in \{-1,+1\}^n$. Also all the measures, which we shall consider, are invariant under any of the mappings

$$(x_1,\dots,x_n) \to (\varepsilon_1 x_1,\dots,\varepsilon_n x_n), \quad \varepsilon \in \{-1,+1\}^n$$

and therefore we restrict our attention only to $\mathbb{R}^n_+$ in the following definition.

Definition 4.2. Let $0 < p \le q \le \infty$ and let $n \ge 2$ and $0 \le m \le n-1$ be natural numbers.

(i) We set

$$\Delta_p^n = \begin{cases} \{(t_1,\dots,t_n) \in \mathbb{R}^n_+ : \sum_{j=1}^n t_j^p = 1\}, & p < \infty, \\ \{(t_1,\dots,t_n) \in \mathbb{R}^n_+ : \max_{j=1,\dots,n} t_j = 1\}, & p = \infty. \end{cases}$$

(ii) Let μ be a Borel probability measure on Δ_p^n. Then

$$\sigma_m^{p,q}(\mu) = \int_{\Delta_p^n} \sigma_m(x)_q d\mu(x)$$

is called *average surface best m-term width of $id : \ell_p^n \to \ell_q^n$ with respect to μ.*

(iii) Let ν be a Borel probability measure on $[0,1]\cdot\Delta_p^n$. Then

$$\sigma_m^{p,q}(\nu) = \int_{[0,1]\cdot\Delta_p^n} \sigma_m(x)_q d\nu(x)$$

is called *average volume best m-term width of $id : \ell_p^n \to \ell_q^n$ with respect to ν.*

Following the classical works from geometry of Banach spaces [8, 9, 47, 78, 79, 81, 82] we were able to characterize these widths for classical measures on Δ_p^n including the normalized Lebesgue measure, the $n-1$ dimensional Hausdorff measure restricted to the surface of Δ_p^n, and for the so-called *cone measure*. We refer to [P9] for the detailed statements of the results.

4.4 Particle systems and kinetic equations modeling interacting agents in high dimension

The starting point of [P10] is the well-known Cucker-Smale model, introduced and analyzed in [30, 31], which is described by

$$\dot{x}_i = v_i \in \mathbb{R}^d, \tag{4.5}$$

$$\dot{v}_i = \frac{1}{N}\sum_{j=1}^{N} g(\|x_i - x_j\|_{\ell_2^d})(v_j - v_i), \quad i = 1, \dots, N. \tag{4.6}$$

The function $g : [0, \infty) \to \mathbb{R}$ is given by $g(s) = \frac{G}{(1+s^2)^\beta}$, for $\beta > 0$, and bounded by $g(0) = G > 0$. This model describes the *emerging of consensus* in a group of interacting agents, trying to *align* (also in terms of abstract consensus) with their neighbors. One of the motivations of the model from Cucker and Smale was to describe the formation and evolution of languages [31, Section 6], although, due to its simplicity, it has been eventually related mainly to the description of the *emergence of flocking* in groups of birds [30]. In the latter case, in fact, spatial and velocity coordinates are sufficient to describe a pointlike agent ($d = 3 + 3$), while for the evolution of languages, one would have to take into account a much broader dictionary of parameters, hence a higher dimension $d \gg 3 + 3$ of parameters, which is in fact was the case of our interest in [P10].

We investigated dynamical systems of the type

$$\dot{x}_i(t) = f_i(\mathcal{D}x(t)) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}x(t))x_j(t), \tag{4.7}$$

where we use the following notation:

- $N \in \mathbb{N}$ - number of agents,
- $x(t) = (x_1(t), \dots, x_N(t)) \in \mathbb{R}^{d\times N}$, where $x_i : [0, T] \to \mathbb{R}^d$, $i = 1, \dots, N$,
- $f_i : \mathbb{R}^{N\times N} \to \mathbb{R}^d, \quad i = 1, \dots, N,$
- $f_{ij} : \mathbb{R}^{N\times N} \to \mathbb{R}, \quad i, j = 1, \dots, N,$
- $\mathcal{D} : \mathbb{R}^{d\times N} \to \mathbb{R}^{N\times N}$, $\mathcal{D}x := (\|x_i - x_j\|_{\ell_2^d})_{i,j=1}^N$ is the *adjacency matrix* of the point cloud x.

We assumed that the governing functions f_i and f_{ij} are Lipschitz. The system (4.7) describes the dynamics of multiple complex agents $x(t) = (x_1(t), \dots, x_N(t)) \in \mathbb{R}^{d\times N}$, interacting on the basis of their mutual "social" distance $\mathcal{D}x(t)$, and its general form includes several models for swarming and collective motion of animals and micro-organisms, aggregation of cells, etc. Several relevant effects can be included in the model by means of the functions f_i and f_{ij}, in particular, fundamental binary mechanisms of *attraction, repulsion, aggregation* and *alignment* [22, 30, 31, 86, 61].

In [P10] we applied the following strategy for dimensionality reduction of such dynamical systems. To decide if some effects occurred during the evolution of the dynamical system, it is often not necessary to know the full trajectory of the system. For the Cucker-Smale system we might be interested, if flocking occurred or not - but this can be very well guessed also from any lowdimensional projection of the system. We therefore first apply Johnson-Lindenstrauss embedding of the initial data and then calculate the solution path in the lower dimension. It turns out that (at least for small period of time) the result of this lies close to the projection of the solution of the original (highdimensional) dynamical system.

5 Ridge functions

It is very well known, cf. [83], that approximation of smooth functions is (at least in some settings) intractable in high dimensions. Therefore, the aim of the next group of papers was to study approximation of well structured multivariate functions, which take a form of a ridge, i.e.

$$f(x) = g(a \cdot x), \quad x \in \mathbb{R}^d, \quad x \in \Omega. \tag{5.1}$$

Here, one assumes that both the *ridge vector* $a \in \mathbb{R}^d$ and the univariate function g (sometimes also called *ridge profile*) are unknown. Although the formula (5.1) is rather simple, it revealed couple of features:

(i) Typical structural assumptions posed on multivariate functions are linear (i.e. the function belongs to some Banach space, which is of course linear). In contrary, (5.1) is non-linear and may serve as a prototype of non-linear function classes useful for multivariate problems.

(ii) Although the formula (5.1) is rather simple, the tractability of the approximation of ridge functions runs through several of the tractability classes considered in the field of *Information Based Complexity*, cf. [84, 85], depending on the assumptions made on a and g (and on the domain Ω).

(iii) For certain assumptions on a, the theory of compressed sensing comes in as an useful tool.

The results reported in this section were based on [27, 39, 119] and were published in the following papers.

[P11] M. Fornasier, K. Schnass, and J. Vybíral, Learning functions of few arbitrary linear parameters in high dimensions, Found. Comput. Math. 12 (2) (2012), 229–262

[P12] A. Kolleck and J. Vybíral, On some aspects of approximation of ridge functions, J. Appr. Theory 194 (2015), 35–61

[P13] S. Mayer, T. Ullrich, and J. Vybíral, Entropy and sampling numbers of classes of ridge functions, Constr. Appr. 42 (2) (2015), 231–264

5.1 Learning functions of few arbitrary linear parameters in high dimensions

Paper [P11] exploited the straightforward formula

$$\frac{\partial f}{\partial \varphi}(\xi) = [g'(a \cdot \xi)]a \cdot \varphi \tag{5.2}$$

to get the access to scalar products of a with carefully chosen directional vectors φ. Furthermore, replacing derivatives with first-order differences allowed for a sampling algorithm based on randomly chosen sampling points and polynomial or even logarithmic complexity in the dimension d.

To be more precise we define two sets $\mathcal{X}, \Phi$ of points. The first

$$\mathcal{X} = \{\xi_j \in \mathbb{S}^{d-1} : j = 1, \dots, m_{\mathcal{X}}\}, \tag{5.3}$$

contains the $m_{\mathcal{X}}$ sampling points and is drawn at random in $\mathbb{S}^{d-1}$ according to the probability measure $\mu_{\mathbb{S}^{d-1}}$. For the second, containing the m_Φ derivative directions, we have

$$\Phi \quad = \quad \Big\{ \varphi_i \in B_{\mathbb{R}^d}(\sqrt{d}/\sqrt{m_\Phi}) : \varphi_{i\ell} = \frac{1}{\sqrt{m_\Phi}} \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2, \end{cases} \\ i = 1, \dots, m_\Phi, \text{ and } \ell = 1, \dots, d \Big\}. \tag{5.4}$$

Actually we identify Φ with the $m_\Phi \times d$ matrix whose rows are the vectors φ_i. To write the $m_{\mathcal{X}} \times m_\Phi$ instances of (5.2) in a concise way we collect the directional derivatives $g'(a \cdot \xi_j)a$, $j = 1, \dots, m_{\mathcal{X}}$ as columns in the $d \times m_{\mathcal{X}}$ matrix X, i.e.,

$$X = (g'(a \cdot \xi_1)a^T, \dots, g'(a \cdot \xi_{m_{\mathcal{X}}})a^T), \tag{5.5}$$

and we define the $m_\Phi \times m_{\mathcal{X}}$ matrices Y and $\mathcal{E}$ entrywise by

$$y_{ij} = \frac{f(\xi_j + \epsilon\varphi_i) - f(\xi_j)}{\epsilon}, \tag{5.6}$$

and

$$\varepsilon_{ij} = \frac{\epsilon}{2}[\varphi_i^T \nabla^2 f(\zeta_{ij})\varphi_i]. \tag{5.7}$$

We denote by y_j the columns of Y and by ε_j the columns of $\mathcal{E}$, $j = 1, \dots, m_{\mathcal{X}}$. With these matrices we can write the following factorization

$$\Phi X = Y - \mathcal{E}. \tag{5.8}$$

Under the additional assumptions that $a \in \mathbb{R}^d$ is sparse, (5.8) may be interpreted as compressive measurements of a with noise, and it is therefore possible to use the methods of sparse recovery to approximate a. We therefore proposed the following algorithm.

Algorithm:

- *Given $m_\Phi, m_{\mathcal{X}}$, draw at random the sets Φ and $\mathcal{X}$ as in* (5.3) *and* (5.4)*, and construct Y according to* (5.6)*.*
- *Set $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j = \Phi z} \|z\|_{\ell_1^d}$.*
- *Find*
$$j_0 = \arg \max_{j=1,\dots,m_{\mathcal{X}}} \|\hat{x}_j\|_{\ell_2^d}. \tag{5.9}$$
- *Set $\hat{a} = \hat{x}_{j_0}/\|\hat{x}_{j_0}\|_{\ell_2^d}$.*
- *Define $\hat{g}(y) := f(\hat{a}^T y)$ and $\hat{f}(x) := \hat{g}(\hat{a} \cdot x)$.*

Using recent Chernoff bounds for sums of positive-semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions, we were able to provide (probabilistic) guarantees on the performance of this algorithm in approximating ridge function (5.1). Furthermore, the general case $f(x) = g(Ax)$, where $A \in \mathbb{R}^{k \times d}$ is a matrix, was also considered.

5.2 On some aspects of approximation of ridge functions

In [P12] we addressed several issues of analysis of ridge functions, which were left open in the previous works. The first aspect was the change of the domain from unit ball to unit cube, i.e.

we considered functions

$$f(x) = g(\langle a, x\rangle), \quad x \in [-1,1]^d.$$

As the unit cube is much larger than the unit ball (a fact which is described in many ways in the analysis of convex bodies) it is usually much more difficult to approximate a function on a unit cube than on a unit ball. With the non-linear class of ridge functions the situation is different - the larger domain can be used to learn the ridge direction a more accurately. The crucial notion of our analysis was the sign of a vector $\mathrm{sign}(x)$, which is taken componentwise. Although this mapping is not continuous, its scalar product with the vector x itself not only gives the ℓ_1-norm of the original vector, but the mapping $y \to \langle y, \mathrm{sign}(x)\rangle$ becomes continuous at x.

The second issue discussed in [P12] was the subject of noisy sampling. As the methods used so far were based on first order differences, their stability was an important question. We proposed an algorithm, which involves the *Dantzig selector* of Candés and Tao [21]. This recovery algorithm can deal with random noise much more effectively than the classical ℓ_1-norm minimization. Especially, the effect of *noise folding* is completely avoided with this approach. As intuitively expected, the distance parameter of the first order differences has to be optimized - if it is too small, any small perturbation of the function values affects heavily the differences. If it is too large, the first order differences do not approximate the first derivatives well any more.

Finally, we considered the class of shifted radial functions $f(x) = g(\|a-x\|_2^2)$. It turned out that the approach developed so far can easily be translated to this setting.

5.3 Entropy and sampling numbers of classes of ridge functions

The paper [P13] discussed the approximation of ridge functions from the point of view of *Information Based Complexity*, paying attention to optimality of the known algorithms and to lower bounds on the error of approximation. We considered ridge functions defined on the unit ball

$$\Omega = \bar{B}_2^d = \{x \in \mathbb{R}^d \ : \ \|x\|_2 \le 1\}.$$

Let $\alpha > 0$ denote the order of Lipschitz smoothness. Further, let $0 < p \le 2$. We define the class of ridge functions with Lipschitz profiles as

$$\mathcal{R}_d^{\alpha,p} = \left\{f : \Omega \to \mathbb{R} \ : \ f(x) = g(a\cdot x), \ \|g\|_{\mathrm{Lip}_\alpha[-1,1]} \le 1, \ \|a\|_p \le 1\right\}. \tag{5.10}$$

In addition, we define the class of ridge functions with infinitely differentiable profiles by

$$\mathcal{R}_d^{\infty,p} = \left\{f : \Omega \to \mathbb{R} \ : \ f(x) = g(a\cdot x), \ \|g\|_{C^\infty[-1,1]} \le 1, \ \|a\|_p \le 1\right\}.$$

The concept of entropy numbers is central to this work. They can be understood as a measure to quantify the compactness of a set w.r.t. some reference space. For a detailed exposure and historical remarks, we refer to the monographs [23, 45]. The k-th entropy number $e_k(K,X)$ of a subset K of a (quasi-)Banach space X is defined as

$$e_k(K,X) = \inf\Big\{\varepsilon > 0 : K \subset \bigcup_{j=1}^{2^{k-1}} (x_j + \varepsilon \bar{B}_X) \text{ for some } x_1,\dots,x_{2^{k-1}} \in X\Big\}. \tag{5.11}$$

Note that $e_k(K,X) = \inf\{\varepsilon > 0 : N_\varepsilon(K,X) \le 2^{k-1}\}$ holds true, where

$$N_\varepsilon(K,X) := \min\Big\{n \in \mathbb{N} : \quad \exists x_1,\dots,x_n \in X : \ K \subset \bigcup_{j=1}^{n} (x_j + \varepsilon \bar{B}_X)\Big\} \tag{5.12}$$

denotes the *covering number* of the set K in the space X, which is the minimal natural number n such that there is an ε-net of K in X of n elements. We can introduce entropy numbers for operators, as well. The k-th entropy number $e_k(T)$ of an operator $T : X \to Y$ between two quasi-Banach spaces X and Y is defined by

$$e_k(T) = e_k(T(\bar{B}_X), Y). \tag{5.13}$$

The main result on entropy numbers of classes of ridge functions obtained in [P13] was the following theorem.

Theorem 5.1. *Let d be a natural number and $\alpha > 0$. For the entropy numbers of $\mathcal{R}_d^{\alpha,2}$ in $L_\infty(\Omega)$ we have*

$$\max(k^{-\alpha}, 2^{-k/d}) \lesssim e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \lesssim \begin{cases} 1 & : k \le c_\alpha d \log d, \\ k^{-\alpha} & : k \ge c_\alpha d \log d\,, \end{cases} \tag{5.14}$$

for some universal constant $c_\alpha > 0$ which does not depend on d.

As the decay of these entropy numbers resembles very much the behaviour of the entropy numbers of univariate Lipschitz functions, we can conclude that, when speaking in terms of entropy, classes of ridge functions with Lipschitz profile are essentially as compact as the class of univariate Lipschitz functions. Consequently, these classes must be much smaller than the class of multivariate Lipschitz functions.

The situation changes dramatically, when we come from entropy numbers to the so-called *sampling numbers.* These numbers describe the minimal worst-case error when approximating functions from a certain class using only a limited budget of function values, which we are allowed to take. It turned out that without any additional assumptions on g and a, the problem is intractable. Interestingly, when changing the assumptions on a and g, the problem belongs to a number of different tractability classes considered in Information Based Complexity. Assuming, on the other hand, that $|g'(0)| \ge \varkappa > 0$ allows to use the techniques of compressed sensing and restore tractability.

6 Applications in machine learning

[P14] A. Kolleck and J. Vybíral, Non-asymptotic analysis of ℓ_1-Support Vector Machines, submitted

[P15] L. M. Ghiringhelli, J. Vybíral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big data of materials science - Critical role of the descriptor, Phys. Rev. Lett. 114, 105503 (2015)

6.1 Non-asymptotic analysis of ℓ_1-Support Vector Machines

Support vector machines (SVM) are a group of popular classification methods in machine learning. Their input is a set of data points $x_1, \dots, x_m \in \mathbb{R}^d$, each equipped with a label $y_i \in \{-1, +1\}$, which assigns each of the data points to one of two groups. SVM aims for binary linear classification based on separating hyperplane between the two groups of training data, choosing a hyperplane with separating gap as large as possible.

Since their introduction by Vapnik and Chervonenkis [115], the subject of SVM was studied intensively. We will concentrate on the so-called soft margin SVM [29], which allow also for misclassification of the training data and are the most used version of SVM nowadays.

In its most common form (and neglecting the bias term), the soft-margin SVM is a convex optimization program

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi\in\mathbb{R}^m}} \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^m \xi_i \quad \text{subject to} \quad y_i\langle x_i, w\rangle \geq 1-\xi_i$$

$$\text{and} \quad \xi_i \geq 0 \tag{6.1}$$

for some tradeoff parameter $\lambda > 0$ and so called slack variables ξ_i. It will be more convenient for us to work with the following equivalent reformulation of (6.1)

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^m [1-y_i\langle x_i, w\rangle]_+ \quad \text{subject to} \quad \|w\|_2 \leq R, \tag{6.2}$$

where $R > 0$ gives the restriction on the size of w.

The aim of [P14] was to analyze the ℓ_1-based variant of SVM, which was introduced in [121] and which performs well when looking for sparse classifiers, i.e. when $w \in \mathbb{R}^d$ is supposed to have only few non-zero coordinates. Hence, we denote by $\hat{a}$ the minimizer of

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^m [1-y_i\langle x_i, w\rangle]_+ \quad \text{subject to} \quad \|w\|_1 \leq R. \tag{6.3}$$

The setting of our work, which we will later on refer to as "Standing assumptions", was the following.

Standing assumptions:

(i) $a \in \mathbb{R}^d$ is the true (nearly) sparse classifier with $\|a\|_2 = 1$, $\|a\|_1 \leq R$, $R \geq 1$, which we want to approximate;

(ii) $x_i = r\tilde{x}_i, \quad \tilde{x}_i \sim \mathcal{N}(0, \mathrm{Id}), i = 1, \dots, m$ are i.i.d. training data points for some constant $r > 0$;

(iii) $y_i = \operatorname{sgn}(\langle x_i, a\rangle), \quad i = 1, \dots, m$ are the labels of the data points;

(iv) $\hat{a}$ is the minimizer of (6.3);

(v) Furthermore, we denote

$$K = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R\}, \tag{6.4}$$

$$f_a(w) = \frac{1}{m}\sum_{i=1}^m [1-y_i\langle x_i, w\rangle]_+, \tag{6.5}$$

where the subindex a denotes the dependency of f_a on a (via y_i).

Using the methods of concentration of measure and of probability theory in Banach spaces [67, 68], we could estimate the performance of (6.3) under the "Standing assumptions".

Theorem 6.1. *Let $d \geq 2$, $0 < \varepsilon < 0.18$, $r > \sqrt{2\pi}(0.57-\pi\varepsilon)^{-1}$ and $m \geq C\varepsilon^{-2}r^2R^2\log(d)$ for some constant C. Under the "Standing assumptions" it holds*

$$\frac{\left\|a - \frac{\hat{a}}{\|\hat{a}\|_2}\right\|_2}{\langle a, \frac{\hat{a}}{\|\hat{a}\|_2}\rangle} \leq C'\left(\varepsilon + \frac{1}{r}\right) \tag{6.6}$$

with probability at least

$$1 - \gamma \exp\big(-C'' \log(d)\big) \tag{6.7}$$

for some positive constants γ, C', C''.

If $a \in \mathbb{R}^d$ is s-sparse, then (simply by Hölder's inequality) $\|a\|_1 \leq \sqrt{s}$ and we may take $R = \sqrt{s}$ in Theorem 6.1. The logarithmic dependence of m on d and the linear dependence of m on s are the main achievements of Theorem 6.1 and explain the practical success of ℓ_1-SVM in many different areas. On the other hand, we conjecture that the dependence of m on ε and r is *not* optimal and could be improved by more detailed analysis.

6.2 Big data of materials science - Critical role of the descriptor

The last paper selected for this cumulative thesis arose from the collaboration with colleagues from Fritz-Haber Institute in Berlin. They have been interested in speeding up the discovery of new materials. Nowadays, important material properties may be calculated *ab initio* from the known molecular structure of the material. Essentially, the only inputs of these calculations are the nuclear numbers of the atoms in the molecule. Nevertheless, any such calculation takes quite long amount of time. As the number of potential new materials is in thousands (and hundreds of thousands), it is not feasible to calculate all of them through.

Instead of that, we would be interested in a very quick (but inaccurate) calculation of such properties, which could (at least roughly) predict, were the interesting materials are to be found. Afterwards, these preselected materials could indeed be treated by the full scale computation.

As a model example we have chosen the prediction of the crystal structure of binary compound semiconductors, which are known to crystallize in zincblende (ZB), wurtzite (WZ), or rocksalt (RS) structures. In 1970 Phillips and van Vechten (Ph-vV) [116, 90] analyzed the prediction or classification challenge and came up with a two-dimensional descriptor, i.e., two numbers that are related to the dielectric constant and the nearest-neighbor distance in the crystal [116, 90]. Figure 2 shows their conclusion. Clearly, in this representation ZB/WZ and RS structures separate nicely: Materials in the upper left part crystallize in the RS structure, those in the lower right part are ZB/WZ. Thus, based on the ingenious descriptor $\boldsymbol{d} = (E_h, C)$ one can predict the structure of unknown compounds without the need of performing explicit experiments or calculations. Several authors have taken up the Ph-vV challenge and identified alternative descriptors [122, 89, 24].

We have therefore selected $N = 82$ binary compounds and calculated the property P - the difference in LDA energy (ΔE) between RS and ZB for the given atom pair AB. Then we were searching for a descriptor that minimizes the Root Mean Square Error (RMSE), given by $\sqrt{(1/N)\|\boldsymbol{P} - \boldsymbol{D}\boldsymbol{c}\|_2^2}$. The order is such that element A is the one with the smallest electronegativity EN, defined according to Mulliken: EN = 1/2 (IP+EA). IP and EA are atomic ionization potential and electron affinity evaluated as the energy of the half-occupied Kohn-Sham orbital in the half positively and half negatively charged LDA atom, respectively. For systematically constructing the feature space, i.e., the candidate components of the descriptor, and then selecting the most relevant of them, we implement an ***iterative*** approach. We start from 7 atomic features for atom A: IP(A) and EA(A), H(A) and L(A), the energies of the highest-occupied and lowest-unoccupied Kohn-Sham (KS) levels, as well as r_s(A), r_p(A), and r_d(A), i.e., the radius where the radial probability density of the valence s, p, and d orbital is maximal. Besides, information regarding the isolated AA, BB, and AB dimers was added to the list, namely their equilibrium distance, binding energy, and HOMO-LUMO KS gap (a total of 9 more features).

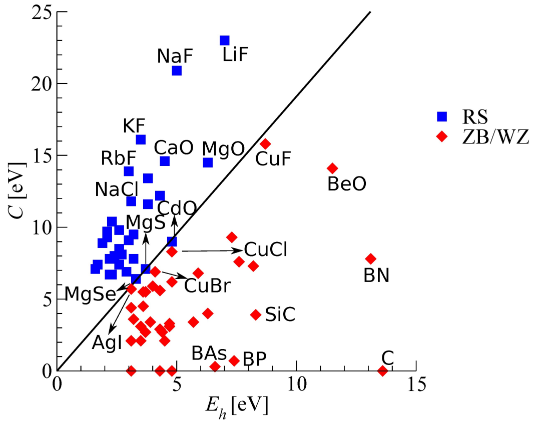


Figure 2: Ground-state structures of 68 octet binary compounds, arranged according to the two-dimensional descriptor introduced by Phillips and van Vechten [116, 90]. Both descriptors and classification derive from experimental data. Because of visibility reasons only 10 materials are labeled for each structure.

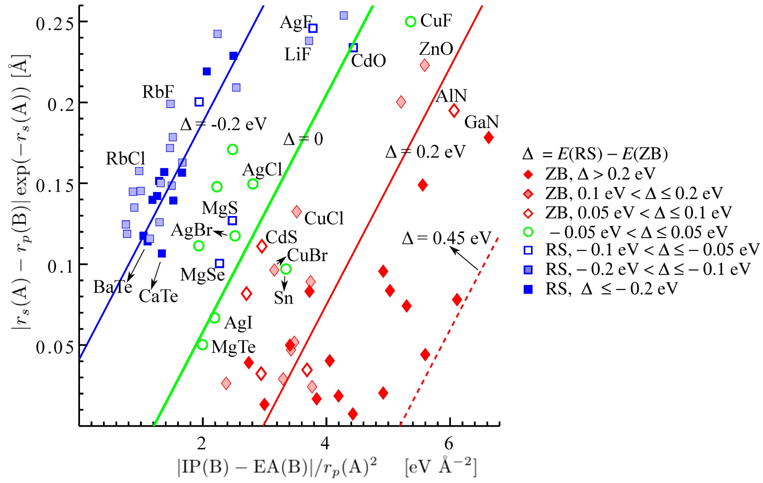


Figure 3: Calculated energy differences of the 82 octet binary materials, arranged according to our optimal two-dimensional descriptor. For visibility reasons, not all materials are labeled. Seven ZB materials with predicted $\Delta E > 0.5$ eV are outside the shown window.

Next, we define rules for linear and non-linear combinations of the just mentioned 23 starting features. One can easily generate a huge number of candidate descriptors, e.g., all thinkable but not violating basic physical rules. In the present study we used about 10 000 candidates subdivided such to be used in different iterations, where we refined the feature space.

We form (non-)linear combinations of the starting features, which we expect to be potentially of some causal significance. In the language of kernel ridge regression we design a kernel and we do it by using physical insight. In this way we can check new mechanisms that are *tested* one against each other. Due to the limited set of data points, the list cannot be exhaustive because LASSO (and actually any other method) has difficulties in selecting between two highly correlated features. In our case, for instance, r_s and r_p for the same atom have a large correlation (Pearson's index larger than 0.95, in other words the two 82-dimensional vectors of the feature r_s and r_p are almost collinear).

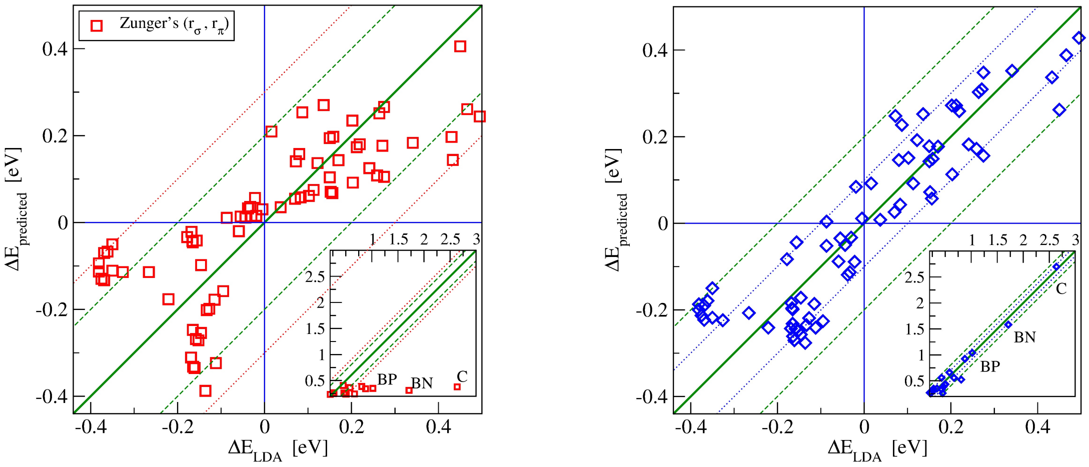


Figure 4: Error of a linear fit for Zunger's descriptors (left figure) and for our best pair (right figure). Each symbol represents one material, which was left out from training and afterwards forecasted by the description found. Especially materials with high ΔE_{LDA} are predicted by our method with much higer accuracy (see the right-bottom zoom of the figures).

Our procedure identifies as best (i.e., yielding the lowest RMSE) one-, two-, and three-dimensional (1D, 2D, and 3D) descriptors. These are the first, the first two, and all three of the following features:

$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2}, \frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}, \frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}) + r_s(\mathrm{B}))}.$$

The extensions of this method to problems closer to real-life questions is currently the subject of further research.

7 Bibliography

[1] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, J. Comput. System Sci. **66**, 671–687 (2003)

[2] R. A. Adams, *Sobolev spaces*, Pure and Applied Mathematics, Vol. 65, Academic Press, New York-London, 1975.

[3] D. R. Adams and L. I. Hedberg, *Function Spaces and Potential Theory*, Springer, Berlin, 1996.

[4] N. Ailon and B. Chazelle, *The fast Johnson-Lindenstrauss transform and approximate nearest neighbors*, SIAM J. Comput. 39 (1):302–322, 2009.

[5] A. Almeida, P. Hästö: *Besov spaces with variable smoothness and integrability*, J. Funct. Anal. **258** (2010), no. 5, 1628–1655.

[6] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*, Cambridge Univ. Press, Cambridge (2009)

[7] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constr. Approx. **28**, 253–263 (2008)

[8] F. Barthe, M. Csörnyei and A. Naor, *A note on simultaneous polar and Cartesian decomposition*, in: Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics, Springer, Berlin, 2003.

[9] F. Barthe, O. Guédon, S. Mendelson and A. Naor, *A probabilistic approach to the geometry of the l_p^n-ball*, Ann. Probab. 33 (2005), no. 2, 480–513.

[10] D. B. Bazarkhanov, *Spaces of functions of variable mixed smoothness I*, Mat. Zh. 6 (2006), no. 4(22), 32–39.

[11] D. B. Bazarkhanov, *Spaces of functions of variable mixed smoothness II*, Mat. Zh. 7 (2007), no. 3(25), 16–27.

[12] C. Bennett and R. Sharpley, *Interpolation of operators*, Academic Press, San Diego, 1988.

[13] J. Bergh and J. Löfström, *Interpolation spaces. An Introduction*, Springer Verlag, 1976.

[14] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, A Survey of Compressed Sensing First chapter in Compressed Sensing and its Applications, Birkhäuser, Springer, 2015

[15] J. Bourgain, S. Dirksen, J. Nelson, *Toward a unified theory of sparse dimensionality reduction in Euclidean space*, Geometric and Functional Analysis 25 (2015), no. 4, 1009-1088

[16] E.J. Candés, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Serie I, **346**, 589–592 (2008)

[17] E.J. Candés, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory **52**, 489–509 (2006)

[18] E.J. Candés and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51**, 4203–4215 (2005)

[19] E.J. Candés, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math. **59**, 1207–1223 (2006)

[20] E.J. Candés and T. Tao, *Near-optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory **52**, 5406–5425 (2006)

[21] E.J. Candés and T. Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35**, 2313–2351 (2007)

[22] J. A. Carrillo, M. Fornasier, G. Toscani, and F. Vecil, *Particle, kinetic, hydrodynamic models of swarming*, in: Mathematical modeling of collective behavior in socio-economic and life-sciences, Birkhäuser, 2010.

[23] B. Carl and I. Stephani, *Entropy, compactness and the approximation of operators*, Cambridge Tracts in Math. **98**, Cambridge Univ. Press, Cambridge, 1990.

[24] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104, 2012

[25] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. **20**, 33–61 (1998)

[26] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22**, 211–231 (2009)

[27] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard, *Capturing ridge functions in high dimensions from point queries*, Constr. Approx. **35**, 225–243 (2012)

[28] R. R. Coifman, *A real variable characterization of H^p*, Studia Math. 51 (1974), 269–274.

[29] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20, no.3, pp. 273–297, 1995.

[30] F. Cucker and S. Smale, *Emergent behavior in flocks*, IEEE Trans. Automat. Control, 52 (2007), pp 852–862.

[31] F. Cucker and S. Smale, *On the mathematics of emergence*, Japan J. Math., 2 (2007), 197–227.

[32] S. Dahlke, E. Novak and W. Sickel, *Optimal approximation of elliptic problems by linear and nonlinear mappings I*, J. Complexity 22 (2006), no. 1, 29–49.

[33] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms **22**, 60–65 (2003)

[34] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok, *Introduction to compressed sensing*, Compressed sensing, 1–64, Cambridge Univ. Press, Cambridge, (2012)

[35] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math. **41** (1988), 909-996.

[36] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.

[37] R. A. DeVore, *Nonlinear approximation*, Acta Num. 51–150, (1998).

[38] R. A. DeVore, B. Jawerth and V. Popov, *Compression of wavelet decompositions*, Amer. J. Math. 114 (1992), no. 4, 737–785.

[39] R. DeVore, G. Petrova, and P. Wojtaszczyk, *Approximation of functions of few variables in high dimensions*, Constr. Approx. **33**, 125–143 (2011)

[40] L. Diening, P. Hästö and S. Roudenko, *Function spaces of variable smoothness and integrability*, J. Funct. Anal. 256 (2009), no. 6, 1731–1768.

[41] D.L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52**, 1289–1306 (2006)

[42] M. Duarte, M. Davenport, D. Takhar, J. Laska, S. Ting, K. Kelly, and R. Baraniuk, *Single-pixel imaging via compressive sampling*, IEEE Signal Process. Mag. **25**, 83–91 (2008)

[43] D. Dung, *Approximation of functions of several variables on a torus by trigonometric polynomials*, Mat. Sb. (N.S.) 131(173) (1986), no. 2, 251–271; translated in Math. USSR-Sb. 59 (1988), no. 1, 247–267.

[44] D. Dung, *Optimal non-linear approximation of functions with a mixed smoothness*, East J. Approx. 4 (1998), no. 1, 75–86.

[45] D.E. Edmunds and H. Triebel. Function Spaces, Entropy Numbers, Differential Operators. Cambridge Tracts in Mathematics, vol. 120, Cambridge University Press, Cambridge, 1996.

[46] C. Fefferman and E. M. Stein, *H^p spaces of several variables*, Acta Math. 129 (1972), no. 3-4, 137-193.

[47] T. Figiel, J. Lindenstrauss and V. D. Milman, *The dimension of almost spherical sections of convex bodies*, Acta Math. 139 (1977), no. 1-2, 53–94.

[48] M. Fornasier and H. Rauhut, *Compressive Sensing*, In: Scherzer, Otmar (Ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, Heidelberg (2011)

[49] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Birkhäuser/Springer, New York (2013)

[50] J. Franke, *On the spaces F^s_{pq} of Triebel-Lizorkin type: pointwise multipliers and spaces on domains*, Math. Nachr. 125 (1986), 29–68.

[51] M. Frazier and B. Jawerth, *Decomposition of Besov spaces*, Indiana Univ. Math. J. 34 (1985), 777–799.

[52] M. Frazier and B. Jawerth, *A discrete transform and decomposition of distribution spaces*, J. Funct. Anal. 93 (1990), 34–170.

[53] K. Friedrichs, *Die Rand- und Eigenwertprobleme aus der Theorie der elastischen Platten*, Math. Ann. 98 (1928), 205–247.

[54] B. Gärtner and J. Matoušek, *Understanding and Using Linear Programming*, Springer, Berlin (2006)

[55] M. Hansen and J. Vybíral, *The Jawerth-Franke embedding of spaces with dominating mixed smoothness*, Georg. Math. J. 16 (2009), No. 4, 667–682.

[56] D.D. Haroske and L. Skrzypczak, *On Sobolev and Franke-Jawerth embeddings of smoothness Morrey spaces*, Rev. matem. complutense 27, no. 2 (2014): 541–573.

[57] P. Indyk and R. Motwani, *Approximate nearest neighbors: Towards removing the curse of dimensionality*, In *Proc. 30th Annual ACM Symposium on Theory of Computing*, pp. 604–613, 1998.

[58] P. Indyk and A. Naor, *Nearest neighbor preserving embeddings*, ACM Trans. Algorithms, 3(3), Article no. 31, 2007.

[59] B. Jawerth, *Some observations on Besov and Lizorkin-Triebel spaces*, Math. Scand. 40 (1977), 94–104.

[60] W.B. Johnson, J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, In: Conf. in Modern Analysis and Probability, pp. 189–206, (1984)

[61] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol. 26 (1970), pp. 399–415.

[62] H. Kempka, *2-microlocal Besov and Triebel-Lizorkin spaces of variable integrability*, Rev. Mat. Complut. **22** (2009), no. 1, 227–251.

[63] O. Kováčik and J. Rákosník, *On spaces $L^{p(x)}$ and $W^{1,p(x)}$*, Czechoslovak Math. J. 41 (116) (1991), 592–618.

[64] F. Krahmer and R. Ward, *New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property*, SIAM J. Math. Anal. **43**, 1269–1281 (2011)

[65] A. Kufner, O. John and S. Fučík, *Function spaces*, Academia, Prague, 1977.

[66] R. H. Latter, *A characterization of $H^p(\mathbf{R}^n)$ in terms of atoms*, Studia Math. 62 (1978), no. 1, 93–101.

[67] M. Ledoux, *The concentration of measure phenomenon*, American Mathematical Society, Providence, (2001)

[68] M. Ledoux and M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes*, Springer, Berlin, (1991)

[69] R. Linde, *s-Numbers of diagonal operators and Besov embeddings*, Proc. 13-th Winter School, Suppl. Rend. Circ. Mat. Palermo (1986).

[70] G. G. Lorentz, M. v. Golitschek and Y. Makovoz, *Constructive approximation. Advanced problems*, Grundlehren der Mathematischen Wissenschaften, 304. Springer-Verlag, Berlin, 1996.

[71] C. Lubitz, *Weylzahlen von Diagonaloperatoren und Sobolev-Einbettungen*, Dissertation, Rheinische Friedrich-Wilhelms-Universität, Bonn, 1982.

[72] M. Lustig, D. Donoho, J.M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magn. Reson. Med. **58**, 1182–1195 (2007)

[73] V. E. Maĭorov, *Discretization of the problem of diameters*, Uspekhi Mat. Nauk **30**, No. 6 (1975), 179–180.

[74] J. Matoušek, *On variants of the Johnson-Lindenstrauss lemma*, Random Structures Algorithms **33** 142–156 (2008)

[75] V. G. Maz'ya, *Sobolev Spaces*, Springer, Berlin, 1985.

[76] V. G. Maz'ya, *Sobolev Spaces: With Applications to Elliptic Partial Differential Equations*, Grundlehren der mathematischen Wissenschaften, Vol. 342, Springer, 2011.

[77] Y. Meyer, *Wavelets and operators*, Cambridge Univ. Press, 1992.

[78] V. D. Milman, *A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies*, Funkcional. Anal. i Priložen. 5 (1971), no. 4, 28–37.

[79] V.D. Milman and G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*, Springer, Berlin (1986)

[80] M. Mishali and Y. Eldar, *From theory to practice: Sub-nyquist sampling of sparse wideband analog signals*, IEEE J. Sel. Top. Signal Process. **4**, 375–391 (2010)

[81] A. Naor, *The surface measure and cone measure on the sphere of ℓ_p^n*, Trans. Amer. Math. Soc. 359 (2007), no. 3, 1045–1079.

[82] A. Naor and D. Romik, *Projecting the surface measure of the sphere of ℓ_p^n*, Ann. Inst. H. Poincaré Probab. Statist. 39 (2003), no. 2, 241–261.

[83] E. Novak and H. Woźniakowski, *Approximation of infinitely differentiable multivariate functions is intractable*, Journal of Complexity **25** (2009), 398–404.

[84] E. Novak and H. Woźniakowski, *Tractability of multivariate problems. Vol. 1: Linear information*, EMS Tracts in Mathematics, 6, European Mathematical Society (EMS), Zürich, 2008.

[85] E. Novak and H. Woźniakowski, *Tractability of multivariate problems, Volume II: Standard information for functionals*, EMS Tracts in Mathematics, 12, European Mathematical Society (EMS), Zürich, 2010.

[86] M. R. D'Orsogna, Y. L. Chuang, A. L. Bertozzi, and L. Chayes, *Self-propelled particles with soft-core interactions: patterns, stability, and collapse*, Phys. Rev. Lett. 96 (2006).

[87] K. Oskolkov, *Polygonal approximation of functions of two variables*, Math. USSR Sbornik 35, 851–861, (1979).

[88] J. Peetre, *New thoughts on Besov spaces*, Duke Univ. Math. Series, Durham, 1976.

[89] D. G. Pettifor, Solid State Commun. 51, 1984

[90] J. C. Phillips, Rev. Mod. Phys. 42, 1970.

[91] A. Pietsch, *Einige neue Klassen von kompakten linearen Operatoren*, Rev. Math. Pures Appl. **8** (1963), 427–447.

[92] A. Pietsch, *Eigenvalues and s-numbers*, Cambridge University Press, 1987, Cambridge.

[93] A. Pinkus, *n-widths in approximation theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete 3.7, Springer, 1985.

[94] F. Rellich, *Ein Satz über mittlere Konvergenz*, Math. Nachr. 31 (1930), 30–35.

[95] V. S. Rychkov, *On a theorem of Bui, Paluszyński and Taibleson*, Steklov Institute of Mathematics **227**, (1999), 280–292.

[96] V. S. Rychkov, *On restrictions and extensions of the Besov and Triebel-Lizorkin spaces with respect to Lipschitz domains*, J. London Math. Soc. (2) **60** (1999), 237–257.

[97] E. Schmidt, *Zur Theorie der linearen und nichtlinearen Integralgleichungen I*, Math. Anal. 63, 433–476, (1907).

[98] C. Schneider, *Spaces of Sobolev type with positive smoothness on $\mathbb{R}^n$, embeddings and growth envelopes*, J. Funct. Spaces 7, no. 3 (2009): 251–288.

[99] S. L. Sobolev, *On some estimates relating to families of functions having derivatives that are square integrable*, Dokl. Akad. Nauk SSSR 1 (1936), 267–270 (in Russian).

[100] S. L. Sobolev, *On theorem in functional analysis*, Sb. Math. 4 (1938), 471–497 (in Russian); English translation: Am. Math. Soc. Trans. 34 (1963), no. 2, 39–68.

[101] S. L. Sobolev, *Applications of Functional Analysis in Mathematical Physics*, Izd. LGU im. A. A. Ždanova, Leningrad, 1950 (in Russian); English translation: Am. Math. Soc. Trans. 7 (1963).

[102] E. M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Univ. Press, Princeton, 1970.

[103] V. N. Temlyakov, *Approximation of periodic functions*, Nova Science, New York, 1993.

[104] V. N. Temlyakov, *Nonlinear methods of approximation*, Found. Comput. Math. 3 (2003), no. 1, 33—107.

[105] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Royal Statist. Soc B **58**, 267–288 (1996)

[106] V. M. Tikhomirov, *Diameters of sets in function spaces and the theory of best approximations*, Uspekhi Mat. Nauk **15**, No. 3 (1960), 81–120; translated in Russ. Math. Survey **15**, No. 3 (1960), 75–111.

[107] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1978.

[108] H. Triebel, *Theory of function spaces*, Birkhäuser, Basel, 1983.

[109] H. Triebel, *Theory of function spaces II*, Birkhäuser, Basel, 1992.

[110] H. Triebel, *Non-smooth atoms and pointwise multipliers in function spaces*, Ann. Mat. Pura Appl. (4) 182 (2003), no. 4, 457–486.

[111] H. Triebel, *Local means and wavelets in function spaces*, Function spaces VIII, 215–234, Banach Center Publ., 79, Polish Acad. Sci. Inst. Math., Warsaw, 2008.

[112] H. Triebel, *Theory of function spaces III*, Birkhäuser, Basel, 2006.

[113] H. Triebel, *Function Spaces and Wavelets on Domains*, EMS Tracts in Mathematics, Vol. 7, EMS Publishing House, Zürich, 2008.

[114] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, *Beyond Nyquist: Efficient sampling of sparse bandlimited signals*, IEEE Trans. Inform. Theor. **56**, 520–544 (2010)

[115] V. Vapnik and A. Chervonenkis, *A note on one class of perceptrons*, Automation and Remote Control, vol. 25, no. 1, 1964.

[116] J. A. van Vechten, Phys. Rev. 182, 1969.

[117] J. Vybíral, *Decomposition methods and their applications in the theory of function spaces*, Habilitation thesis, Friedrich-Schiller-Universtät Jena, 2011.

[118] P. Wojtaszczyk, *A mathematical introduction to wavelets*, London Math. Soc. Student Text **37**, Cambridge Univ. Press, 1997.

[119] P. Wojtaszczyk, *Complexity of approximation of functions of few variables in high dimensions*, J. Complexity **27**, 141–150 (2011)

[120] W. Yuan, W. Sickel and D. Yang, *Morrey and Campanato meet Besov, Lizorkin and Triebel*, Lecture Notes in Math. 2005, Springer, Berlin 2010.

[121] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines”, In Proc. Advances in Neural Information Processing Systems, vol. 16, pp. 49–56, 2004.

[122] A. Zunger, Phys. Rev. B 22, 1980.

A New Proof of the Jawerth-Franke Embedding

Jan VYBÍRAL

Mathematisches Institut
Friedrich-Schiller-Universität Jena
Ernst-Abbe-Platz 3
07740 Jena — Germany
vybiral@minet.uni-jena.de

Received: May 28, 2007
Accepted: July 2, 2007

ABSTRACT

We present an alternative proof of the Jawerth embedding

$$F^{s_0}_{p_0 q}(\mathbb{R}^n) \hookrightarrow B^{s_1}_{p_1 p_0}(\mathbb{R}^n),$$

where

$$-\infty < s_1 < s_0 < \infty, \quad 0 < p_0 < p_1 \leq \infty, \quad 0 < q \leq \infty$$

and

$$s_0 - \frac{n}{p_0} = s_1 - \frac{n}{p_1}.$$

The original proof given in [3] uses interpolation theory. Our proof relies on wavelet decompositions and transfers the problem from function spaces to sequence spaces. Using similar techniques, we also recover the embedding of Franke [2].

Key words: Besov spaces, Triebel-Lizorkin spaces, Sobolev embedding, Jawerth-Franke embedding.

2000 Mathematics Subject Classification: 46E35.

Introduction

Let $B^s_{pq}(\mathbb{R}^n)$ and $F^s_{pq}(\mathbb{R}^n)$ denote the Besov and Triebel-Lizorkin function spaces, respectively. The classical Sobolev embedding theorem can be extended to these two scales.

Theorem 0.1. *Let* $-\infty < s_1 < s_0 < \infty$ *and* $0 < p_0 < p_1 \le \infty$ *with*

$$s_0 - \frac{n}{p_0} = s_1 - \frac{n}{p_1}. \tag{1}$$

(i) *If* $0 < q_0 \le q_1 \le \infty$, *then*

$$B^{s_0}_{p_0q_0}(\mathbb{R}^n) \hookrightarrow B^{s_1}_{p_1q_1}(\mathbb{R}^n).$$

(ii) *If* $0 < q_0, q_1 \le \infty$ *and* $p_1 < \infty$, *then*

$$F^{s_0}_{p_0q_0}(\mathbb{R}^n) \hookrightarrow F^{s_1}_{p_1q_1}(\mathbb{R}^n). \tag{2}$$

We observe that there is no condition on the fine paramters q_0, q_1 in (2). This surprising effect was first observed in full generality by Jawerth, [3]. Using (2), we may prove

$$F^{s_0}_{p_0q}(\mathbb{R}^n) \hookrightarrow F^{s_1}_{p_1p_1}(\mathbb{R}^n) = B^{s_1}_{p_1p_1}(\mathbb{R}^n)$$

and

$$B^{s_0}_{p_0p_0}(\mathbb{R}^n) = F^{s_0}_{p_0p_0}(\mathbb{R}^n) \hookrightarrow F^{s_1}_{p_1q}(\mathbb{R}^n)$$

for every $0 < q \le \infty$. But Jawerth [3] and Franke [2] showed that these embeddings are not optimal and may be improved.

Theorem 0.2. *Let* $-\infty < s_1 < s_0 < \infty$, $0 < p_0 < p_1 \le \infty$, *and* $0 < q \le \infty$ *with* (1).

(i) *Then*

$$F^{s_0}_{p_0q}(\mathbb{R}^n) \hookrightarrow B^{s_1}_{p_1p_0}(\mathbb{R}^n). \tag{3}$$

.

(ii) *If* $p_1 < \infty$, *then*

$$B^{s_0}_{p_0p_1}(\mathbb{R}^n) \hookrightarrow F^{s_1}_{p_1q}(\mathbb{R}^n). \tag{4}$$

The original proofs (see [2, 3]) use interpolation techniques. We rely on a different method. First, we observe that using (for example) the wavelet decomposition method, (3) and (4) are equivalent to

$$f^{s_0}_{p_0q} \hookrightarrow b^{s_1}_{p_1p_0} \quad \text{and} \quad b^{s_0}_{p_0p_1} \hookrightarrow f^{s_1}_{p_1q} \tag{5}$$

under the same restrictions on parameters s_0, s_1, p_0, p_1, q as in Theorem 0.2. Here, b^s_{pq} and f^s_{pq} stands for the sequence spaces of Besov and Triebel-Lizorkin type. We prove (5) directly using the technique of non-increasing rearrangement on a rather elementary level.

All the unimportant constants are denoted by the letter c, whose meaning may differ from one occurrence to another. If $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ are two sequences of positive real numbers, we write $a_n \lesssim b_n$ if, and only if, there is a positive real number $c > 0$ such that $a_n \le c\, b_n, n \in \mathbb{N}$. Furthermore, $a_n \approx b_n$ means that $a_n \lesssim b_n$ and simultaneously $b_n \lesssim a_n$.

1. Notation and definitions

We introduce the sequence spaces associated with the Besov and Triebel-Lizrokin spaces. Let $m \in \mathbb{Z}^n$ and $\nu \in \mathbb{N}_0$. Then $Q_{\nu\, m}$ denotes the closed cube in $\mathbb{R}^n$ with sides parallel to the coordinate axes, centred at $2^{-\nu}m$, and with side length $2^{-\nu}$. By $\chi_{\nu\, m} = \chi_{Q_{\nu\, m}}$ we denote the characteristic function of $Q_{\nu\, m}$. If

$$\lambda = \{\, \lambda_{\nu\, m} : \nu \in \mathbb{N}_0, m \in \mathbb{Z}^n \,\},$$

$-\infty < s < \infty$, and $0 < p, q \le \infty$, we set

$$\|\lambda \mid b^s_{pq}\| = \left(\sum_{\nu=0}^{\infty} 2^{\nu(s-\frac{n}{p})q} \Big(\sum_{m\in\mathbb{Z}^n} |\lambda_{\nu\, m}|^p \Big)^{\frac{q}{p}} \right)^{\frac{1}{q}},$$

appropriately modified if $p = \infty$ and/or $q = \infty$. If $p < \infty$, we define also

$$\|\lambda | f^s_{pq}\| = \left\| \left(\sum_{\nu=0}^{\infty} \sum_{m\in\mathbb{Z}^n} |2^{\nu s}\lambda_{\nu\, m}\chi_{\nu\, m}(\cdot)|^q \right)^{1/q} \;\middle|\; L_p(\mathbb{R}^n) \right\|.$$

The connection between the function spaces $B^s_{pq}(\mathbb{R}^n)$, $F^s_{pq}(\mathbb{R}^n)$ and the sequence spaces b^s_{pq}, f^s_{pq} may be given by various decomposition techniques, we refer to [7, chapters 2 and 3] for details and further references.

As a result of these characterizations, (3) is equivalent to (5).

We use the technique of non-increasing rearrangement. We refer to [1, chapter 2] for details.

Definition 1.1. Let μ be the Lebesgue measure in $\mathbb{R}^n$. If h is a measurable function on $\mathbb{R}^n$, we define the non-increasing rearrangement of h through

$$h^*(t) = \sup\{\, \lambda > 0 : \mu\{x \in \mathbb{R}^n : |h(x)| > \lambda\} > t \,\}, \qquad t \in (0, \infty).$$

We denote its averages by

$$h^{**}(t) = \frac{1}{t}\int_0^t h^*(s)\, ds, \quad t > 0.$$

We shall use the following properties. The first two are very well known and their proofs may be found in [1, Proposition 1.8 in chapter 2, Theorem 3.10 in chapter 3].

Lemma 1.2. *If* $0 < p \le \infty$, *then*

$$\|h \mid L_p(\mathbb{R}^n)\| = \|h^* \mid L_p(0,\infty)\|$$

for every measurable function h.

Lemma 1.3. *If* $1 < p \le \infty$, *then there is a constant* c_p *such that*

$$\|h^{**} \mid L_p(0,\infty)\| \le c_p \|h^* \mid L_p(0,\infty)\|$$

for every measurable function h.

Lemma 1.4. *Let* h_1 *and* h_2 *be two non-negative measurable functions on* $\mathbb{R}^n$. *If* $1 \le p \le \infty$, *then*

$$\|h_1 + h_2 \mid L_p(\mathbb{R}^n)\| \le \|h_1^* + h_2^* \mid L_p(0,\infty)\|.$$

Proof. The proof follows from Theorems 3.4 and 4.6 in [1, chapter2]. □

2. Main results

In this part, we present a direct proof of the discrete versions of Jawerth and Franke embedding. We start with the Jawerth embedding.

Theorem 2.1. *Let* $-\infty < s_1 < s_0 < \infty$, $0 < p_0 < p_1 \le \infty$, *and* $0 < q \le \infty$. *Then*

$$f^{s_0}_{p_0 q} \hookrightarrow b^{s_1}_{p_1 p_0} \quad \textit{if} \quad s_0 - \frac{n}{p_0} = s_1 - \frac{n}{p_1}.$$

Proof. Using the elementary embedding

$$f^s_{pq_0} \hookrightarrow f^s_{pq_1} \quad \text{if} \quad 0 < q_0 \le q_1 \le \infty \tag{6}$$

and the lifting property of Besov and Triebel-Lizorkin spaces (which is even simpler in the language of sequence spaces), we may restrict ourselves to the proof of

$$f^s_{p_0\infty} \hookrightarrow b^0_{p_1 p_0}, \quad \text{where} \quad s = n\Big(\frac{1}{p_0} - \frac{1}{p_1}\Big).$$

Let $\lambda \in f^s_{p_0\infty}$ and set

$$h(x) = \sup_{\nu\in\mathbb{N}_0} 2^{\nu s} \sum_{m\in\mathbb{Z}^n} |\lambda_{\nu m}|\, \chi_{\nu m}(x).$$

Hence

$$|\lambda_{\nu\, m}| \le 2^{-\nu s} \inf_{x\in Q_{\nu\, m}} h(x), \quad \nu\in\mathbb{N}_0, \quad m\in\mathbb{Z}^n.$$

Using this notation,

$$\|\lambda \mid f^s_{p_0\infty}\| = \|h \mid L_{p_0}(\mathbb{R}^n)\|$$

and

$$\begin{aligned}\|\lambda \mid b^0_{p_1p_0}\|^{p_0} &\le \sum_{\nu=0}^\infty 2^{-\nu n}\Big(\sum_{m\in\mathbb{Z}^n} \inf_{x\in Q_{\nu m}} h(x)^{p_1}\Big)^{p_0/p_1}\\ &\le \sum_{\nu=0}^\infty 2^{-\nu n}\Big(\sum_{k=1}^\infty h^*(2^{-\nu n}k)^{p_1}\Big)^{p_0/p_1}.\end{aligned}$$

Using the monotonicity of h^* and $p_0<p_1$ we get

$$\begin{aligned}\|\lambda \mid b^0_{p_1p_0}\|^{p_0} &\lesssim \sum_{\nu=0}^\infty 2^{-\nu n}\Big(\sum_{l=0}^\infty 2^{nl}\cdot(2^n-1)\cdot h^*(2^{-\nu n}2^{nl})^{p_1}\Big)^{p_0/p_1}\\ &\lesssim \sum_{\nu=0}^\infty 2^{-\nu n}\sum_{l=0}^\infty 2^{nl\frac{p_0}{p_1}} h^*(2^{-\nu n}2^{nl})^{p_0}.\end{aligned}$$

We substitute $j=l-\nu$ and obtain

$$\begin{aligned}\|\lambda \mid b^0_{p_1p_0}\|^{p_0} &\lesssim \sum_{j=-\infty}^\infty \sum_{\nu=-j}^\infty 2^{-\nu n}2^{n(\nu+j)\frac{p_0}{p_1}} h^*(2^{jn})^{p_0}\\ &= \sum_{j=-\infty}^\infty 2^{nj\frac{p_0}{p_1}} h^*(2^{jn})^{p_0} \sum_{\nu=-j}^\infty 2^{n\nu\left(\frac{p_0}{p_1}-1\right)}\\ &\approx \sum_{j=-\infty}^\infty 2^{nj}h^*(2^{nj})^{p_0} \approx \|h^* \mid L_{p_0}(0,\infty)\|^{p_0} = \|h \mid L_{p_0}(\mathbb{R}^n)\|^{p_0}.\end{aligned}$$

If $p_1=\infty$, only notational changes are necessary. □

Theorem 2.2. *Let* $-\infty<s_1<s_0<\infty, 0<p_0<p_1<\infty$, *and* $0<q\le\infty$. *Then*

$$b^{s_0}_{p_0p_1} \hookrightarrow f^{s_1}_{p_1q} \quad \textit{if} \quad s_0-\frac{n}{p_0}=s_1-\frac{n}{p_1}.$$

Proof. Using the lifting property and (6), we may suppose that $s_1=0$ and $0<q<p_0$.

By Lemma 1.4, we observe that

$$\|\lambda|f^0_{p_1q}\| = \left\|\Big(\sum_{\nu=0}^\infty \sum_{m\in\mathbb{Z}^n} |\lambda_{\nu m}|^q\chi_{\nu m}(x)\Big)^{1/q} \;\Big|\; L_{p_1}(\mathbb{R}^n)\right\|$$

may be estimated from above by

$$\left\|\sum_{\nu=0}^{\infty}\sum_{m=0}^{\infty}\tilde{\lambda}_{\nu m}^{q}\tilde{\chi}_{\nu m}(\cdot)\;\Big|\;L_{\frac{p_1}{q}}(0,\infty)\right\|^{1/q}, \tag{7}$$

where $\tilde{\lambda}_\nu=\{\tilde{\lambda}_{\nu m}\}_{m=0}^{\infty}$ is a non-increasing rearrangement of $\lambda_\nu=\{\lambda_{\nu m}\}_{m\in\mathbb{Z}^n}$ and $\tilde{\chi}_{\nu m}$ is a characteristic function of the interval $(2^{-\nu n}m,2^{-\nu n}(m+1))$.

Using duality, (7) may be rewritten as

$$\sup_g\left(\int_0^{\infty}g(x)\left(\sum_{\nu=0}^{\infty}\sum_{m=0}^{\infty}\tilde{\lambda}_{\nu m}^{q}\tilde{\chi}_{\nu m}(x)\right)dx\right)^{1/q}=\sup_g\left(\sum_{\nu=0}^{\infty}\sum_{m=0}^{\infty}2^{-\nu n}\tilde{\lambda}_{\nu m}^{q}g_{\nu m}\right)^{1/q}, \tag{8}$$

where the supremum is taken over all non-increasing non-negative measurable functions g with $\|g\mid L_\beta(0,\infty)\|\le 1$ and $g_{\nu m}=2^{\nu n}\int g(x)\tilde{\chi}_{\nu m}(x)\,dx$. Here, β is the conjugated index to $\frac{p_1}{q}$. Similarly, α stands for the conjugated index to $\frac{p_0}{q}$.

We use twice Hölder's inequality and estimate (8) from above by

$$\left(\sum_{\nu=0}^{\infty}2^{-\nu n}\left(\sum_{m=0}^{\infty}\tilde{\lambda}_{\nu m}^{p_0}\right)^{\frac{p_1}{p_0}}\right)^{1/p_1}\cdot\sup_g\left(\sum_{\nu=0}^{\infty}2^{-\nu n}\left(\sum_{m=0}^{\infty}g_{\nu m}^{\alpha}\right)^{\frac{\beta}{\alpha}}\right)^{\frac{1}{\beta q}} \tag{9}$$

Since $s_0=n\big(\frac{1}{p_0}-\frac{1}{p_1}\big)$ and $p_1\big(s_0-\frac{n}{p_0}\big)=-n$, the first factor in (9) is equal to $\|\lambda\mid b_{p_0p_1}^{s_0}\|$. To finish the proof, we have to show that there is a number $c>0$ such that

$$\left(\sum_{\nu=0}^{\infty}2^{-\nu n}\left(\sum_{m=0}^{\infty}g_{\nu m}^{\alpha}\right)^{\frac{\beta}{\alpha}}\right)^{\frac{1}{\beta q}}\le c \tag{10}$$

holds for every non-increasing non-negative measurable functions g with $\|g\mid L_\beta(0,\infty)\|\le 1$. We fix such a function g. Using the monotonicity of g, we get

$$\begin{aligned}\sum_{m=0}^{\infty}g_{\nu m}^{\alpha}&=\sum_{l=0}^{\infty}\sum_{m=2^{ln}-1}^{2^{(l+1)n}}\left(2^{\nu n}\int_{2^{-\nu n}m}^{2^{-\nu n}(m+1)}g(x)\,dx\right)^{\alpha}\\&\lesssim\sum_{l=0}^{\infty}2^{ln}\left(2^{\nu n}\int_{2^{-\nu n}(2^{ln}-1)}^{2^{-\nu n}2^{ln}}g(x)\,dx\right)^{\alpha}\le\sum_{l=0}^{\infty}2^{ln}(g^{**})^{\alpha}(2^{(l-\nu)n}).\end{aligned}$$

We use $1 < \beta < \alpha$, Lemma 1.3 and obtain

$$\begin{aligned}\left(\sum_{\nu=0}^{\infty} 2^{-\nu n}\left(\sum_{m=0}^{\infty} g_{\nu m}^{\alpha}\right)^{\frac{\beta}{\alpha}}\right)^{1/\beta} &\le \left(\sum_{\nu=0}^{\infty} 2^{-\nu n}\left(\sum_{l=0}^{\infty} 2^{ln}(g^{**})^{\alpha}(2^{(l-\nu)n})\right)^{\frac{\beta}{\alpha}}\right)^{1/\beta}\\ &\le \left(\sum_{\nu=0}^{\infty} 2^{-\nu n}\sum_{l=0}^{\infty} 2^{ln\frac{\beta}{\alpha}}(g^{**})^{\beta}(2^{(l-\nu)n})\right)^{1/\beta}\\ &\le \left(\sum_{k=-\infty}^{\infty} 2^{kn\frac{\beta}{\alpha}}\sum_{\nu=-k}^{\infty} 2^{\nu n(\frac{\beta}{\alpha}-1)}(g^{**})^{\beta}(2^{kn})\right)^{1/\beta}\\ &\lesssim \left(\sum_{k=-\infty}^{\infty} 2^{kn}(g^{**})^{\beta}(2^{kn})\right)^{1/\beta}\\ &\lesssim \|g^{**} \mid L_{\beta}(0,\infty)\| \le c\,\|g \mid L_{\beta}(0,\infty)\| \le c.\end{aligned}$$

Taking the $\frac{1}{q}$-power of this estimate, we finish the proof of (10). □

The Theorems 2.1 and 2.2 are sharp in the following sense.

Theorem 2.3. *Let $-\infty < s_1 < s_0 < \infty$, $0 < p_0 < p_1 \le \infty$, and $0 < q_0, q_1 \le \infty$ with*

$$s_0 - \frac{n}{p_0} = s_1 - \frac{n}{p_1}.$$

(i) *If*

$$f^{s_0}_{p_0 q_0} \hookrightarrow b^{s_1}_{p_1 q_1}, \tag{11}$$

then $q_1 \ge p_0$.

(ii) *If $p_1 < \infty$ and*

$$b^{s_0}_{p_0 q_0} \hookrightarrow f^{s_1}_{p_1 q_1}, \tag{12}$$

then $q_0 \le p_1$.

Remark 2.4. Using (any of) the usual decomposition techniques, the same statements hold true also for the function spaces. These results were first proved in [4].

Proof. (i) Suppose that $0 < q_1 < p_0 < \infty$ and set

$$\lambda_{\nu m} = \begin{cases} \nu^{-\frac{1}{q_1}} 2^{\nu(\frac{n}{p_1}-s_1)} & \text{if} \quad \nu \in \mathbb{N} \quad \text{and} \quad m = 0,\\ 0, & \text{otherwise.}\end{cases}$$

A simple calculation shows that $\|\lambda \mid f^{s_0}_{p_0 q_0}\| < \infty$ and $\|\lambda \mid b^{s_1}_{p_1 q_1}\| = \infty$. Hence, (11) does not hold.

(ii) Suppose that $0 < p_1 < q_0 \le \infty$ and set

$$\lambda_{\nu m} = \begin{cases} \nu^{-\frac{1}{p_1}} 2^{\nu(\frac{n}{p_1}-s_1)} & \text{if} \quad \nu \in \mathbb{N} \quad \text{and} \quad m = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Again, it is a matter of simple calculation to show, that $\|\lambda \mid b^{s_0}_{p_0q_0}\| < \infty$ and $\|\lambda \mid f^{s_1}_{p_1q_1}\| = \infty$. Hence, (12) is not true. □

References

[1] C. Bennett and R. Sharpley, *Interpolation of operators*, Pure and Applied Mathematics, vol. 129, Academic Press Inc., Boston, MA, 1988.

[2] J. Franke, *On the spaces $\mathbf{F}^s_{pq}$ of Triebel-Lizorkin type: pointwise multipliers and spaces on domains*, Math. Nachr. **125** (1986), 29–68.

[3] B. Jawerth, *Some observations on Besov and Lizorkin-Triebel spaces*, Math. Scand. **40** (1977), no. 1, 94–104.

[4] W. Sickel and H. Triebel, *Hölder inequalities and sharp embeddings in function spaces of B^s_{pq} and F^s_{pq} type*, Z. Anal. Anwendungen **14** (1995), no. 1, 105–140.

[5] H. Triebel, *Theory of function spaces*, Monographs in Mathematics, vol. 78, Birkhäuser Verlag, Basel, 1983.

[6] ———, *Theory of function spaces*, II, Monographs in Mathematics, vol. 84, Birkhäuser Verlag, Basel, 1992.

[7] ———, *Theory of function spaces*, III, Monographs in Mathematics, vol. 100, Birkhäuser Verlag, Basel, 2006.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Complexity 24 (2008) 545 – 570
Journal of COMPLEXITY
www.elsevier.com/locate/jco

Widths of embeddings in function spaces

Jan Vybíral

Mathematisches Institut, Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

Received 27 November 2007; accepted 4 January 2008
Available online 17 January 2008

Abstract

We study the approximation, Gelfand and Kolmogorov numbers of embeddings in function spaces of Besov and Triebel-Lizorkin type. Our aim here is to provide sharp estimates in several cases left open in the literature and give a complete overview of the known results. We also add some historical remarks.

MSC: 41A45; 41A46; 46E35

Keywords: Approximation numbers; Gelfand numbers; Kolmogorov numbers; Besov spaces; Sobolev embeddings

1. Introduction

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, $1 \leqslant p \leqslant \infty$ and let k be a natural number. We denote by $W_p^k(\Omega)$ the Sobolev spaces of functions from $L_p(\Omega)$ with all distributive derivatives of order smaller or equal to k in $L_p(\Omega)$. If

$$k_1 - k_2 \geqslant d\left(\frac{1}{p_1} - \frac{1}{p_2}\right)_+, \tag{1.1}$$

and the boundary of Ω is Lipschitz then $W_{p_1}^{k_1}(\Omega)$ is continuously embedded into $W_{p_2}^{k_2}(\Omega)$. This theorem goes back to Sobolev [55].

If the inequality in (1.1) is strict, the embedding is even compact, cf. [48,31]. During the second half of the last century, this fact (and its numerous generalizations) found its applications in many areas of modern analysis, especially in connection with partial differential (and pseudo-differential) equations.

E-mail address: vybiral@minet.uni-jena.de.

0885-064X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2008.01.002

Later on, mathematicians started to be interested in measuring the *quality of compactness* of the embedding

$$I : W^{k_1}_{p_1}(\Omega) \hookrightarrow W^{k_2}_{p_2}(\Omega).$$

The very first question is, of course, how to measure compactness. During the years, several methods were developed. The most popular one assigns to I a non-increasing sequence of non-negative real numbers, say $\{s_n(I)\}_{n\in\mathbb{N}}$, often based on specific approximation quantities, and measures the decay of s_n as n tends to infinity.

Let us present this approach on the following example. Let X and Y be Banach spaces and let $T : X \to Y$ be a bounded linear operator between them. Then the nth approximation number of T is defined by

$$a_n(T) = \inf\{\|T - L\| : L \in \mathcal{L}(X, Y),\ \text{rank}\ (L) < n\}, \quad n \in \mathbb{N}, \tag{1.2}$$

where $\mathcal{L}(X, Y)$ is the space of all bounded linear operators mapping X into Y endowed with the classical operator norm and rank L denotes the dimension of $L(X)$. Hence, we measure how well the operator T may be approximated by finite rank operators. If $\lim_{n\to\infty} a_n(T) = 0$, then T is compact. And in some sense, the faster the sequence $\{a_n(T)\}_{n\in\mathbb{N}}$ tends to zero, the more compact T is.

There are many other ways, how to define a sequence $\{s_n(T)\}_{n\in\mathbb{N}}$ for an operator $T \in \mathcal{L}(X, Y)$ such that the decay of $\{s_n\}$ describes in some sense the compactness of T; we refer to [43,44,6], where the axiomatic theory of the so-called s-numbers can be found.

It was observed by many authors, that even in the most simple case

$$id : \ell^m_{p_1} \to \ell^m_{p_2}, \quad m \in \mathbb{N}$$

it is surprisingly difficult to calculate (or at least estimate) the approximation numbers, as well as the other s-numbers, corresponding to id. The complexity of the problem may be demonstrated by the fact that in several cases the proofs are based on probabilistic arguments and no optimal constructive approximation procedure is known up to now.

As a part of the good news is that these results may be combined with the discretization technique of Maĭorov [37] to get direct counterparts for embeddings between function spaces. Nowadays, there are many discretization techniques well known and studied in the literature. Let us mention at least spline and wavelet decompositions and the φ-transform, cf. [8,7,49,64,23,11, 16,17].

The research in this area was complicated also by another regretful phenomena, namely communication problems between several groups working on the field. This effect was already pointed out by Caetano [4] and Pietsch [45, Section 6.2.6]. Also the separation of the Russian mathematical school causes some obstacles. Many breakthroughs achieved by Kashin, Gluskin and others were published in Russian. The nicely written dissertation of Lubitz [36] was written in German, never translated into English and never published.

The aim of this paper is rather extensive. We wish to

- give an overview of known results in this area,
- collect some historical references,
- close several minor gaps left open until now,
- present the power of the discretization method, but also its limits,
- provide an easy reference to the results about function spaces.

Several overviews may already be found in the literature, cf. [46,34,35,45]. Unfortunately, they sometimes restrict themselves to $d = 1$, state the results only implicitly, or deal only with integer smoothness parameters $s_1, s_2 \in \mathbb{N}$. Here, leaded by the needs of possible applications, we shall study three types of s-numbers, namely approximation, Kolmogorov and Gelfand numbers, with respect to embeddings of function spaces defined on Lipschitz domains. This generalization is not particularly interesting from the standpoint of functional analysis, but is of course crucial as far as the applications are concerned.

2. Function and sequence spaces

2.1. Notation

We use standard notation: $\mathbb{N}$ denotes the collection of all natural numbers, $\mathbb{Z}$ the collection of all integers, $\mathbb{R}^d$ is the Euclidean d-dimensional space, where $d \in \mathbb{N}$, and $\mathbb{C}$ stands for the complex plane. Let $S(\mathbb{R}^d)$ be the Schwartz space of all complex-valued rapidly decreasing, infinitely differentiable functions on $\mathbb{R}^d$ and let $S'(\mathbb{R}^d)$ be its dual, the space of all tempered distributions.

Furthermore, $L_p(\mathbb{R}^d)$ with $0 < p \leqslant \infty$ are the classical Lebesgue spaces endowed with the (quasi-)norm

$$\|f|L_p(\mathbb{R}^d)\| = \begin{cases} \left(\int_{\mathbb{R}^d} |f(x)|^p dx\right)^{1/p}, & 0 < p < \infty, \\ \operatorname*{ess\,sup}_{x \in \mathbb{R}^d} |f(x)|, & p = \infty. \end{cases}$$

For $\psi \in S(\mathbb{R}^d)$ we denote by

$$\widehat{\psi}(\xi) = (F\psi)(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i<x,\xi>} \psi(x)\, dx, \quad x \in \mathbb{R}^d,$$

its Fourier transform and by $\psi^\vee$ or $F^{-1}\psi$ its inverse Fourier transform. Through duality, F and F^{-1} are extended to $S'(\mathbb{R}^d)$.

If $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$ are two sequences of non-negative real numbers, we write $a_n \lesssim b_n$ if there is a constant $c > 0$, such that $a_n \leqslant c\, b_n$ for all natural numbers n. The symbols $a_n \gtrsim b_n$ and $a_n \approx b_n$ are defined similarly.

2.2. Function spaces

We give a Fourier-analytic definition of Besov and Triebel-Lizorkin spaces, which relies on the so-called *smooth dyadic resolution of unity*. Let $\varphi \in S(\mathbb{R}^d)$ with

$$\varphi(x) = 1 \text{ if } |x| \leqslant 1 \quad \text{and} \quad \varphi(x) = 0 \quad \text{if } |x| \geqslant \tfrac{3}{2}. \tag{2.1}$$

We put $\varphi_0 = \varphi$ and $\varphi_j(x) = \varphi(2^{-j}x) - \varphi(2^{-j+1}x)$ for $j \in \mathbb{N}$ and $x \in \mathbb{R}^d$. This leads to the identity

$$\sum_{j=0}^{\infty} \varphi_j(x) = 1, \quad x \in \mathbb{R}^d.$$

Definition 2.1. (i) Let $s \in \mathbb{R}, 0 < p, q \leqslant \infty$. Then $B^s_{pq}(\mathbb{R}^d)$ is the collection of all $f \in S'(\mathbb{R}^d)$ such that

$$\|f|B^s_{pq}(\mathbb{R}^d)\| = \left(\sum_{j=0}^{\infty} 2^{jsq}\|(\varphi_j\widehat{f})^\vee|L_p(\mathbb{R}^d)\|^q\right)^{1/q} < \infty \tag{2.2}$$

(with the usual modification for $q = \infty$).

(ii) Let $s \in \mathbb{R}, 0 < p < \infty, 0 < q \leqslant \infty$. Then $F^s_{pq}(\mathbb{R}^d)$ is the collection of all $f \in S'(\mathbb{R}^d)$ such that

$$\|f|F^s_{pq}(\mathbb{R}^d)\| = \left\|\left(\sum_{j=0}^{\infty} 2^{jsq}|(\varphi_j\widehat{f})^\vee(\cdot)|^q\right)^{1/q}|L_p(\mathbb{R}^d)\right\| < \infty \tag{2.3}$$

(with the usual modification for $q = \infty$).

Remark 2.2. We recommend [40,59,60,51,61] as standard references with respect to these classes of distributions. Extensive historical overviews, remarks and comments may be found in [60, Chapter 1], [61, Chapter 1] and [45, Chapter 6.7]. Let us mention that the spaces $B^s_{pq}(\mathbb{R}^d)$ and $F^s_{pq}(\mathbb{R}^d)$ do not depend on the choice of φ in the sense of equivalent (quasi-)norms. Many classical function spaces are included in these two scales.

1. If $1 < p < \infty$, then the Littlewood–Paley theorem states that

   $$F^0_{p2}(\mathbb{R}^d) = L_p(\mathbb{R}^d).$$

2. Let $1 < p < \infty$ and $s \in \mathbb{N}$. Then

   $$F^s_{p2}(\mathbb{R}^d) = W^s_p(\mathbb{R}^d)$$

   are the classical Sobolev spaces.
3. Let $s > 0, s \notin \mathbb{N}$. Then

   $$B^s_{\infty\infty}(\mathbb{R}^d) = \mathcal{C}^s(\mathbb{R}^d)$$

   are the Hölder–Zygmund spaces.

On the other hand, many important function spaces (especially $L_1(\mathbb{R}^d)$, $L_\infty(\mathbb{R}^d)$, $BV(\mathbb{R})$—the space of functions with bounded variation and $C^k(\mathbb{R}^d)$—the space of functions with all partial derivatives of order smaller or equal to k uniformly continuous and bounded) are *not* included.

If X and Y are two topological vector spaces, we write $X \hookrightarrow Y$ if X is continuously embedded in Y. The following embeddings describe the interplay between these function spaces and the Besov scale.

$$\begin{aligned}
&B^0_{11}(\mathbb{R}^d) \hookrightarrow L_1(\mathbb{R}^d) \hookrightarrow B^0_{1\infty}(\mathbb{R}^d),\\
&B^0_{\infty 1}(\mathbb{R}^d) \hookrightarrow C(\mathbb{R}^d) \hookrightarrow L_\infty(\mathbb{R}^d) \hookrightarrow B^0_{\infty\infty}(\mathbb{R}^d),\\
&B^k_{\infty 1}(\mathbb{R}^d) \hookrightarrow C^k(\mathbb{R}^d) \hookrightarrow B^k_{\infty\infty}(\mathbb{R}^d).
\end{aligned} \tag{2.4}$$

In many cases it will be possible to use the Fourier-analytical methods in the framework of Besov spaces and afterwards, simply by applying these simple continuous embeddings, to derive the same results also for the "bad" spaces $L_1(\mathbb{R}^d)$, $L_\infty(\mathbb{R}^d)$ and $C^k(\mathbb{R}^d)$. The same procedure may be used also for the Triebel-Lizorkin scale because of

$$B^s_{p,\min(p,q)}(\mathbb{R}^d) \hookrightarrow F^s_{pq}(\mathbb{R}^d) \hookrightarrow B^s_{p,\max(p,q)}(\mathbb{R}^d). \tag{2.5}$$

Remark 2.3. If $0 < p_1 \leqslant p_2 \leqslant \infty$, $0 < q_1, q_2 \leqslant \infty$ and $s_2 \leqslant s_1$, then the following version of the Sobolev embedding is true, see [2], [40, Chapters 3 and 11] and [58, Section 2.8.1]:

$$B^{s_1}_{p_1,q_1}(\mathbb{R}^d) \hookrightarrow B^{s_2}_{p_2,q_2}(\mathbb{R}^d), \quad \text{if } s_1 - \frac{d}{p_1} > s_2 - \frac{d}{p_2}.$$

There are several modifications of this embedding, which result in compact mappings. The first possibility is to restrict to function spaces on smooth bounded domains, the second involves *weighted spaces* and another one considers the so-called *radial spaces*, i.e. spaces of radial symmetric functions. We concentrate on the first possibility and refer to [61, Chapter 6], [54] for the second and third approach.

Let Ω be a bounded domain. Let $D(\Omega) = C_0^\infty(\Omega)$ be the collection of all complex-valued infinitely differentiable functions with compact support in Ω and let $D'(\Omega)$ be its dual—the space of all complex-valued distributions on Ω.

Let $g \in S'(\mathbb{R}^d)$. Then we denote by $g|\Omega$ its restriction to Ω:

$$(g|\Omega) \in D'(\Omega), \quad (g|\Omega)(\psi) = g(\psi) \quad \text{for} \quad \psi \in D(\Omega).$$

Definition 2.4. Let Ω be a bounded domain in $\mathbb{R}^d$. Let $s \in \mathbb{R}$, $0 < p, q \leqslant \infty$ with $p < \infty$ in the F-case. Let A^s_{pq} stand either for B^s_{pq} or F^s_{pq}. Then

$$A^s_{pq}(\Omega) = \{f \in D'(\Omega) : \exists g \in A^s_{pq}(\mathbb{R}^d) : g|\Omega = f\}$$

and

$$\|f|A^s_{pq}(\Omega)\| = \inf \|g|A^s_{pq}(\mathbb{R}^d)\|,$$

where the infimum is taken over all $g \in A^s_{pq}(\mathbb{R}^d)$ such that $g|\Omega = f$.

Intrinsic characterization of $B^s_{p,q}(\Omega)$, $s > \sigma_p = d\left(\frac{1}{p} - 1\right)_+ = d \max\left(\frac{1}{p} - 1, 0\right)$ are known to exist in case of Lipschitz domains, see [12–14] and [61, Section 1.11.9].

2.3. Sequence spaces

In this section we comment on the discretization techniques mentioned in the Introduction.

First, we describe the situation on $\mathbb{R}^d$. Therefore, we introduce the sequence spaces b^s_{pq} and give a wavelet decomposition theorem for Besov spaces on $\mathbb{R}^d$. Good references in our context are [8,11,23,38,39,63,64].

Second, we deal with bounded domains $\Omega \subset \mathbb{R}^d$. The wavelet decomposition techniques may be adapted also to these function spaces, cf. [9,61], but unfortunately, there are still open problems in this setting. To avoid these gaps, we use the theory on $\mathbb{R}^d$ and combine it with suitable extension and restriction operators.

Theorem 2.5. *For any $k \in \mathbb{N}$ there are real-valued compactly supported functions*

$$\psi_0, \psi_1 \in C^k(\mathbb{R})$$

satisfying

$$\int_{\mathbb{R}} t^\alpha \psi_1(t)\, dt = 0, \quad \alpha = 0, 1, \dots, k-1,$$

such that

$$\{2^{v/2}\psi_{vm} : v \in \mathbb{N}_0, m \in \mathbb{Z}\}$$

with

$$\psi_{vm}(t) = \begin{cases} \psi_0(t-m) & \text{if } v = 0, m \in \mathbb{Z}, \\ 2^{-\frac{1}{2}}\psi_1(2^{v-1}t - m) & \text{if } v \in \mathbb{N}, m \in \mathbb{Z} \end{cases}$$

is an orthonormal basis in $L_2(\mathbb{R})$.

Remark 2.6. This theorem was first proven by Daubechies in [10]. The functions ψ_0 and ψ_1 are therefore usually called Daubechies wavelets. We refer to [63, Theorem 19] for the proof of the next theorem.

Theorem 2.7. *Let $0 < p, q \leqslant \infty$, $s \in \mathbb{R}$ and $k \in \mathbb{N}$ with $k > \max(s, \sigma_p - s)$. Let ψ_0, ψ_1 be the Daubechies wavelets of smoothness k. Let $E = \{0,1\}^d \setminus (0, \dots, 0)$. For $e = (e_1, \dots, e_d) \in E$ let*

$$\Psi_e(x) = \prod_{j=1}^d \psi_{e_j}(x_j), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

(i) *Then*

$$\begin{cases} \Psi(x-m) = \prod_{j=1}^d \psi_0(x_j - m_j), & m = (m_1, \dots, m_d) \in \mathbb{Z}^d, \\ 2^{\frac{v-1}{2}d}\Psi_e(2^{v-1}x - m), & e \in E, v \in \mathbb{N}, m \in \mathbb{Z}^d \end{cases}$$

is an orthonormal basis in $L_2(\mathbb{R}^d)$.

(ii) *Let $f \in S'(\mathbb{R}^d)$. Then $f \in B^s_{pq}(\mathbb{R}^d)$ if, and only if, it can be represented as*

$$f = \sum_{m \in \mathbb{Z}^d} \lambda_m \Psi(x-m) + \sum_{v \in \mathbb{N}} \sum_{e \in E} \sum_{m \in \mathbb{Z}^d} \lambda^e_{vm} 2^{-vd/2} \Psi_e(2^{v-1}x - m) \tag{2.6}$$

with

$$\|\lambda|\mathrm{b}^s_{pq}\| = \left(\sum_{m \in \mathbb{Z}^d} |\lambda_m|^p\right)^{\frac{1}{p}} + \left(\sum_{v=1}^{\infty} 2^{v(s-\frac{d}{p})q} \sum_{e \in E} \left(\sum_{m \in \mathbb{Z}^d} |\lambda^e_{vm}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} < \infty$$

appropriately modified if $p = \infty$ and/or $q = \infty$. The representation in (2.6) *is unique, the complex coefficients $\{\lambda_m\}_{m \in \mathbb{Z}^d}$ and $\{\lambda^e_{vm}\}_{e \in E, v \in \mathbb{N}_0, m \in \mathbb{Z}^d}$ depend linearly on f and the* *mapping, which associates to $f \in B^s_{pq}(\mathbb{R}^d)$ the sequence of coefficients, is an isomorphic map of $B^s_{pq}(\mathbb{R}^d)$ onto b^s_{pq}.*

2.4. *s-Numbers*

Given $p \in (0, 1]$, we say that the quasi-Banach space Y is a p-Banach space if the inequality

$$\|x + y|Y\|^p \leqslant \|x|Y\|^p + \|y|Y\|^p, \quad x, y \in Y$$

is satisfied.

We recall a few basic facts of the theory of s-numbers. We refer to [44,6] for further details. In this theory, one associates to every linear operator $T : X \to Y$ (X and Y quasi-Banach spaces) a sequence of scalars

$$s_1(T) \geqslant s_2(T) \geqslant \cdots \geqslant 0.$$

Let W, X, Y, Z be (quasi-)Banach spaces and let Y be a p-Banach space, $0 < p \leqslant 1$. If the rule $s : T \to \{s_n(T)\}_{n\in\mathbb{N}}$ satisfies

(S1) $\|T\| = s_1(T) \geqslant s_2(T) \geqslant \cdots \geqslant 0.$
(S2) $s^p_{m+n-1}(S + T) \leqslant s^p_m(T) + s^p_n(S)$ for all $S, T \in \mathcal{L}(X, Y)$ and $m, n \in \mathbb{N}$.
(S1) $s_n(STU) \leqslant \|S\| s_n(T) \|U\|$ for all $U \in \mathcal{L}(W, X)$, $T \in \mathcal{L}(X, Y)$, $S \in \mathcal{L}(Y, Z)$ and $n \in \mathbb{N}$.
(S4) If rank $T < n$, then $s_n(T) = 0$.
(S5) $s_n(I : \ell_2(n) \to \ell_2(n)) = 1$

then the $s_n(T)$ are called s-numbers of the operator T.

Let us point out, that we shall not use (**S4**) and (**S5**) in what follows. Hence, our approach applies also to rules $s : T \to \{s_n(T)\}_{n\in\mathbb{N}}$ which satisfy only (**S1**)–(**S3**). Such rules are called *pseudo-s-numbers* in [43, Chapter 12] and cover also the concept of entropy numbers with $\|T\| \geqslant s_1(T)$ in (**S1**).

Let

$$\mathcal{I}d : B^{s_1}_{p_1q_1}(\Omega) \to B^{s_2}_{p_2q_2}(\Omega) \tag{2.7}$$

be compact, i.e.

$$s_1 - s_2 > d\left(\frac{1}{p_1} - \frac{1}{p_2}\right)_+ . \tag{2.8}$$

We denote by

$$\mathrm{ext} : B^{s_1}_{p_1q_1}(\Omega) \to B^{s_1}_{p_1q_1}(\mathbb{R}^d) \tag{2.9}$$

a bounded linear extension operator. A convenient reference for this is Rychkov, cf. [52], but see also the references given there. Here we use the Lipschitz smoothness of $\partial\Omega$. The natural restriction will be denoted by

$$\mathrm{re} : B^{s_2}_{p_2q_2}(\mathbb{R}^d) \to B^{s_2}_{p_2q_2}(\Omega).$$

Clearly, it also represents a bounded linear operator.

Let $k > \max(s_1, \sigma_{p_1} - s_1, s_2, \sigma_{p_2} - s_2)$ be a natural number and let $\mathcal{W}$ be the mapping which associates to each $f \in B^{s_1}_{p_1q_1}(\mathbb{R}^d)$ its wavelet coefficients with respect to the Daubechies wavelets of smoothness k, as described in Theorem 2.7. Our choice of k ensures that Theorem 2.7 may be applied to both, $B^{s_1}_{p_1q_1}(\mathbb{R}^d)$ and $B^{s_2}_{p_2q_2}(\mathbb{R}^d)$, simultaneously and that $\mathcal{W}^{-1}$ is a bounded linear operator, which maps $\mathrm{b}^{s_2}_{p_2q_2}$ isomorphically onto $B^{s_2}_{p_2q_2}(\mathbb{R}^d)$.

Finally, we adapt the sequence spaces b^{s}_{pq} to the function spaces on domains.

Definition 2.8. (i) Let $M = \{M_\nu\}_{\nu=0}^\infty$ be a sequence of non-negative integers. We say that M is admissible, if there is some $\nu_0 \in \mathbb{N}_0$ and two positive real constants c_1, c_2 such that

$$M_\nu = 0 \quad \text{for all } \nu < \nu_0$$

and

$$c_1 2^{\nu d} \leqslant M_\nu \leqslant c_2 2^{\nu d}, \qquad \nu \geqslant \nu_0.$$

(ii) If $0 < p, q \leqslant \infty$, $s \in \mathbb{R}$, $E = \{0,1\}^d \setminus (0,\dots,0)$, $M = \{M_\nu\}_{\nu=0}^\infty$ is an admissible sequence and

$$\lambda = \{\lambda_k : k = 1,\dots,M_0\} \cup \{\lambda^e_{\nu k} : e \in E, \nu \in \mathbb{N}, k \in M_\nu\},$$

we set

$$\|\lambda|\mathrm{b}^{s,M}_{pq}\| = \left(\sum_{k=1}^{M_0} |\lambda_k|^p\right)^{\frac{1}{p}} + \left(\sum_{\nu=1}^{\infty} 2^{\nu(s-\frac{d}{p})q} \sum_{e\in E} \left(\sum_{k=1}^{M_\nu} |\lambda^e_{\nu k}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}, \tag{2.10}$$

again appropriately modified if $p = \infty$ and/or $q = \infty$.

Let now Ω be a bounded Lipschitz domain in $\mathbb{R}^d$ and let the number $k \in \mathbb{N}$ describing the smoothness of the wavelets be fixed. Then we collect those wavelets, whose support intersects $\overline{\Omega}$:

$$\mathcal{M}_\nu = \begin{cases} \{m \in \mathbb{Z}^d : \operatorname{supp} \Psi(\cdot - m) \cap \overline{\Omega} \neq \emptyset\} & \text{if } \nu = 0, \\ \{m \in \mathbb{Z}^d : \exists e \in E : \operatorname{supp} \Psi_e(2^{\nu-1} \cdot - m) \cap \overline{\Omega} \neq \emptyset\} & \text{if } \nu \geqslant 1. \end{cases}$$

We observe that the sequence $M = \{M_\nu\}_{\nu=0}^\infty$ with

$$M_\nu = \#(\mathcal{M}_\nu) = \text{number of elements of } \mathcal{M}_\nu, \qquad \nu \in \mathbb{N}_0$$

is an admissible sequence in the sense of Definition 2.8.

With a slight abuse of notation, there is a natural projection operator $P : \mathrm{b}^{s}_{pq} \to \mathrm{b}^{s,M}_{pq}$ and a natural embedding operator $Q : \mathrm{b}^{s,M}_{pq} \to \mathrm{b}^{s}_{pq}$.

Using the weak multiplicativity property **(S3)** of s-numbers and the commutative diagram

$$\begin{tikzcd}
B^{s_1}_{p_1q_1}(\Omega) \arrow[r, "\mathrm{ext}"] \arrow[d, "\mathcal{I}d"'] & B^{s_1}_{p_1q_1}(\mathbb{R}^d) \arrow[r, "\mathcal{W}"] & \mathrm{b}^{s_1}_{p_1q_1} \arrow[r, "P"] & \mathrm{b}^{s_1,M}_{p_1q_1} \arrow[d, "id"] \\
B^{s_2}_{p_2q_2}(\Omega) & B^{s_2}_{p_2q_2}(\mathbb{R}^d) \arrow[l, "\mathrm{re}"'] & \mathrm{b}^{s_2}_{p_2q_2} \arrow[l, "\mathcal{W}^{-1}"'] & \mathrm{b}^{s_2,M}_{p_2q_2} \arrow[l, "Q"']
\end{tikzcd}$$

we conclude that

$$s_n(\mathcal{I}d) \lesssim s_n(id), \qquad n \in \mathbb{N}.$$

To obtain the reverse inequality, we first set

$$\mathcal{M}'_\nu = \begin{cases} \{m \in \mathbb{Z}^d : \operatorname{supp} \Psi(\cdot - m) \subset \Omega\} & \text{if } \nu = 0, \\ \{m \in \mathbb{Z}^d : \forall e \in E : \operatorname{supp} \Psi_e(2^{\nu-1} \cdot - m) \subset \Omega\} & \text{if } \nu \geqslant 1. \end{cases} \tag{2.11}$$

Again, we observe that the sequence $M' = \{M'_\nu\}_{\nu=0}^\infty$ with

$$M'_\nu = \#(\mathcal{M}'_\nu) = \text{number of elements of } \mathcal{M}'_\nu, \quad \nu \in \mathbb{N}_0$$

is an admissible sequence in the sense of Definition 2.8.

If we use **(S3)** and

$$\begin{tikzcd}
\mathrm{b}^{s_1,M'}_{p_1q_1} \arrow[r, "Q'"] \arrow[d, "id'"'] & \mathrm{b}^{s_1}_{p_1q_1} \arrow[r, "\mathcal{W}^{-1}"] & B^{s_1}_{p_1q_1}(\mathbb{R}^d) \arrow[r, "\mathrm{re}"] & B^{s_1}_{p_1q_1}(\Omega) \arrow[d, "\mathcal{I}d"] \\
\mathrm{b}^{s_2,M'}_{p_2q_2} & \mathrm{b}^{s_2}_{p_2q_2} \arrow[l, "P'"'] & B^{s_2}_{p_2q_2}(\mathbb{R}^d) \arrow[l, "\mathcal{W}"'] & B^{s_2}_{p_2q_2}(\Omega), \arrow[l, "\mathrm{ext}"']
\end{tikzcd}$$

we get the inequality.

$$s_n(id') \lesssim s_n(\mathcal{I}d), \quad n \in \mathbb{N}.$$

Hence

$$s_n(id') \lesssim s_n(\mathcal{I}d) \lesssim s_n(id), \quad n \in \mathbb{N}. \tag{2.12}$$

It tells us, roughly speaking, that we may restrict ourselves to sequence spaces and all the results translate also into the language of function spaces. Before we start with the study of $s_n(id)$ and $s_n(id')$, we make another simplification. The (finite) sum over $e \in E$ in (2.10) comes from the theory of multivariate wavelet decompositions, but has no influence on the s-numbers.

If $M = \{M_\nu\}_{\nu=0}^\infty$ is an admissible sequence, we set

$$\|\lambda | b^{s,M}_{pq}\| = \left(\sum_{\nu=0}^\infty 2^{\nu(s-\frac{d}{p})q} \left(\sum_{k=1}^{M_\nu} |\lambda_{\nu k}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}.$$

It follows that

$$\begin{aligned} s_n(\mathcal{I}d : B^{s_1}_{p_1q_1}(\Omega) \to B^{s_2}_{p_2q_2}(\Omega)) &\approx s_n(id : \mathrm{b}^{s,M}_{pq} \to \mathrm{b}^{s,M}_{pq}) \\ &\approx s_n(id : b^{s,M}_{pq} \to b^{s,M}_{pq}). \end{aligned} \tag{2.13}$$

Remark 2.9. Formulas (2.12) and (2.13) represent the main result of this section and is of a crucial importance for our study of s-numbers of (2.7). We have proved (2.13) under the assumption that Ω is a bounded domain in $\mathbb{R}^d$ with Lipschitz boundary. Using more sophisticated tools from the theory of function spaces, it may be proven that (2.13) holds also for more general classes of domains, at least under some restrictions on the parameters $s_1, s_2, p_1, p_2, q_1, q_2$. A detailed inspection of our proof shows that (2.13) is true anytime there is a bounded linear extension operator (2.9) and its counterpart for $B^{s_2}_{p_2q_2}(\Omega)$. We refer to [62, Section 4.3.4] for a detailed treatment of these questions.

3. Approximation numbers

Definition 3.1. Let X, Y be two quasi-Banach spaces and let $T \in \mathcal{L}(X,Y)$. For $n \in \mathbb{N}$, we define the nth approximation number by

$$a_n(T) = \inf\{\|T - L\| : L \in \mathcal{L}(X,Y), \operatorname{rank}(L) < n\}.$$

In the setting of Banach spaces, this definition goes back to Pietsch [41] and Tikhomirov [57]. The generalization to quasi-Banach spaces may be found in [15, Section 1.3.1]. In this section, we characterize the approximation numbers of (2.7) with (2.8).

First, we recall some lemmas which we shall need on the sequence space level. Lemma 3.2 is taken from [22] and Lemma 3.3 in the case $1 \leqslant p_2 \leqslant p_1 \leqslant \infty$ may be found in [43, Section 11.11.5]. The proof may be directly generalized to the quasi-Banach setting $0 < p_2 \leqslant p_1 \leqslant \infty$.

For $0 < p \leqslant \infty$, we set

$$p' = \begin{cases} \dfrac{p}{p-1} & \text{if } 1 < p < \infty, \\ 1 & \text{if } p = \infty, \\ \infty & \text{if } 0 < p \leqslant 1. \end{cases}$$

Lemma 3.2. *For $1 \leqslant n \leqslant m < \infty$ and $1 \leqslant p_1 < p_2 \leqslant \infty$, we define*

$$\Phi(m,n,p_1,p_2) := \begin{cases} \left(\min\{1, m^{\frac{1}{p_2}} n^{-\frac{1}{2}}\}\right)^{\frac{\frac{1}{p_1} - \frac{1}{p_2}}{\frac{1}{2} - \frac{1}{p_2}}} & \text{if } 2 \leqslant p_1 < p_2 \leqslant \infty, \\ \max\{m^{\frac{1}{p_2} - \frac{1}{p_1}}, \min\{1, m^{\frac{1}{p_2}} n^{-\frac{1}{2}}\} \cdot \sqrt{1 - \frac{n}{m}}\} & \text{if } 1 \leqslant p_1 < 2 \leqslant p_2 \leqslant \infty, \\ \max\{m^{\frac{1}{p_2} - \frac{1}{p_1}}, \sqrt{1 - \frac{n}{m}}^{\frac{\frac{1}{p_1} - \frac{1}{p_2}}{\frac{1}{p_1} - \frac{1}{2}}}\} & \text{if } 1 \leqslant p_1 < p_2 \leqslant 2 \end{cases}$$

and

$$\Psi(m,n,p_1,p_2) := \begin{cases} \Phi(m,n,p_1,p_2) & \text{if } 1 \leqslant p_1 < p_2 \leqslant p_1', \\ \Phi(m,n,p_2',p_1') & \text{if } \max(p_1,p_1') < p_2 \leqslant \infty. \end{cases}$$

Then if $1 \leqslant p_1 < p_2 \leqslant \infty$ and $(p_1,p_2) \neq (1,\infty)$

$$a_n(id : \ell_{p_1}^m \to \ell_{p_2}^m) \approx \Phi(m,n,p_1,p_2), \quad 1 \leqslant n \leqslant m < \infty.$$

The constants of equivalence may depend on p_1 and p_2 but are independent of m and n.

Lemma 3.3. *If $1 \leqslant n \leqslant m < \infty$ and $0 < p_2 \leqslant p_1 \leqslant \infty$, then*

$$a_n(id : \ell_{p_1}^m \to \ell_{p_2}^m) = (m - n + 1)^{\frac{1}{p_2} - \frac{1}{p_1}}.$$

Lemma 3.4. *Let $0 < p \leqslant 1$.*

(i) *Let $0 < \lambda < 1$. Then there is a number $c_\lambda > 0$ such that*

$$a_n(id : \ell_p^m \to \ell_\infty^m) \leqslant \frac{c_\lambda}{\sqrt{n}} \tag{3.1}$$

holds for all natural numbers n and m with $m^\lambda < n \leqslant m$.

(ii) *There is a number* $c > 0$ *such that*

$$a_n(id : \ell_p^{2n} \to \ell_\infty^{2n}) \geqslant \frac{c}{\sqrt{n}}, \quad n \geqslant 1. \tag{3.2}$$

Proof. Let $A = (a_{i,j})_{i,j=1}^m$ be an $m \times m$ matrix. Then

$$\|A|\mathcal{L}(\ell_1^m, \ell_\infty^m)\| = \|A|\mathcal{L}(\ell_p^m, \ell_\infty^m)\| = \max_{i,j=1,\ldots,m} |a_{i,j}|$$

for every $0 < p \leqslant 1$. Hence, the approximation numbers of $id : \ell_p^m \to \ell_\infty^m$ do not depend on $0 < p \leqslant 1$ and it is enough, when we prove Lemma 3.4 only for $p = 1$.

The first part follows from a combinatorial result of Kashin, cf. [26,27] and [43, Section 11.11.11]:

Let $0 < \lambda < 1$ and $m^\lambda \leqslant n \leqslant m$ be natural numbers. Then there are m ℓ_2^n-unit vectors $\{f_i\}_{i=1}^m \subset \mathbb{R}^n$, such that

$$|(f_i, f_j)| \leqslant \frac{c_\lambda}{\sqrt{n}} \quad \text{if } i \neq j.$$

We set $A = (a_{i,j})_{i,j=1}^m$ with $a_{i,j} = (f_i, f_j)$. Then A is a matrix with rank $A \leqslant n$ and $\|I - A|\mathcal{L}(\ell_1^m, \ell_\infty^m)\| \leqslant \frac{c_\lambda}{\sqrt{n}}$.

The proof of the second part follows trivially from the result of Stechkin, cf. [56] and [43, Section 11.11.8]:

$$a_n(id : \ell_1^m \to \ell_2^m) = \left(\frac{m-n+1}{m}\right)^{1/2}$$

and

$$\|id : \ell_\infty^m \to \ell_2^m\| = \sqrt{m}. \qquad \square$$

Theorem 3.5. *Let* $-\infty < s_2 < s_1 < \infty$ *and* $0 < p_1, p_2, q_1, q_2 \leqslant \infty$ *with* (2.8). *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipschitz domain. Then* (2.7) *is compact and for* $n \in \mathbb{N}$

$$a_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}+\left(\frac{1}{p_1}-\frac{1}{p_2}\right)_+} \quad \textit{if} \begin{cases} \textit{either} & 0 < p_1 \leqslant p_2 \leqslant 2, \\ \textit{or} & 2 \leqslant p_1 \leqslant p_2 \leqslant \infty, \\ \textit{or} & 0 < p_2 \leqslant p_1 \leqslant \infty, \end{cases} \tag{3.3}$$

$$a_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}+\frac{1}{p}-\frac{1}{2}} \quad \textit{if } 0 < p_1 < 2 < p_2 < \infty \text{ } \textit{and } \frac{s_1-s_2}{d} > \frac{1}{p} = \max\left(1-\frac{1}{p_2}, \frac{1}{p_1}\right), \tag{3.4}$$

$$a_n(\mathcal{I}d) \approx n^{\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{p_2}\right)\cdot\frac{\min(p_1',p_2)}{2}} \quad \textit{if } \frac{s_1-s_2}{d} < \frac{1}{p} = \max\left(1-\frac{1}{p_2}, \frac{1}{p_1}\right) \text{ } \textit{and either } 1 < p_1 < 2 < p_2 = \infty \text{ } \textit{or } 0 < p_1 < 2 < p_2 < \infty, \tag{3.5}$$

$$a_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{2}} \quad \textit{if } \ 0 < p_1 \leqslant 1 < p_2 = \infty. \tag{3.6}$$

Proof. Approximation numbers form an additive and multiplicative scale of s-numbers. This fact may be verified directly, or the reader may consult [43, Section 11.2] in the Banach space settings and [15, Section 1.3] for the extension to quasi-Banach spaces.

Hence (2.12) applies to approximation numbers and we may restrict ourselves to sequence spaces.

The estimates covered by (3.3)–(3.5) are known. We refer to [15, Section 3.3.4] and [4]. The proof given in [15] is rather complicated, but [4] uses an approach very similar to ours.

It remains to prove the only missing case (3.6). We use Lemma 3.4 to estimate the approximation numbers of

$$id : b^{s_1,M}_{p_1q_1} = \ell_{q_1}(2^{\nu(s_1-\frac{d}{p_1})}\ell^{M_\nu}_{p_1}) \to \ell_{q_2}(2^{\nu s_2}\ell^{M_\nu}_{\infty}) = b^{s_2,M}_{\infty q_2},$$

where $M = \{M_\nu\}_{\nu=0}^{\infty}$ is an admissible sequence. Let

$$id_\nu : 2^{\nu(s_1-\frac{d}{p_1})}\ell^{M_\nu}_{p_1} \to 2^{\nu s_2}\ell^{M_\nu}_{\infty}, \quad \nu = 0, 1, 2, \ldots$$

denote the identity operator between the finite dimensional building blocks of the considered sequence spaces. With a slight abuse of notation, we get

$$id = \sum_{\nu=0}^{\infty} id_\nu, \tag{3.7}$$

which, combined with the additivity of approximation numbers, leads to

$$a^{\omega}_{n'}(id) \leqslant \sum_{\nu=0}^{N_1} a^{\omega}_{n_\nu}(id_\nu) + \sum_{\nu=N_1+1}^{N_2} a^{\omega}_{n_\nu}(id_\nu) + \sum_{\nu=N_2+1}^{\infty} \|id_\nu\|^{\omega},$$

where $N_1 < N_2$ are natural numbers, $n' - 1 = \sum_{\nu=0}^{N_2}(n_\nu - 1)$ and $\omega = \min(1, q_2)$. We set

$$n_\nu = \begin{cases} M_\nu + 1 & \text{if } 0 \leqslant \nu \leqslant N_1, \\ n^{1+\alpha}2^{-\alpha\nu d} & \text{if } N_1 + 1 \leqslant \nu \leqslant N_2, \end{cases}$$

where

$$0 < \alpha < 2\left(\frac{s}{d} - \frac{1}{p_1}\right) \tag{3.8}$$

and

$$N_1 = \left[\frac{\log_2 n}{d}\right], \quad N_2 = \left[\frac{\frac{s}{d} - \frac{1}{p} + \frac{1}{2}}{\frac{s}{d} - \frac{1}{p}} \cdot \frac{\log_2 n}{d}\right] \geqslant N_1.$$

Here, $[a]$ denotes the integer part of a real number a.

For this choice we get

$$n' = \sum_{\nu=0}^{N_2}(n_\nu - 1) + 1 \approx 2^{\nu N_1 d} + N_1^{1+\alpha}2^{-\alpha\nu d} \approx n.$$

A simple calculation shows that there is a number $\lambda > 0$ such that $M_\nu^\lambda \leqslant n_\nu \leqslant M_\nu$. Hence

$$a_{n_\nu}(id_\nu) \leqslant \begin{cases} 0 & \text{if } 0 \leqslant \nu \leqslant N_1, \\ \dfrac{c_\lambda}{\sqrt{n_\nu}} 2^{-\nu(s-\frac{d}{p_1})} & \text{if } N_1 + 1 \leqslant \nu \leqslant N_2 \end{cases}$$

and

$$\sum_{\nu=0}^{N_1} a_{n_\nu}^{\omega}(id_\nu) = 0,$$

$$\sum_{\nu=N_1+1}^{N_2} a_{n_\nu}^{\omega}(id_\nu) \leqslant \sum_{\nu=N_1+1}^{N_2} \frac{c_\lambda^{\omega}}{\sqrt{n_\nu^{\omega}}} \leqslant c n^{-\frac{1+\alpha}{2}\omega} \sum_{\nu=N_1+1}^{N_2} 2^{-\nu d\omega(\frac{s}{d} - \frac{1}{p_1} - \frac{\alpha}{2})} \lesssim n^{-\omega\left(\frac{s}{d} - \frac{1}{p_1} + \frac{1}{2}\right)},$$

$$\sum_{\nu=N_2+1}^{\infty} \|id_\nu\|^{\omega} \leqslant \sum_{\nu=N_2+1}^{\infty} 2^{-\nu\omega(s-\frac{d}{p_1})} \lesssim n^{-\omega\left(\frac{s}{d} - \frac{1}{p_1} + \frac{1}{2}\right)}.$$

It follows, that there is a constant $c > 0$ such that

$$a_{cn}(id) \lesssim n^{-\left(\frac{s}{d} - \frac{1}{p_1} + \frac{1}{2}\right)}, \quad n \geqslant 1,$$

which is equivalent to

$$a_n(id) \lesssim n^{-\left(\frac{s}{d} - \frac{1}{p_1} + \frac{1}{2}\right)}, \quad n \geqslant 1. \tag{3.9}$$

The proof of the reverse inequality to (3.9) follows easily from the second part of Lemma 3.4.

Let $M' = \{M'_\nu\}_{\nu=0}^{\infty}$ be an admissible sequence. Then, for $\nu \geqslant \nu_0$

$$a_n(id) \geqslant a_n(id_\nu) \gtrsim 2^{-\nu(s-\frac{d}{p_1})} \cdot \frac{1}{\sqrt{n}}$$

if $n = \left[\frac{M_\nu}{2}\right]$. This leads to

$$a_n(id) \gtrsim n^{-\left(\frac{s}{d} - \frac{1}{p_1} + \frac{1}{2}\right)}, \quad n = \left[\frac{M_\nu}{2}\right], \quad \nu \geqslant \nu_0$$

and by means of the monotonicity of the approximation numbers the result follows. □

Remark 3.6. We have used the open case (3.6) to demonstrate the typical use of the wavelet decomposition method and (2.12). Also (3.3)–(3.5) could be proven exactly in the same manner. For example, the proof of (3.5) in [4] follows along this line.

Remark 3.7. Although the results were stated only for Besov spaces, with the aid of (2.4) and (2.5) we may extend them also to Triebel-Lizorkin spaces, Sobolev and Lebesgue spaces and $C(\Omega)$, $L_1(\Omega)$ and $L_\infty(\Omega)$. We return to this point later on.

Remark 3.8. The first estimates on approximation numbers of Sobolev embeddings of function spaces were obtained by Kolmogorov [30], who dealt with the Hilbert space case $p_1 = q_1 = p_2 = q_2 = 2$. Later on, Birman and Solomyak [3] studied the embeddings of Sobolev spaces. Finally, Kashin [29] observed the effect of “small smoothness" expressed by (3.5). In the framework of

Besov spaces the results are contained in [15,4]. Nowadays, the proof of (3.3)–(3.5) could be done very similar to the proof of (3.6), only using Lemmas 3.2 and 3.3 instead of Lemma 3.4.

4. Kolmogorov and Gelfand numbers

In this chapter we deal with Kolmogorov and Gelfand numbers. To begin with we recall their definition and describe their decay in connection with Sobolev embeddings of Besov spaces. We use the symbol $A \subset\subset B$ if A is a closed subspace of a topological vector space B.

Definition 4.1. Let X, Y be two quasi-Banach spaces and let $T \in \mathcal{L}(X, Y)$.

(i) For $n \in \mathbb{N}$, we define the nth Kolmogorov number by

$$d_n(T) = \inf\{\|Q_N^Y T\| : N \subset\subset Y, \dim(N) < n\}.$$

Here, Q_N^Y stands for the natural surjection of Y onto the quotient space Y/N.

(ii) For $n \in \mathbb{N}$, we define the nth Gelfand number by

$$c_n(T) = \inf\{\|T J_M^X\| : M \subset\subset X, \operatorname{codim}(M) < n\}.$$

Here, J_M^X stands for the natural injection of M into X.

Clearly, the notion *dimension of a subspace* is purely algebraic and may be freely used also in the setting of quasi-Banach spaces. We refer to [50, Section 1.40] for the definition of a quotient subspace in the framework of general topological vector spaces (including quasi-Banach spaces as a special case). Finally, the codimension of a subspace may be defined as the dimension of the quotient space.

Both, Gelfand and Kolmogorov numbers, are additive and multiplicative s-scales. This follows directly from Definition 4.1, but the reader may wish to consult [44, Sections 2.4, 2.5] for the proof in the Banach space case. The generalization to p-Banach spaces is obvious and causes no complications. Also the following relations are trivial:

$$c_n(T) \leqslant a_n(T), \quad d_n(T) \leqslant a_n(T), \quad n \in \mathbb{N}. \tag{4.1}$$

The Gelfand and Kolmogorov numbers are dual to each other in the following sense, cf. [44, Section 11.7.6-7]: If X and Y are Banach spaces, then

$$c_n(T^*) = d_n(T) \tag{4.2}$$

for all compact operators $T \in \mathcal{L}(X, Y)$ and

$$d_n(T^*) = c_n(T) \tag{4.3}$$

for all $T \in \mathcal{L}(X, Y)$.

The following result is due to Gluskin, cf. [21,22] with [56,24,26,27] as forerunners. It gives a very precise information on the behaviour of $d_n(id : \ell_{p_1}^m \to \ell_{p_2}^m)$ in the Banach space setting.

Lemma 4.2. *For $1\leqslant n\leqslant m < \infty$ and $1\leqslant p_1, p_2\leqslant\infty$, we define*

$$\Phi(m,n,p_1,p_2) := \begin{cases} (m-n+1)^{\frac{1}{p_2}-\frac{1}{p_1}} & \text{if } 1\leqslant p_2\leqslant p_1\leqslant\infty, \\ \left(\min\{1, m^{\frac{1}{p_2}} n^{-\frac{1}{2}}\}\right)^{\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{2}-\frac{1}{p_2}}} & \text{if } 2\leqslant p_1 < p_2\leqslant\infty, \\ \max\{m^{\frac{1}{p_2}-\frac{1}{p_1}}, \sqrt{1-\frac{n}{m}}^{\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{p_1}-\frac{1}{2}}}\} & \text{if } 1\leqslant p_1 < p_2\leqslant 2, \\ \max\{m^{\frac{1}{p_2}-\frac{1}{p_1}}, & \text{if } 1\leqslant p_1 < 2 < p_2\leqslant\infty. \\ \quad \min\{1, m^{\frac{1}{p_2}} n^{-\frac{1}{2}}\}\cdot\sqrt{1-\frac{n}{m}}\} & \end{cases}$$

Then

$$d_n(id : \ell_{p_1}^m \to \ell_{p_2}^m) \approx \Phi(m,n,p_1,p_2), \quad 1\leqslant n\leqslant m < \infty,$$

if $p_2 < \infty$. The constants of equivalence may depend on p_1 and p_2 but are independent of m and n.

Furthermore, there are two constants c_{p_1} and C_{p_1} such that

$$c_{p_1}\Phi(m,n,p_1,\infty)\leqslant d_n(id : \ell_{p_1}^m \to \ell_\infty^m)\leqslant C_{p_1}\Phi(m,n,p_1,\infty)\left(\log\left(\frac{em}{n}\right)\right)^{3/2}$$

for $1\leqslant p_1\leqslant\infty$.

Again we shall add some estimates which apply to quasi-Banach spaces.

Lemma 4.3. *If $0 < p_2\leqslant p_1\leqslant\infty$, then there is a constant $c > 0$ such that*

$$d_{[cn]+1}(\ell_{p_1}^{2n}, \ell_{p_2}^{2n}) \gtrsim n^{\frac{1}{p_2}-\frac{1}{p_1}}, \quad n\in\mathbb{N},$$

where $[cn]$ denotes the upper integer part of cn.

Proof. If $p_2\geqslant 1$, then the result is a special case of [43, Section 11.11.4], which states that

$$d_n(\ell_{p_1}^m, \ell_{p_2}^m) = (m-n+1)^{\frac{1}{p_2}-\frac{1}{p_1}}, \quad 1\leqslant n\leqslant m.$$

Let us mention that (in contrast to Lemmas 3.3 and 4.8) the estimate

$$d_n(\ell_{p_1}^m, \ell_{p_2}^m) = (m-n+1)^{\frac{1}{p_2}-\frac{1}{p_1}}, \quad 1\leqslant n\leqslant m\leqslant\infty$$

is *not* true for Kolmogorov numbers if $0 < p_2\leqslant p_1\leqslant\infty$ and $p_2 < 1$. Simple counterexamples can be constructed directly.

If $p_2 < 1$ the proof is based on an inequality between entropy numbers and Kolmogorov numbers. First, we recall the basic facts about entropy numbers. Let $T : X \to Y$ be a bounded linear operator between two quasi-Banach spaces X and Y and let U_X and U_Y be the unit ball of X and Y, respectively. If $k\in\mathbb{N}$, we define the kth entropy number $e_k(T)$ as the infimum of all $\varepsilon > 0$

such that

$$T(U_X) \subset \bigcup_{j=1}^{2^{k-1}} (y_j + \varepsilon U_Y) \quad \text{for some } y_1, \dots, y^{2^{k-1}} \in Y.$$

We refer to [43,15] for detailed discussions of this concept, its history and further references.

The following Lemma may be found in [1], cf. also [5] and [47, Section 5].

Lemma 4.4. *If* $\alpha > 0$ *and* $0 < p < 1$, *then there is a constant* $c_{\alpha,p} > 0$ *such that for all p-Banach spaces X and Y, all linear mappings* $T : X \to Y$ *and all* $n \in \mathbb{N}$ *we have*

$$\sup_{k \leqslant n} k^\alpha e_k(T) \leqslant c_{\alpha,p} \sup_{k \leqslant n} k^\alpha d_k(T).$$

We apply this lemma to $T = id : \ell_{p_1}^{2n} \to \ell_{p_2}^{2n}$ and combine it with the estimate (cf. [53])

$$e_k(T) \gtrsim 2^{-\frac{k}{4n}} (2n)^{\frac{1}{p_2} - \frac{1}{p_1}}, \quad k, n \in \mathbb{N}.$$

This leads to

$$n^\alpha n^{\frac{1}{p_2} - \frac{1}{p_1}} \lesssim \sup_{k \leqslant n} k^\alpha d_k(T).$$

Hence, for every $n \in \mathbb{N}$ there is a $k_n \leqslant n$ such that

$$n^\alpha n^{\frac{1}{p_2} - \frac{1}{p_1}} \lesssim k_n^\alpha d_{k_n}(T) \leqslant k_n^\alpha (2n)^{\frac{1}{p_2} - \frac{1}{p_1}}. \tag{4.4}$$

We conclude that there is a constant $1 \geqslant c > 0$ such that $n \geqslant k_n \geqslant cn$ for all $n \in \mathbb{N}$. Finally, we insert this estimate into (4.4) and the result follows. □

It is an obvious fact that the convex hull of the unit ball of ℓ_p^m, $0 < p < 1$, is the unit ball of ℓ_1^m. This can be combined with the following simple observation, cf. [35, Section 13.1].

Lemma 4.5. *Let X be a Banach space and let* $K \subset X$. *We define by*

$$d_n(K, X) = \inf \left\{ \sup_{x \in K} \inf_{y \in N} \|x - y\| : N \subset\subset Y, \dim(N) < n \right\}$$

the nth Kolmogorov number of the set K.

Then

$$d_n(K, X) = d_n(\operatorname{conv} K, X),$$

where conv K *is the convex hull of K.*

Theorem 4.6. *Let* $-\infty < s_2 < s_1 < \infty$ *and* $0 < p_1, p_2, q_1, q_2 \leqslant \infty$ *with* (2.8). *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipschitz domain. Then* (2.7) *is compact and for* $n \in \mathbb{N}$

$$d_n(\mathcal{I}d) \approx n^{-\frac{s_1 - s_2}{d} + \left(\frac{1}{p_1} - \frac{1}{p_2}\right)_+} \quad \text{if} \begin{cases} \text{either} & 0 < p_1 \leqslant p_2 \leqslant 2, \\ \text{or} & 0 < p_2 \leqslant p_1 \leqslant \infty, \end{cases} \tag{4.5}$$

$$d_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}} \quad \textit{if } 2 < p_1 \leqslant p_2 \leqslant \infty \ \textit{and } \frac{s_1-s_2}{d} > \frac{1}{2}\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{2}-\frac{1}{p_2}}, \tag{4.6}$$

$$d_n(\mathcal{I}d) \approx n^{\frac{p_2}{2}\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{p_2}\right)} \quad \textit{if } 2 < p_1 \leqslant p_2 \leqslant \infty \ \textit{and } \frac{s_1-s_2}{d} < \frac{1}{2}\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{2}-\frac{1}{p_2}}, \tag{4.7}$$

$$d_n(\mathcal{I}d) \approx n^{\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{2}\right)} \quad \textit{if } 0 < p_1 < 2 < p_2 \leqslant \infty \ \textit{and } \frac{s_1-s_2}{d} > \frac{1}{p_1}, \tag{4.8}$$

$$d_n(\mathcal{I}d) \approx n^{\frac{p_2}{2}\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{p_2}\right)} \quad \textit{if } 0 < p_1 < 2 < p_2 < \infty \ \textit{and } \frac{1}{p_1}-\frac{1}{p_2} < \frac{s_1-s_2}{d} < \frac{1}{p_1}. \tag{4.9}$$

Proof. Lubitz [36] used the results of [21] and was able to prove (4.5)–(4.9) if $1 \leqslant p_1, p_2 \leqslant \infty$ up to a certain logarithmic gap. This gap originates from using only the weaker results of [21] instead of the sharp inequalities in [22]. Using [22] and the method of Lubitz (which is very similar to the discretization method presented above), the proof of (4.5)–(4.9) in the Banach space setting follows immediately.

Hence, we concentrate on the proof of

(♣) (4.5) if $0 < p_2 \leqslant p_1 \leqslant \infty$ and $0 < p_2 < 1$,
(♡) (4.5) if $0 < p_1 < p_2 \leqslant 2$ and $0 < p_1 < 1$,
(♠) (4.8) if $0 < p_1 < 1, 2 < p_2 \leqslant \infty$ and $\dfrac{s_1-s_2}{d} > \dfrac{1}{p_1}$,
(◇) (4.8) if $0 < p_1 < 1, 2 < p_2 < \infty$ and $\dfrac{1}{p_1}-\dfrac{1}{p_2} < \dfrac{s_1-s_2}{d} < \dfrac{1}{p_1}$.

Let us mention that all the estimates from above follow from the estimates given in Theorem 3.5 and (4.1). We shall give the proof of the estimates from below in following three steps.

Step 1: *Proof of* (♣).

The proof of (4.5) can be finished in the same manner as in the proof of Theorem 3.5. Namely, if $M' = \{M'_v\}_{v=0}^{\infty}$ is an admissible sequence, we get for $v \geqslant v_0$

$$d_n(id) \geqslant d_n(id_v) \gtrsim 2^{-v(s_1-s_2-\frac{d}{p_1}+\frac{d}{p_2})} \cdot M_v^{\frac{1}{p_2}-\frac{1}{p_1}}$$

for $n = \left[\frac{c}{2} \cdot M'_v\right]$, where c is the constant from Lemma 4.3. This leads to

$$d_n(id) \gtrsim n^{-\frac{s_1-s_2}{d}}, \quad n = \left[\frac{c}{2} \cdot M'_v\right], \quad v \geqslant v_0.$$

Again the monotonicity of the Kolmogorov numbers completes the proof.

Step 2: *Proof of* (♠) *and* (◇).

It follows from Lemma 4.5, that if $0 < p_1 < 1$ and $2 < p_2 \leqslant \infty$

$$d_n(\ell_{p_1}^m, \ell_{p_2}^m) = d_n(\ell_1^m, \ell_{p_2}^m), \quad 1 \leqslant n \leqslant m < \infty. \tag{4.10}$$

The proof of (♠) follows from (4.10), (4.2), Lemma (4.2) and the choice $n = \left[\frac{M'_\nu}{2}\right]$.

The proof of (◇) follows in the same way, but with $n = \left[\left(M'_\nu\right)^{\frac{2}{p_2}}\right]$.

Step 3: *Proof of* (♡).

We generalize the idea of Lemma 4.5 to p-Banach spaces, namely we show that for $0 < p_1 < p_2 \leqslant 2$

$$d_n(\ell_{p_1}^m, \ell_{p_2}^m) = d_n(\ell_{\min(1,p_2)}^m, \ell_{p_2}^m), \quad 1 \leqslant n \leqslant m < \infty. \tag{4.11}$$

If $p_2 \geqslant 1$, this follows immediately from Lemma 4.5. If $p_2 \leqslant 1$, we show that

$$d_n(\ell_{p_1}^m, \ell_{p_2}^m) \geqslant d_n(E_m, \ell_{p_2}^m) \geqslant d_n(\ell_{p_2}^m, \ell_{p_2}^m). \tag{4.12}$$

Here, $E_m = \{e_i\}_{i=1}^m \subset \mathbb{R}^m$ and $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ are the canonical unit vectors having all but one components 0 and the ith component 1.

Of course, (4.12) implies one half of (4.11), the second one being obvious. From (4.12), only the second inequality needs a proof. Let $N \subset\subset \ell_{p_2}^m = Y$ be such that

$$\sup_{i=1,\dots,n} \inf_{y \in N} \|e_i - y\|_{p_2} \leqslant (1+\varepsilon) d_n(E_m, \ell_{p_2}^m)$$

with $\dim N < n$. Hence, to every $e_i \in E_m$ there is an $f_i \in N$ such that

$$\|e_i - f_i\|_Y \leqslant (1+\varepsilon)^2 d_n(E_m, \ell_{p_2}^m).$$

To every $x \in \ell_{p_2}^m$, $x = \sum_{i=1}^m x_i e_i$ with $\sum_{i=1}^m |x_i|^{p_2} \leqslant 1$ we associate $\tilde{x}(x) = \sum_{i=1}^m x_i f_i \in N$. The estimate

$$\begin{aligned} d_n(id\colon \ell_{p_2}^m \to \ell_{p_2}^m)^{p_2} &\leqslant \sup_{\|x\|_{p_2} \leqslant 1} \inf_{y \in N} \|x - y\|_{p_2}^{p_2} \\ &\leqslant \sup_{\|x\|_{p_2} \leqslant 1} \|x - \tilde{x}(x)\|_{p_2}^{p_2} = \sup_{\|x\|_{p_2} \leqslant 1} \left\| \sum_{i=1}^m x_i(e_i - f_i) \right\|_{p_2}^{p_2} \\ &\leqslant \sup_{\|x\|_{p_2} \leqslant 1} \sum_{i=1}^m \|x_i(e_i - f_i)\|_{p_2}^{p_2} = \sup_{\|x\|_{p_2} \leqslant 1} \sum_{i=1}^m |x_i|^{p_2} \|e_i - f_i\|_{p_2}^{p_2} \\ &\leqslant \sup_{\|x\|_{p_2} \leqslant 1} \sum_{i=1}^m |x_i|^{p_2} (1+\varepsilon)^{2p_2} d_n(E_m, \ell_{p_2}^m)^{p_2} \\ &\leqslant (1+\varepsilon)^{2p_2} d_n(E_m, \ell_{p_2}^m)^{p_2} \end{aligned}$$

finishes the proof of (4.12).

The proof of (♡) follows in the same way as in the first and the second step. □

Now, we turn our attention to Gelfand numbers. First, we collect some information about $c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m)$, cf. [22], (4.2) and (4.3).

Lemma 4.7. *For* $1 \leqslant n \leqslant m < \infty$ *and* $1 \leqslant p_1, p_2 \leqslant \infty$, *we define*

$$\Phi(m,n,p_1,p_2) := \begin{cases} (m-n+1)^{\frac{1}{p_2}-\frac{1}{p_1}} & \text{if } 1 \leqslant p_2 \leqslant p_1 \leqslant \infty, \\ \left(\min\{1, m^{1-\frac{1}{p_1}} n^{-\frac{1}{2}}\}\right)^{\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{p_1}-\frac{1}{2}}} & \text{if } 1 < p_1 < p_2 \leqslant 2, \\ \max\{m^{\frac{1}{p_2}-\frac{1}{p_1}}, \sqrt{1-\frac{n}{m}}^{\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{2}-\frac{1}{p_2}}}\} & \text{if } 2 \leqslant p_1 < p_2 \leqslant \infty, \\ \max\{m^{\frac{1}{p_2}-\frac{1}{p_1}}, \\ \quad \min\{1, m^{1-\frac{1}{p_1}} n^{-\frac{1}{2}}\} \times \sqrt{1-\frac{n}{m}}\} & \text{if } 1 < p_1 \leqslant 2 < p_2 \leqslant \infty. \end{cases}$$

Then, *if* $p_1 > 1$,

$$c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m) \approx \Phi(m,n,p_1,p_2), \quad 1 \leqslant n \leqslant m < \infty.$$

Furthermore, *there are two constants* c_{p_2} *and* C_{p_2} *such that*

$$c_{p_2} \Psi(m,n,p_2) \leqslant c_n(id\colon \ell_1^m \to \ell_{p_2}^m) \leqslant C_{p_2} \Psi(m,n,p_2) \left(\log\left(\frac{em}{n}\right)\right)^{3/2},$$

where

$$\Psi(m,n,p_2) := \begin{cases} n^{1-\frac{1}{p_2}} & \text{if } 1 < p_2 \leqslant 2, \\ \min\left\{1, \max\left\{m^{1-\frac{1}{p_2}}, m^{-\frac{1}{2}}\sqrt{\frac{m}{n}-1}\right\}\right\} & \text{if } 2 \leqslant p_2 \leqslant \infty. \end{cases}$$

The proof of this lemma follows by (4.2) or (4.3) and Lemma 4.2.

Lemma 4.8. *If* $0 < p_2 \leqslant p_1 \leqslant \infty$, *then*

$$c_n(\ell_{p_1}^m, \ell_{p_2}^m) = (m-n+1)^{\frac{1}{p_2}-\frac{1}{p_1}}.$$

The proof of this lemma follows literally [44, Section 11.11.4].

Lemma 4.9. *Let* $0 < p < 1$. *Then there is a real constant* $c > 0$ *such that*

$$c_n(id\colon \ell_p^m \to \ell_2^m) \leqslant c \left[\frac{n}{\log\left(1+\frac{m}{n}\right)}\right]^{\frac{1}{2}-\frac{1}{p}}, \quad 1 \leqslant n \leqslant m < \infty.$$

Proof. This lemma slightly generalizes a result of Kashin [28], which was later improved by Gluskin [22] and Garnaev and Gluskin [20]. We closely follow the presentation given in [35, Chapter 14].

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a multivector, with $y_1, \ldots, y_n \in S^{m-1}$, the unit sphere of $\mathbb{R}^m$. We set

$$F_{m,n}(x, \mathbf{y}) = \frac{|(x, y_1)| + \cdots + |(x, y_n)|}{n}, \quad x \in \mathbb{R}^m.$$

We equip the space

$$\Sigma_{m,n} = \underbrace{S^{m-1} \times \cdots \times S^{m-1}}_{n \text{ times}}$$

with the natural rotation invariant probability measure P. Then (cf. [35, Lemma 4.1, Chapter 14]) we have the following.

Lemma 4.10. *For any $x \in S^{m-1}$ and $m, n \geqslant 2$*

$$P\left\{\mathbf{y} \in \Sigma_{m,n}\colon \frac{1}{8\sqrt{m}} \leqslant F(x, \mathbf{y}) \leqslant \frac{3}{\sqrt{m}}\right\} > \begin{cases} 1 - e^{-n}, & n > 2, \\ \frac{1}{2}, & n = 2. \end{cases}$$

Let l and m be natural numbers with $1 \leqslant l \leqslant m$. Let b_p^m denote the unit ball of ℓ_p^m. We denote by $b_p^{m,l}$ the subset of all vectors from b_p^m whose coordinates are of the form $\frac{k}{l}$, $k \in \mathbb{Z}$. Then there is a real constant $\tilde{c} > 0$ such that for any natural number $n \leqslant m$ with

$$l = \left\lceil \frac{1}{2\tilde{c}} \left(\frac{n}{\log\left(1 + \frac{m}{n}\right)} \right)^{1/p} \right\rceil \geqslant 1$$

there exists a multivector $\mathbf{y} = (y_1, \ldots, y_n)$ such that for all $x \in b_p^{m,l}$

$$\frac{1}{8\sqrt{m}} \|x\|_2 \leqslant F(x, \mathbf{y}) \leqslant \frac{3}{\sqrt{m}} \|x\|_2. \tag{4.13}$$

To prove it, we need to estimate the number of the elements of $b_p^{m,l}$ from above. It could be done directly, but we prefer to use known results. Observe that the mutual ℓ_∞^m distance of the points in $b_p^{m,l}$ is at least $\frac{1}{l}$. Hence, if $M_p^{m,l} = \#b_p^{m,l}$ (i.e. the number of elements of $b_p^{m,l}$) is greater than 2^n for some natural number n, then

$$e_n(id\colon \ell_p^m \to \ell_\infty^m) \geqslant \frac{1}{2l}. \tag{4.14}$$

But, according to [53] and [15, Section 3.2.2], there is a constant $\tilde{c}$ such that

$$e_n(id\colon \ell_p^m \to \ell_\infty^m) \leqslant \tilde{c} \left(\frac{\log\left(1 + \frac{m}{n}\right)}{n} \right)^{1/p}, \quad 1 \leqslant n \leqslant m. \tag{4.15}$$

Note that according to [65], this estimate is known to be even an equivalence if $\log m \leqslant n \leqslant m$.

From (4.14) and (4.15), it follows that if

$$\frac{1}{2l} > \tilde{c} \left(\frac{\log\left(1 + \frac{m}{n}\right)}{n} \right)^{1/p},$$

then $M_p^{m,l} \leqslant 2^n < e^n$. This, combined with Lemma 4.10 ensures the existence of the multivector $\mathbf{y}$.

Let $b_p^{m,l}$ be as above and let b_∞^m be a unit ball of ℓ_∞^m. Let $V_p^{m,l} = b_p^{m,l} \cap (\frac{1}{l} b_\infty^m)$ be the set of all vectors in $\mathbb{R}^m$ with the ℓ_p^m-quasi-norm at most one and with components in $\{0, \pm\frac{1}{l}\}$. Then we claim that

$$b_p^m \cap \left(\frac{1}{l} b_\infty^m\right) = \operatorname{conv}_p(V_p^{m,l}) \subset \operatorname{conv}(V_p^{m,l}), \tag{4.16}$$

where $\operatorname{conv}_p(V_p^{m,l})$ is the so-called p-convex hull of $V_p^{m,l}$. We refer to [18,19,25] for the notion of p-convexity, p-extreme points and the quasi-convex variant of the Krein–Milman theorem, which gives the identity in (4.16). The inclusion is a simple consequence of the fact that $p < 1$.

To prove Lemma 4.9, we need to find $N \subset\subset \mathbb{R}^m$ of codimension at most n such that for each point $x \in N \cap b_p^m$ we have $\|x\|_2 \leqslant c l^{\frac{p}{2}-1}$.

Let $\mathbf{y}$ be one multivector with (4.13). We set

$$N = \left\{x \in \mathbb{R}^m : F(x, \mathbf{y}) = 0\right\}.$$

Let $x \in N \cap b_p^m$ and let $x' \in b_p^{m,l}$ be the closest point to x, hence $\|x - x'\|_\infty \leqslant \frac{1}{l}$. We set $x'' = x - x'$. Then

$$\|x''\|_2 \leqslant \|x''\|_p^{\frac{p}{2}} \cdot \|x''\|_\infty^{1-\frac{p}{2}} \leqslant l^{\frac{p}{2}-1}. \tag{4.17}$$

It remains to estimate $\|x'\|_2$. This will be done by estimating the value of $F(x', \mathbf{y})$. The estimate

$$F(x', \mathbf{y}) \geqslant \frac{1}{8\sqrt{m}} \|x'\|_2 \tag{4.18}$$

follows from (4.13) and the fact that $x' \in b_p^{m,l}$. On the other hand, because of $x \in N$ and F is subadditive,

$$F(x', \mathbf{y}) \leqslant F(x, \mathbf{y}) + F(x'', \mathbf{y}) = F(x'', \mathbf{y}). \tag{4.19}$$

For all $\tilde{x} \in V_p^{m,l} \subset b_p^{m,l}$, we have

$$F(\tilde{x}, \mathbf{y}) \leqslant \frac{3}{\sqrt{m}} \|\tilde{x}\|_2 \leqslant 3 m^{-\frac{1}{2}} l^{\frac{p}{2}-1} \tag{4.20}$$

and by subadditivity of F and (4.16), the same holds also for $x'' \in b_p^m \cap \left(\frac{1}{l} b_\infty^m\right)$.

We insert (4.20) into (4.19) and (4.18) and get $\|x'\|_2 \leqslant 24 l^{\frac{p}{2}-1}$, and together with (4.17), $\|x\| \leqslant 25 l^{\frac{p}{2}-1}$. □

Following lemma follows from Lemma 4.9 by interpolation.

Lemma 4.11. *Let $0 < p_1 < 1$ and $p_1 < p_2 \leqslant \infty$. Then there is a real constant $c > 0$ such that*

$$c_n(id : \ell_{p_1}^m \to \ell_{p_2}^m) \leqslant c \left[\frac{n}{\log\left(1 + \frac{m}{n}\right)}\right]^{\frac{1}{\min(p_2,2)} - \frac{1}{p_1}}, \quad 1 \leqslant n \leqslant m < \infty.$$

Theorem 4.12. *Let* $-\infty < s_2 < s_1 < \infty$ *and* $0 < p_1, p_2, q_1, q_2 \leqslant \infty$ *with* (2.8). *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipschitz domain. Then* (2.7) *is compact and for* $n \in \mathbb{N}$

$$c_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}+\left(\frac{1}{p_1}-\frac{1}{p_2}\right)_+} \quad \textit{if} \begin{cases} \textit{either} & 2 \leqslant p_1 < p_2 \leqslant \infty, \\ \textit{or} & 0 < p_2 \leqslant p_1 \leqslant \infty, \end{cases} \tag{4.21}$$

$$c_n(\mathcal{I}d) \approx n^{-\frac{s_1-s_2}{d}} \quad \textit{if } 0 < p_1 < p_2 \leqslant 2 \quad \textit{and} \quad \frac{s_1-s_2}{d} > \frac{1}{2}\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{p_1}-\frac{1}{2}}, \tag{4.22}$$

$$c_n(\mathcal{I}d) \approx n^{\frac{p_1'}{2}\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{p_2}\right)} \quad \textit{if } 1 < p_1 < p_2 \leqslant 2 \quad \textit{and} \quad \frac{s_1-s_2}{d} < \frac{1}{2}\frac{\frac{1}{p_1}-\frac{1}{p_2}}{\frac{1}{p_1}-\frac{1}{2}}, \tag{4.23}$$

$$c_n(\mathcal{I}d) \approx n^{\left(-\frac{s_1-s_2}{d}+\frac{1}{2}-\frac{1}{p_2}\right)} \quad \textit{if } 0 < p_1 < 2 < p_2 \leqslant \infty \quad \textit{and} \quad \frac{s_1-s_2}{d} > 1-\frac{1}{p_2}, \tag{4.24}$$

$$c_n(\mathcal{I}d) \approx n^{\frac{p_1'}{2}\left(-\frac{s_1-s_2}{d}+\frac{1}{p_1}-\frac{1}{p_2}\right)} \quad \textit{if } 1 < p_1 < 2 < p_2 \leqslant \infty \quad \textit{and} \quad \frac{1}{p_1}-\frac{1}{p_2} < \frac{s_1-s_2}{d} < 1-\frac{1}{p_2}. \tag{4.25}$$

Proof. As Gelfand numbers are multiplicative and additive s-numbers, we may invoke (2.12) and restrict again to sequence spaces. Then, the method of the proof of Theorem 3.5 applies. The estimates on the sequence space side are given by Lemma 4.2 and (4.2). This approach finishes the proof in case $1 \leqslant p_1, p_2 \leqslant \infty$.

In the cases, when $p_1 < 1$ and/or $p_2 < 1$, (4.2) and (4.3) fail and Lemma 4.2 does not provide suitable estimates for $c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m)$. Hence, we are forced to treat these cases separately.

(♣) (4.21) if $0 < p_2 \leqslant p_1 \leqslant \infty$ and $0 < p_2 < 1$,

(♡) (4.22) if $0 < p_1 < p_2 \leqslant 2$ and $0 < p_1 < 1$,

(♠) (4.24) if $0 < p_1 < 1$ and $2 < p_2 \leqslant \infty$.

Step 1: *Proof of* (♣).

The proof of the estimate from below in (♣) follows exactly as in the proof of Theorem 4.6 with Lemma 4.3 replaced by Lemma 4.8.

The estimate from above in (♣) is provided by the corresponding statement about approximation numbers, cf. Theorem 3.5 and (4.1).

Step 2: *Proof of the estimates from below in* (♡) *and* (♠).

If $1 \leqslant p_2 \leqslant \infty$, we use the estimate

$$c_n(id\colon \ell_1^m \to \ell_{p_2}^m) \leqslant \|id\colon \ell_1^m \to \ell_{p_1}^m\| \cdot c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m) \tag{4.26}$$

and if $p_2 < 1$, we use the estimate

$$c_n(id\colon \ell_{p_2}^m \to \ell_{p_2}^m) \leqslant \|id\colon \ell_{p_2}^m \to \ell_{p_1}^m\| \cdot c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m). \tag{4.27}$$

This leads to

$$c_n(id\colon \ell_{p_1}^{2n} \to \ell_{p_2}^{2n}) \gtrsim \begin{cases} n^{\frac{1}{2}-\frac{1}{p_1}} & \text{if } 2\leqslant p_2\leqslant\infty, \\ n^{\frac{1}{p_2}-\frac{1}{p_1}} & \text{if } 0 < p_2\leqslant 2 \end{cases} \tag{4.28}$$

and the proof of the estimates from below included in (♡) and (♠) may be again finished as in the proof of Theorem 4.6.

Step 3: *Proof of the estimates from above in* (♡) *and* (♠).

Again, the knowledge of the behaviour of $c_n(id\colon \ell_{p_1}^m \to \ell_{p_2}^m)$ is of a crucial importance. Lemma 4.11 contains already the necessary information and the proof can be finished using the standard discretization method. □

5. Conclusion

In Theorems 3.5, 4.6 and 4.12 we gave an overview of the behaviour of approximation, Kolmogorov and Gelfand numbers of

$$\mathcal{I}d\colon B^{s_1}_{p_1q_1}(\Omega) \to B^{s_2}_{p_2q_2}(\Omega),$$

where Ω is a bounded domain in $\mathbb{R}^d$ with smooth (i.e. Lipschitz) boundary and the parameters satisfy

$$s_1 - s_2 > d\left(\frac{1}{p_1} - \frac{1}{p_2}\right)_+ .$$

The reader has surely noticed that all the obtained results about the asymptotic decay of $a_n(\mathcal{I}d)$, $d_n(\mathcal{I}d)$ and $c_n(\mathcal{I}d)$ do not depend on the fine parameters $0 < q_1, q_2\leqslant\infty$. This is of course no coincidence. The reason lies in the roots of the method we have used, namely in (3.7).

Nevertheless, the presented bounds from above and from below coincide in all "non-limiting" cases. Unfortunately, this method has also its natural bounds. For example, if $0 < p_1 < 2 < p_2\leqslant\infty$ and $s_1 - s_2 = d\max(1-\frac{1}{p_2}, \frac{1}{p_1})$, then Theorem 3.5 fails to characterize the decay of $a_n(\mathcal{I}d)$. One observes that in this case both (3.4) and (3.5) meet at $n^{-\frac{1}{2}}$, but (in general) this is not the exact speed of the decay of $a_n(\mathcal{I}d)$. It was shown by Kulanin [33], that additional logarithmic factors come into play. Their exact order seems to be unknown, but we believe that it depends on q_1 and q_2. So, for principle reasons, the decomposition method cannot be extended to this "limiting" case.

Using the elementary embeddings (2.4), we conclude that all the results hold for Triebel-Lizorkin spaces, Lebesgue spaces, Sobolev spaces, Bessel potential spaces and Hölder–Zygmund spaces as well.

For example, Theorem 3.5 may be stated in the framework of Bessel potential spaces and their embeddings into $C(\Omega)$ and $L_\infty(\Omega)$.

Theorem 5.1. *Let* $1\leqslant p\leqslant\infty$, $s > \frac{d}{p}$ *and let* $\Omega\subset\mathbb{R}^d$ *be a bounded Lipschitz domain. Then the embeddings*

$$\mathcal{I}d_1\colon H^s_p(\Omega) \to \mathcal{C}(\Omega), \tag{5.1}$$

$$\mathcal{I}d_2\colon H^s_p(\Omega) \to L_\infty(\Omega) \tag{5.2}$$

are compact and

$$
\begin{aligned}
&a_n(\mathcal{I}d_1) \approx a_n(\mathcal{I}d_2) \approx n^{-\frac{s}{d}+\frac{1}{p}} \quad \textit{if } 2 \leqslant p \leqslant \infty,\\
&a_n(\mathcal{I}d_1) \approx a_n(\mathcal{I}d_2) \approx n^{-\frac{s}{d}+\frac{1}{\tilde{p}}-\frac{1}{2}} \quad \textit{if } 0 < p < 2 \textit{ and } \frac{s}{d} > \frac{1}{\tilde{p}} = \max\left(1, \frac{1}{p}\right),\\
&a_n(\mathcal{I}d_1) \approx a_n(\mathcal{I}d_2) \approx n^{\left(-\frac{s}{d}+\frac{1}{p}\right)\cdot\frac{p'}{2}} \quad \textit{if } 1 < p < 2 \textit{ and } \frac{1}{p} < \frac{s}{d} < 1.
\end{aligned}
$$

Acknowledgment

I would like to thank to my colleagues from Jena, Aicke Hinrichs, Erich Novak, Winfried Sickel and Hans Triebel, for many valuable discussions on the topic.

Note added in proof

Recently, it was brought to our attention, that Gelfand numbers play an interesting role in Compressed Sensing. For example, our Lemma 4.9 covers the contents of Theorem 1 in [66].

References

[1] J. Bastero, J. Bernués, A. Peña, An extension of Milman's reverse Brunn–Minkowski inequality, Geom. Funct. Anal. 5 (3) (1995) 572–581.

[2] O. Besov, *Исследование одного семейства функциональных пространств в связи с теоремами вложения и продолжения*, Trudy. Mat. Inst. Steklova 60 (1961) 42–81. (Engl. transl.: Investigation of a family of function spaces in connection with theorems of imbedding and extension, Amer. Math. Soc. Transl. (2) 40 (1964) 85–126).

[3] M.S. Birman, M.Z. Solomyak, *Кусочно-полиномиальные приближения функций классов* W_p^α, Mat. Sbo. 73 (1967) 331–355 (Engl. transl.: Piecewise polynomial approximation of functions of the class W_p^α, Math. USSR Sb. 2 (1967) 295–317).

[4] A.M. Caetano, About approximation numbers in function spaces, J. Approx. Theory 94 (1998) 383–395.

[5] B. Carl, Entropy numbers, s-numbers, and eigenvalue problems, J. Funct. Anal. 41 (1981) 290–306.

[6] B. Carl, I. Stephani, Entropy, compactness and the approximation of operators, Cambridge Tracts in Mathematics, vol. 98, Cambridge University Press, Cambridge, 1990.

[7] Z. Ciesielski, T. Figiel, Construction of Schauder bases in function spaces on smooth compact manifolds. Approximation and function spaces, in: Proceedings of the International Conference, Gdansk, 1979, 1981, pp. 217–232.

[8] A. Cohen, Numerical Analysis of Wavelet Methods, Studies in Mathematics and its Applications, vol. 32, North-Holland, Elsevier, Amsterdam, 2003.

[9] A. Cohen, W. Dahmen, R. DeVore, Multiscale decompositions on bounded domains, Trans. Amer. Math. Soc. 352 (2000) 3651–3685.

[10] I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. 41 (1988) 909–996.

[11] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.

[12] R.A. DeVore, R.C. Sharpley, Besov spaces on Domains in $\mathbb{R}^d$, Trans. Amer. Math. Soc. 335 (1993) 843–864.

[13] S. Dispa, Intrinsic characterizations of Besov spaces on Lipschitz domains, Math. Nachr. 260 (2003) 21–33.

[14] S. Dispa, Intrinsic descriptions using means of differences for Besov spaces on Lipschitz domains, in: Function Spaces, Differential Operators and Nonlinear Analysis, Birkhäuser, Basel, Boston, Stuttgart, 2003, pp. 279–287.

[15] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, Differential Operators, Cambridge Tracts in Mathematics, vol. 120, Cambridge University Press, Cambridge, 1996.

[16] M. Frazier, B. Jawerth, A discrete transform and decompositions of distribution spaces, J. Funct. Anal. 93 (1) (1990) 34–170.

[17] M. Frazier, B. Jawerth, G. Weiss, Littlewood-Paley theory and the study of function spaces, in: Regional Conference Series, vol. 79, AMS, Providence, 1991.

[18] B. Fuchssteiner, Verallgemeinerte Konvexitätsbegriffe und der Satz von Krein–Milman, Math. Ann. 186 (1970) 149–154.

[19] B. Fuchssteiner, Verallgemeinerte Konvexitätsbegriffe und L^p-Räume, Math. Ann. 186 (1970) 171–176.

[20] A.Yu. Garnaev, E.D. Gluskin, *О поперечниках евклидова шара*,Dokl. Akad. Nauk SSSR 277 (1984) 1048 –1052 (Engl. transl.: On widths of the Euclidean ball, Soviet Math. Dokl. 30 (1984) 200–204).

[21] E.D. Gluskin, *О некоторых конечномерных задачах теории поперечников*, Vestnik Leningrad. Univ. Seria Mat. 13 (1981) 5–10 (Engl. transl.: On some finite-dimensional problems of in the theory of widths., Vestnik Leningrad Univ. Math. 14 (1982) 163–170).

[22] E.D. Gluskin, *Нормы влучайных матриц и поперечники конечномерных множеств*, Mat. Sb. 120 (1983) 180–189 (Engl. transl.: Norms of random matrices and widths of finite-dimensional sets, Math. USSR Sb. 48 (1984) 173–182).

[23] E. Hernández, G. Weiss, A First Course on Wavelets, CRC Press, Boca Raton, 1996.

[24] R.S. Ismagilov, *Поперечники множеств в линейных нормированных прос транствах и приближение функций тригонометрическими полиномами*, Uspekhi Mat. Nauk 29 (3) (1974) 161–178 (Engl. transl.: Diameters of sets in normed linear spaces and approximation of functions by trigonometric polynomials, Russian Math. Surveys 29(3) (1974) 169–186).

[25] N.J. Kalton, Compact p-convex sets, Quart. J. Math. Oxford (2) 28 (1977) 301–308.

[26] B.S. Kashin, *О колмогоровских поперечниках октаедров*, Dokl. Akad. Nauk SSSR 214 (1974) 1024–1026 (Engl. transl.: On Kolmogorov diameters of octahedra, Soviet Math. Dokl. 15 (1974) 304–307).

[27] B.S. Kashin, *О поперечниках октаедров* (The diameters of octahedra), Uspekhi Mat. Nauk 30 (4) (1975) 251–252.

[28] B.S. Kashin, *Поперечники некоторых конечномерных монжеств и классоб гладких функций*, Izv. Akad. Nauk, Seria Mat. 41 (1977) 334–351 (Engl. transl.: Diameters of some finite-dimensional sets and classes of smooth functions, Math. USSR, Izv. 11 (1977) 317–333).

[29] B.S. Kashin, *О поперечниках классов Соболева малой гладкости* (On the diameters of Sobolev classes of small smoothness), Vestnik Mosk. Univ. Seria Mat. 5 (1981) 50–54.

[30] A.N. Kolmogorov, Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse, Ann. of Math. 37 (1936) 107–110.

[31] V.I. Kondrashov, Sur certaines propriétés des fonctions dans l'espace L_p^ν, Doklady Akad. Nauk SSSR 48 (1945) 535–538.

[32] H. König, Eigenvalue Distribution of Compact Operators, Birkhäuser, Basel, Boston, Stuttgart, 1986.

[33] E.D. Kulanin, *О поперечниках класса функций ограниченной вариации в пространстве* $L^q(0, 1), 2 < q < \infty$, Uspekhi Mat. Nauk 38 (5) (1983) 191–192 (Engl. transl.: Diameters of a class of functions of bounded variation in the space $L^q(0, 1), 2 < q < \infty$, Russ. Math. Survey 38(5) (1983) 146–147).

[34] R. Linde, s-Numbers of diagonal operators and Besov embeddings, in: Proceedings of the 13th Winter School, Suppl. Rend. Circ. Mat. Palermo (1986).

[35] G.G. Lorentz, M.v. Golitschek, Y. Makovoz, Constructive approximation. Advanced problems, Grundlehren der Mathematischen Wissenschaften, vol. 304, Springer, Berlin, 1996.

[36] C. Lubitz, Weylzahlen von Diagonaloperatoren und Sobolev-Einbettungen, Dissertation, Rheinische Friedrich-Wilhelms-Universität, Bonn, 1982.

[37] V.E. Maĭorov, *Дискретизация задачи о поперечниках* (Discretization of the problem of diameters), Uspekhi Mat. Nauk 30 (6) (1975) 179–180.

[38] S.G. Mallat, Multiresolution approximation and wavelet orthonormal bases in $L_2(\mathbb{R})$, Trans. Amer. Math. Soc. 315 (1989) 69–87.

[39] Y. Meyer, Wavelets and Operators, Cambridge University Press, Cambridge, 1992.

[40] J. Peetre, New Thoughts on Besov Spaces, Duke University Mathematics Series, Duke University, Durham, 1976.

[41] A. Pietsch, Einige neue Klassen von kompakten linearen Operatoren, Rev. Math. Pures Appl. 8 (1963) 427–447.

[42] A. Pietsch, s-Numbers of operators in Banach spaces, Studia Math. 51 (1974) 201–223.

[43] A. Pietsch, Operator ideals, Deutsch. Verlag Wiss., Berlin, 1978; North-Holland, Amsterdam, London, New York, Tokyo, 1980.

[44] A. Pietsch, Eigenvalues and s-Numbers, Cambridge University Press, Cambridge, 1987.

[45] A. Pietsch, History of Banach Spaces and Linear Operators, Birkhäuser, Boston, Basel, Berlin, 2007.

[46] A. Pinkus, n-Widths in Approximation Theory, Ergebnisse der Mathematik und ihrer Grenzgebiete 3.7, Springer, Berlin, 1985.

[47] G. Pisier, The Volume of Convex Bodies and Banach Space Geometry, Cambridge Tracts in Mathematics, vol. 94, Cambridge University Press, Cambridge, 1989.

[48] F. Rellich, Ein Satz über mittlere Konvergenz, Nach. Wiss. Gesell. Göttingen, Math.-Phys. Kl. (1930) 30–35.

[49] S. Ropela, Spline bases in Besov spaces, Bull. Acad. Pol. Sci., S. Sci. Math. Astron. Phys. 24 (1976) 319–325.

[50] W. Rudin, Functional Analysis, second ed., McGraw-Hill, New York, St.Louis, San Francisco, 1991.

[51] T. Runst, W. Sickel, Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations, W. de Gruyter, Berlin, 1996.

[52] V.S. Rychkov, On restrictions and extensions of the Besov and Triebel-Lizorkin spaces with respect to Lipschitz domains, J. London Math. Soc. (2) 60 (1999) 237–257.

[53] C. Schütt, Entropy numbers of diagonal operators between symmetric Banach spaces, J. Approx. Theory 40 (1984) 121–128.

[54] L. Skrzypczak, Approximation and Entropy Numbers of Compact Sobolev Embeddings, Approximation and Probability, vol. 72, Banach Center Publications, Warszawa, 2006, pp. 309–326 (Papers of the Conference Held on the Occasion of the 70th anniversary of Prof. Zbigniew Ciesielski, Bedlewo, Poland, September 20–24, 2004).

[55] S.L. Sobolev, *Об одной теореме функционального анализа*, Mat. Sb. 4 (1938) 471–497 (Engl. transl.: On a theorem of functional analysis, Amer. Math. Soc. Transl. (2) 34 (1963) 39–68).

[56] S.B. Stechkin, *О наилучшем приближении заданных классов функций любыми полиномами* (On the best approximation of given classes of functions by arbitrary polynomials), Uspekhi Mat. Nauk 9 (1954) 133–134.

[57] V.M. Tikhomirov, *Поперечники множеств в функциональных пространствах и теория наилучших приближений*, Uspekhi Mat. Nauk 15 (3) (1960) 81–120 (Engl. title: Diameters of sets in function spaces and the theory of best approximations, Russ. Math. Survey 15(3) (1960) 75–111).

[58] H. Triebel, Interpolation Theory, Function Spaces, Differential Operators, North-Holland, Amsterdam, 1978.

[59] H. Triebel, Theory of Function Spaces, Geest & Portig, Leipzig, Birkhäuser, Basel, Boston, Berlin, 1983.

[60] H. Triebel, Theory of Function Spaces II, Birkhäuser, Basel, Boston, Berlin, 1992.

[61] H. Triebel, Theory of Function Spaces III, Birkhäuser, Basel, Boston, Berlin, 2006.

[62] H. Triebel, Function spaces and wavelets on domains, to appear.

[63] H. Triebel, Local Means and Wavelets in Function Spaces, Banach Center Publications, to appear.

[64] P. Wojtaszczyk, A Mathematical Introduction to Wavelets, London Mathematical Society Student Text, vol. 37, Cambridge University Press, Cambridge, 1997.

[65] T. Kühn, A lower estimate for entropy numbers, J. Approx. Theory 110 (1) (2001) 120–124.

[66] D.L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (4) (2006) 1289–1306.

Annales Academiæ Scientiarum Fennicæ
Mathematica
Volumen 34, 2009, 529–544

SOBOLEV AND JAWERTH EMBEDDINGS FOR SPACES WITH VARIABLE SMOOTHNESS AND INTEGRABILITY

Jan Vybíral

Universität Jena, Mathematisches Institut
Ernst-Abbe-Platz 2, 07740 Jena, Germany; vybiral@mathematik.uni-jena.de

Abstract. We consider the Triebel–Lizorkin spaces $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)$ of variable smoothness and integrability as introduced recently by Diening, Hästö and Roudenko in [9]. Under certain regularity conditions on the function parameters involved we show that

$$F^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}(\mathbf{R}^n)$$

if

$$s_0(x) \ge s_1(x) \ \text{ and } \ s_0(x) - \frac{n}{p_0(x)} = s_1(x) - \frac{n}{p_1(x)} \ \text{ for all } x \in \mathbf{R}^n$$

with embeddings of Sobolev and Bessel potential spaces included as special cases.

If $\inf_{x\in\mathbf{R}^n}(s_0(x) - s_1(x)) > 0$ we recover also the analogue of the Jawerth embedding

$$F^{s_0(\cdot)}_{p_0(\cdot),q_0(\cdot)}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q_1(\cdot)}(\mathbf{R}^n)$$

for any q_0, q_1.

The proofs are based on the decomposition techniques of [9] and work exclusively with the associated sequence spaces $f^{s(\cdot)}_{p(\cdot),q(\cdot)}$.

1. Introduction

The interplay between smoothness and integrability constitutes one of the corner stones of the theory of function spaces. It can be traced back as far as to Hardy and Littlewood [17, 18], but the decisive breakthrough was achieved by Sobolev [33], who proved the famous embedding

$$W^m_p(\Omega) \hookrightarrow L_q(\Omega), \tag{1.1}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded domain with Lipschitz boundary, $L_q(\Omega)$ stands for the usual Lebesgue space and $W^m_p(\Omega)$ denotes the Sobolev space of functions with all distributive derivatives of order smaller or equal to m bounded in the $L_p(\Omega)$ norm. The crucial relation between the involved parameters $m \in \mathbf{N}$, $1 < p < n/m$ and $1 < q < \infty$ is

$$\frac{1}{q} = \frac{1}{p} - \frac{m}{n}. \tag{1.2}$$

During the last seventy years, many scales of spaces of smooth functions were defined using various techniques (e.g. derivatives, differences, Fourier coefficients or Fourier transform) with the corresponding analogues of (1.1) and (1.2) playing usually an important role in most of the applications. Actually, it seems that any new scale

2000 Mathematics Subject Classification: Primary 46E35, 46E30.

Key words: Triebel–Lizorkin spaces, variable smoothness, variable integrability, Jawerth embedding, Sobolev embedding.

of spaces of smooth functions needs to exhibit some kind of interaction between smoothness and integrability to be accepted by the mathematical audience.

In recent years there has been a growing interest in function spaces describing local regularity properties of functions. The first spaces of this type are the spaces of variable integrability, which were introduced by Orlicz [27] already in 1931 and studied in detail by Kováčik and Rákosník [24] in 1991 together with the corresponding Sobolev spaces of variable integrability. During 1990's these spaces found applications in the study of variational integrals with non-standard growth, but it was probably the work of Ružička [29, 30, 31] on electrorheological fluids what promoted an enormous interest in these spaces. Since then, more than one hundred papers on this topic appeared. We refer to [8] for a brief overview and an extensive collection of references.

Another way how to describe the local properties of a function was outlined already by Peetre in [28, p. 266] in Chapter 12 named "Some strange new spaces" and resulted in the concept of 2-microlocal spaces, cf. [5] and [20]. Along a different line of study, Leopold [25] introduced spaces of Besov-type with variable smoothness, but constant integrability. This approach was further developed by Besov [3, 4].

The Sobolev embedding for the spaces with variable integrability was addressed already by Kováčik and Rákosník [24] and later on by Ružička [31]. But their results failed to cover the optimal exponent according to (1.1). Edmunds and Rákosník [10, 11] proved the optimal Sobolev embedding theorem under Lipschitz and Hölder continuity of the exponents, cf. also [13]. Finally, Diening [7] and Samko [32] showed, that log-Hölder continuity is sufficient.

The embeddings of Besov and Triebel–Lizorkin spaces of variable smoothness were obtained by Besov [4] in a fairly general form. It seems that Leopold [26] was the only one up to now who tried to connect the function spaces with variable smoothness with spaces of variable integrability. Unfortunately, he also failed to recover the optimal exponent.

The last step (up to now) was done by Diening, Hästö and Roudenko in [9]. These authors combined the concept of spaces with variable integrability of Orlicz, Kováčik and Rákosník with the concept of variable smoothness of Leopold and Besov (which is in some sense very similar to the ideas of Peetre, Bony and Jaffard) and proposed the function spaces of Triebel–Lizorkin type of variable smoothness *and* integrability, cf. Definition 2.5. They proved (under some restrictions on the function parameters involved), that these spaces include the Lebesgue and Sobolev spaces of variable integrability and the spaces of variable smoothness as special cases. They proved also a certain version of the atomic decomposition theorem, which is a well known tool in the theory of function spaces of Besov and Triebel–Lizorkin type. Finally, they proved an analogue of the usual trace theorem, which exhibits the interplay between smoothness and integrability. The reader may consult also [12], [19] and references given there for other versions of the trace embedding theorem for Sobolev spaces with varying integrability.

Although mentioned on several places in [9] (and even in the abstract), the authors have not presented any version of Sobolev embedding, which would not only result in a generalization of (1.1) with (1.2) holding pointwise, but would (in the sense described above) help to justify the existence of this scale of function spaces—at least until this promising line of research finds any applications.

Our aim is to fill this gap. In the frame of Triebel–Lizorkin spaces with constant parameters, the following analogue of Sobolev embedding is true.

Theorem 1.1. (Jawerth, [21]) *Let*

$$-\infty < s_1 < s_0 < \infty, \quad 0 < p_0 < p_1 < \infty, \quad 0 < q \le \infty \tag{1.3}$$

with

$$s_0 - \frac{n}{p_0} = s_1 - \frac{n}{p_1}. \tag{1.4}$$

Then

$$F^{s_0}_{p_0,\infty}(\mathbf{R}^n) \hookrightarrow F^{s_1}_{p_1,q}(\mathbf{R}^n). \tag{1.5}$$

The remarkable effect, which was first observed by Jawerth and which is in some sense unique to the Triebel–Lizorkin spaces, is the improvement in the third fine parameter $q > 0$, which may be chosen arbitrarily small. Of course, (1.5) holds only for $q = \infty$ if $s_0 = s_1$ (or, equivalently, $p_0 = p_1$). If the smoothness and integrability parameters s and p become functions of $x \in \mathbf{R}^n$, then it seems to be appropriate to assume that (1.4) holds pointwise, i.e.,

$$s_0(x) - \frac{n}{p_0(x)} = s_1(x) - \frac{n}{p_1(x)}, \quad x \in \mathbf{R}^n \tag{1.6}$$

and if the improvement in the fine parameter is to be achieved, that also

$$\inf_{x\in\mathbf{R}^n}(s_0(x) - s_1(x)) = n \inf_{x\in\mathbf{R}^n}\Big(\frac{1}{p_0(x)} - \frac{1}{p_1(x)}\Big) > 0. \tag{1.7}$$

We prove that these "natural" assumptions (combined with appropriate regularity conditions) are really sufficient. We show, that if $s_1(x) \le s_0(x)$ and $p_0(x) \le p_1(x)$ with (1.6) and $0 < q(x) \le \infty$ for all $x \in \mathbf{R}^n$, then

$$F^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}(\mathbf{R}^n). \tag{1.8}$$

If also (1.7) is satisfied, then even

$$F^{s_0(\cdot)}_{p_0(\cdot),\infty}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}(\mathbf{R}^n)$$

holds.

2. Preliminaries

Let $S(\mathbf{R}^n)$ be the Schwartz space of all complex-valued rapidly decreasing, infinitely differentiable functions on $\mathbf{R}^n$ and let $S'(\mathbf{R}^n)$ be its dual—the space of all tempered distributions. For $f \in S'(\mathbf{R}^n)$ we denote by $\widehat{f} = Ff$ its Fourier transform and by $f^\vee$ or $F^{-1}f$ its inverse Fourier transform. We give a Fourier-analytic definition of Triebel–Lizorkin spaces, which relies on the so-called *dyadic resolution of unity.* Let $\varphi \in S(\mathbf{R}^n)$ with

$$\varphi(x) = 1 \quad \text{if} \quad |x| \le 1 \quad \text{and} \quad \varphi(x) = 0 \quad \text{if} \quad |x| \ge \frac{3}{2}. \tag{2.1}$$

We put $\varphi_0 = \varphi$ and $\varphi_j(x) = \varphi(2^{-j}x) - \varphi(2^{-j+1}x)$ for $j \in \mathbf{N}$ and $x \in \mathbf{R}^n$. This leads to the identity

$$\sum_{j=0}^{\infty} \varphi_j(x) = 1, \quad x \in \mathbf{R}^n.$$

Definition 2.1. Let $s \in \mathbf{R}$, $0 < p < \infty$, $0 < q \leq \infty$. Then $F^s_{pq}(\mathbf{R}^n)$ is the collection of all $f \in S'(\mathbf{R}^n)$ such that

$$(2.2) \qquad ||f|F^s_{pq}(\mathbf{R}^n)|| = \left\| \left(\sum_{j=0}^{\infty} 2^{jsq} |(\varphi_j \widehat{f})^\vee(\cdot)|^q \right)^{1/q} |L_p(\mathbf{R}^n) \right\| < \infty$$

(with the usual modification for $q = \infty$).

Remark 2.2. (i) These spaces have a long history. In this context we recommend [28, 34, 35, 37] as standard references. We point out that the spaces $F^s_{pq}(\mathbf{R}^n)$ are independent of the choice of φ in the sense of equivalent (quasi-)norms. Special cases of this scale include Lebesgue spaces, Sobolev spaces and inhomogeneous Hardy spaces.

(ii) Interchanging the order of L_p and ℓ_q norm in (2.2) would lead to the Fourier-analytic definition of Besov spaces. Unfortunately, they seem to be less convenient for describing local regularity properties of distributions, because they lack the so-called *localization principle*, cf. [35, Theorem 2.4.7]. Hence (also in correspondence with [9]) we study only the F-scale.

Next, we introduce the Lebesgue spaces of variable integrability.

Definition 2.3. Let $p\colon \mathbf{R}^n \to (0, \infty)$ be a measurable function. Then the space $L_{p(\cdot)}(\mathbf{R}^n)$ consists of all measurable functions $f\colon \mathbf{R}^n \to [-\infty, \infty]$ such that $||f|L_{p(\cdot)}(\mathbf{R}^n)|| < \infty$, where

$$||f|L_{p(\cdot)}(\mathbf{R}^n)|| = \inf\{\lambda > 0 : \int_{\mathbf{R}^n} \left(\frac{|f(x)|}{\lambda} \right)^{p(x)} dx \leq 1\}$$

is the Minkowski functional of the absolutely convex set $\{f : \int_{\mathbf{R}^n} |f(x)|^{p(x)}\, dx \leq 1\}$.

Remark 2.4. (i) One could also consider (and it was done so already by Kováčik and Rákosník in [24]) that $p(x) = \infty$ on a set of a positive measure. But Definition 2.3 is already sufficient for our purpose, cf. also Remark 2.6.

(ii) If $p(x) \geq 1$ for all $x \in \mathbf{R}^n$, then $L_{p(\cdot)}(\mathbf{R}^n)$ are Banach spaces. To ensure, that $L_{p(\cdot)}(\mathbf{R}^n)$ are at least quasi-Banach spaces, we assume that

$$p^- := \inf_{x \in \mathbf{R}^n} p(x) > 0.$$

The generalization of Definition 2.1 to the setting of variable smoothness and integrability as it was given by [9] is surprisingly simple.

Definition 2.5. Let $-\infty < s(x) < +\infty$, $0 < p(x) < \infty$, $0 < q(x) \leq \infty$. Then $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)$ is the collection of all $f \in S'(\mathbf{R}^n)$ such that

$$(2.3) \qquad ||f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)|| = \left\| \left(\sum_{j=0}^{\infty} 2^{js(\cdot)q(\cdot)} |(\varphi_j \widehat{f})^\vee(\cdot)|^{q(\cdot)} \right)^{1/q(\cdot)} |L_{p(\cdot)}(\mathbf{R}^n) \right\| < \infty$$

(with the usual modification for $q(x) = \infty$).

Remark 2.6. This definition introduces the Triebel–Lizorkin spaces of variable smoothness, integrability and summability under almost no conditions on $s(\cdot)$, $p(\cdot)$ and $q(\cdot)$. Unfortunately, these spaces may depend on the choice of the function φ as described in (2.1). This is the case already when s and $q < \infty$ are constant and

$p = \infty$. We refer to [34, Chapter 2.3.4] for a detailed discussion of related aspects. So, a first natural restriction seems to be the condition

$$p^+ = \sup_{x\in\mathbf{R}^n} p(x) < \infty.$$

Together with Remark 2.4(ii) this leads to

$$0 < p^- := \inf_{z\in\mathbf{R}^n} p(z) \le p(x) \le \sup_{z\in\mathbf{R}^n} p(z) =: p^+ < \infty, \quad x \in \mathbf{R}^n. \tag{2.4}$$

Next we present the regularity assumptions of [9].

Definition 2.7. Let g be a continuous function on $\mathbf{R}^n$.

(i) We say, that g is *1-locally* log-*Hölder continuous*, abbreviated $g \in C^{\log}_{1-\mathrm{loc}}(\mathbf{R}^n)$, if there exists $c > 0$ such that

$$|g(x) - g(y)| \le \frac{c}{\log(e + 1/||x-y||_\infty)} \quad \text{for all } x, y \in \mathbf{R}^n \text{ with } ||x-y||_\infty \le 1.$$

Here, $||z||_\infty = \max\{|z_1|, \dots, |z_n|\}$ denotes the maximum norm of $z \in \mathbf{R}^n$.

(ii) We say, that g is *locally* log-*Hölder continuous*, abbreviated $g \in C^{\log}_{\mathrm{loc}}(\mathbf{R}^n)$, if there exists $c > 0$ such that

$$|g(x) - g(y)| \le \frac{c}{\log(e + 1/|x-y|)}, \quad x, y \in \mathbf{R}^n.$$

(iii) We say, that g is *globally* log-*Hölder continuous*, abbreviated $g \in C^{\log}(\mathbf{R}^n)$, if it is locally log-Hölder continuous and there exists $c > 0$ and $g_\infty \in \mathbf{R}$ such that

$$|g(x) - g_\infty| \le \frac{c}{\log(e + |x|)}, \quad x \in \mathbf{R}^n.$$

Remark 2.8. (i) The conditions (ii) and (iii) are overtaken literally from [9] and we shall need them only for the transference of our results from sequence spaces to function spaces. It is the less restrictive condition (i), which we shall involve in our proofs.

(ii) The condition (i) is very similar to the original condition of Diening used in [6] to show the boundedness of the maximal operator.

We shall use the property (i) in the form formulated in next Lemma. We leave out the trivial proof.

Lemma 2.9. *Let $g \in C^{\log}_{1-\mathrm{loc}}(\mathbf{R}^n)$. Then there exists a constant $c > 0$ such that for every $j \in \mathbf{N}_0$ and every $x, y \in \mathbf{R}^n$ with $||x-y||_\infty \le 2^{-j}$ the following inequalities hold:*

$$\frac{1}{c} \le 2^{-j|g(x)-g(y)|} \le 2^{j(g(x)-g(y))} \le 2^{j|g(x)-g(y)|} \le c.$$

Definition 2.10. (Standing assumptions of [9]) Let p and q be positive functions on $\mathbf{R}^n$ such that $\frac{1}{p}, \frac{1}{q} \in C^{\log}(\mathbf{R}^n)$ and let $s \in C^{\log}_{\mathrm{loc}}(\mathbf{R}^n) \cap L^\infty(\mathbf{R}^n)$ with $s(x) \ge 0$ and let $s(x)$ have a limit at infinity.

Remark 2.11. (i) Let us note, that the *standing assumptions* imply in particular (2.4) and a similar chain of inequalities for $q(x)$.

We introduce the sequence spaces associated with the Triebel–Lizorkin spaces of variable smoothness and integrability. Let $j \in \mathbf{N}_0$ and $m \in \mathbf{Z}^n$. Then Q_{jm} denotes the closed cube in $\mathbf{R}^n$ with sides parallel to the coordinate axes, centered at $2^{-j}m$, and with side length 2^{-j}. By $\chi_{jm} = \chi_{Q_{jm}}$ we denote the characteristic function of Q_{jm}. If

$$\gamma = \{\gamma_{jm} \in \mathbf{C} : j \in \mathbf{N}_0, m \in \mathbf{Z}^n\},$$

$-\infty < s(x) < \infty$, $0 < p(x) < \infty$ and $0 < q(x) \leq \infty$ for all $x \in \mathbf{R}^n$, we define

$$\begin{aligned} ||\gamma|f^{s(\cdot)}_{p(\cdot),q(\cdot)}|| &= \left|\left|\left(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{js(\cdot)q(\cdot)}|\gamma_{jm}|^{q(\cdot)}\chi_{jm}(\cdot)\right)^{1/q(\cdot)}|L_{p(\cdot)}(\mathbf{R}^n)\right|\right| \\ &= \left|\left|\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{js(\cdot)}|\gamma_{jm}|\chi_{jm}(\cdot)\right|L_{p(\cdot)}(\ell_{q(\cdot)})\right|\right|. \end{aligned} \tag{2.5}$$

Establishing the connection between the function spaces $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)$ and the sequence spaces $f^{s(\cdot)}_{p(\cdot),q(\cdot)}$ was the main aim of [9]. Following [14] and [15], these authors investigated the properties of the so-called φ-transform (denoted by S_φ) and obtained the following result.

Theorem 2.12. *Under the standing assumptions of* [9]

$$||f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)|| \approx ||S_\varphi f|f^{s(\cdot)}_{p(\cdot),q(\cdot)}||$$

with constants independent of $f \in F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)$.

Remark 2.13. (i) The assumptions on s in the Theorem 2.12 seem to be too restrictive. It seems, that several authors now try to prove similar results also for $s(x)$, which are not necessarily positive or convergent at infinity. We refer at least to [23] and [39].

From this reason we formulate the theorems of embeddings of sequence spaces under minimal assumptions, which shall really be needed in the proof. If later on any improved version of Theorem 2.12 should appear, the results may then be easily taken over.

(ii) We shall use only a corollary of Theorem 2.12, namely that (under the *standing assumptions*) the space $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbf{R}^n)$ is isomorphic to a subspace of $f^{s(\cdot)}_{p(\cdot),q(\cdot)}$ via the S_φ transform.

3. Main results

First, we state the results in the form of embeddings of sequence spaces under those assumptions really needed in the proof. Later on, we combine those with the *standing assumptions of* [9] and obtain similar results also for the embeddings of function spaces. Finally, we state separately the embeddings of Sobolev and Bessel potential spaces.

Theorem 3.1. *Let* $-\infty < s_1(x) \leq s_0(x) < \infty$, $0 < p_0(x) \leq p_1(x) < \infty$ *for all* $x \in \mathbf{R}^n$ *with* $0 < p_0^- \leq p_1^- \leq p_1^+ < \infty$ *and*

$$s_0(x) - \frac{n}{p_0(x)} = s_1(x) - \frac{n}{p_1(x)}, \quad x \in \mathbf{R}^n.$$

Let $q(x) = \infty$ *for all* $x \in \mathbf{R}^n$ *or* $0 < q^- \leq q(x) < \infty$ *for all* $x \in \mathbf{R}^n$ *and* $s_0, \frac{1}{p_0} \in C^{\log}_{1-\mathrm{loc}}(\mathbf{R}^n)$. *Then*

$$f^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}.$$

Proof. Step 1. $q(x)=\infty$ for all $x\in\mathbf{R}^n$. We set

$$h(x)=\sup_{j,m}2^{js_0(x)}|\gamma_{jm}|\chi_{jm}(x),\quad x\in\mathbf{R}^n. \tag{3.1}$$

Here, and later on, the supremum is taken over all $j\in\mathbf{N}_0$ and $m\in\mathbf{Z}^n$. Then by (2.5),

$$||\gamma|f^{s_0(\cdot)}_{p_0(\cdot),\infty}||=||h|L_{p_0(\cdot)}(\mathbf{R}^n)|| \tag{3.2}$$

and trivially

$$2^{js_0(x)}|\gamma_{jm}|\le h(x),\quad x\in Q_{jm}, \tag{3.3}$$

which leads to

$$|\gamma_{jm}|\le\inf_{x\in Q_{jm}}2^{-js_0(x)}h(x),\quad j\in\mathbf{N}_0,\ m\in\mathbf{Z}^n. \tag{3.4}$$

Using consequently (2.5), (3.4) and Lemma 2.9 for s_0 we may estimate

$$\begin{aligned}||\gamma|f^{s_1(\cdot)}_{p_1(\cdot),\infty}||&=\left\|\sup_{j,m}2^{js_1(x)}|\gamma_{jm}|\chi_{jm}(x)|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\&\le\left\|\sup_{j,m}2^{js_1(x)}\Big(\inf_{y\in Q_{jm}}2^{-js_0(y)}h(y)\Big)\chi_{jm}(x)|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\&=\left\|\sup_{j,m}2^{j(s_1(x)-s_0(x))}\Big(\inf_{y\in Q_{jm}}2^{j(s_0(x)-s_0(y))}h(y)\Big)\chi_{jm}(x)|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\&\le c\left\|\sup_{j,m}2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)}\Big(\inf_{y\in Q_{jm}}h(y)\Big)\chi_{jm}(x)|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|.\end{aligned}$$

Let $A_{-1}\subset\mathbf{R}^n$ stand for those x, where

$$\sup_{j,m}2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)}\Big(\inf_{y\in Q_{jm}}h(y)\Big)\chi_{jm}(x)=0. \tag{3.5}$$

For each $x\in\mathbf{R}^n\setminus A_{-1}$ we denote by $J=J_x\in\mathbf{N}_0$ the smallest non-negative integer such that

$$\begin{aligned}&\sup_{j,m}2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)}\Big(\inf_{y\in Q_{jm}}h(y)\Big)\chi_{jm}(x)\\&\le 2\cdot 2^{Jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)}\sum_{m\in\mathbf{Z}^n}\Big(\inf_{y\in Q_{Jm}}h(y)\Big)\chi_{Jm}(x).\end{aligned} \tag{3.6}$$

We may assume, that for almost all $x\in\mathbf{R}^n$ the left-hand side of (3.5) is finite. Otherwise $h(x)=\infty$ on a set of positive measure and there is nothing to prove. Furthermore, we denote by $A_J\subset\mathbf{R}^n$ those x with $J_x=J\in\mathbf{N}_0$.

Let $\lambda>0$ be a positive real number such that

$$\begin{aligned}1\ge\int_{\mathbf{R}^n}\left(\frac{h(x)}{\lambda}\right)^{p_0(x)}dx&=\sum_{J=-1}^{\infty}\int_{A_J}\left(\frac{h(x)}{\lambda}\right)^{p_0(x)}dx\\&\ge\sum_{J=0}^{\infty}\sum_{m\in\mathbf{Z}^n}\int_{A_J\cap Q_{Jm}}\left(\frac{h(x)}{\lambda}\right)^{p_0(x)}dx.\end{aligned} \tag{3.7}$$

We set

$$h_{jm} := \frac{\inf\limits_{y\in Q_{jm}} h(y)}{\lambda}, \quad j \in \mathbf{N}_0, \ \ m \in \mathbf{Z}^n$$

and show, that there is a constant $C > 0$ such that

$$\int_{\mathbf{R}^n} \left(C^{-1} \sup_{j,m} 2^{jn\left(\frac{1}{p_1(x)} - \frac{1}{p_0(x)}\right)} h_{jm}\chi_{jm}(x) \right)^{p_1(x)} dx \le 1.$$

We split the integration over $\mathbf{R}^n$ into integrals over A_J and use (3.6).

$$\begin{aligned} &\int_{\mathbf{R}^n} \left(C^{-1} \sup_{j,m} 2^{jn\left(\frac{1}{p_1(x)} - \frac{1}{p_0(x)}\right)} h_{jm}\chi_{jm}(x) \right)^{p_1(x)} dx \\ &\le \sum_{J=0}^{\infty} \int_{A_J} \left((C/2)^{-1} \sum_{m\in\mathbf{Z}^n} 2^{Jn\left(\frac{1}{p_1(x)} - \frac{1}{p_0(x)}\right)} h_{Jm}\chi_{Jm}(x) \right)^{p_1(x)} dx \\ &= \sum_{J=0}^{\infty} \sum_{m\in\mathbf{Z}^n} \int_{A_J} \left((C/2)^{-1} 2^{Jn\left(\frac{1}{p_1(x)} - \frac{1}{p_0(x)}\right)} h_{Jm} \right)^{p_1(x)} \chi_{Jm}(x)\, dx \\ &= \sum_{J=0}^{\infty} \sum_{m\in\mathbf{Z}^n} \int_{A_J\cap Q_{Jm}} (C/2)^{-p_1(x)} 2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)} h_{Jm}^{p_1(x)} dx \end{aligned} \tag{3.8}$$

Let us fix $(J,m) \in \mathbf{N}_0 \times \mathbf{Z}^n$. We shall distinguish two cases.

1. case: $h_{Jm} \le 1$. Then (as $p_0(x) \le p_1(x)$)

$$2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)} \le 1$$

and

$$h_{Jm}^{p_1(x)} \le h_{Jm}^{p_0(x)}.$$

Hence for $C \ge 2$ we obtain

$$\begin{aligned} \int_{A_J\cap Q_{Jm}} (C/2)^{-p_1(x)} 2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)} h_{Jm}^{p_1(x)} dx &\le \int_{A_J\cap Q_{Jm}} h_{Jm}^{p_0(x)} dx \\ &\le \int_{A_J\cap Q_{Jm}} \left(\frac{h(x)}{\lambda}\right)^{p_0(x)} dx. \end{aligned} \tag{3.9}$$

2. case: $h_{Jm} > 1$. Then

$$1 \ge \int_{Q_{Jm}} \left(\frac{h(x)}{\lambda}\right)^{p_0(x)} dx \ge \int_{Q_{Jm}} h_{Jm}^{p_0(x)} dx \ge 2^{-Jn} h_{Jm}^{p_0^{Jm}},$$

where $p_0^{Jm} = \inf\limits_{x\in Q_{Jm}} p_0(x) > 0$. Hence

$$1 < h_{Jm} \le 2^{Jn/p_0^{Jm}}. \tag{3.10}$$

We rewrite the integrals in (3.8) as

$$\begin{aligned} &\int_{A_J\cap Q_{Jm}} (C/2)^{-p_1(x)} 2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)} h_{Jm}^{p_1(x)} dx \\ &= \int_{A_J\cap Q_{Jm}} \underbrace{(C/2)^{-p_1(x)} 2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)} h_{Jm}^{p_1(x)-p_0(x)}}_{(\star)} h_{Jm}^{p_0(x)} dx \end{aligned} \tag{3.11}$$

and show that the estimate $(\star) \le 1$ for $C \ge 2$ large enough and $x \in Q_{Jm}$ finishes immediately the proof. By (3.9) and (3.11) combined with $(\star) \le 1$ and (3.7)

$$\begin{aligned}&\sum_{J=0}^{\infty}\sum_{m\in\mathbf{Z}^n}\int_{A_J\cap Q_{Jm}}(C/2)^{-p_1(x)}2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)}h_{Jm}^{p_1(x)}dx = \sum_{(J,m):h_{Jm}\le 1}\cdots + \sum_{(J,m):h_{Jm}>1}\cdots\\ &\le \sum_{(J,m):h_{Jm}\le 1}\int_{A_J\cap Q_{Jm}}\left(\frac{h(x)}{\lambda}\right)^{p_0(x)}dx + \sum_{(J,m):h_{Jm}>1}\int_{A_J\cap Q_{Jm}}h_{Jm}^{p_0(x)}dx\\ &\le \sum_{J=0}^{\infty}\sum_{m\in\mathbf{Z}^n}\int_{A_J\cap Q_{Jm}}\left(\frac{h(x)}{\lambda}\right)^{p_0(x)}dx \le 1.\end{aligned}$$

Hence, it remains to prove that $(\star) \le 1$ for all $x \in Q_{Jm}$. By (3.10), it is enough to show that

$$(C/2)^{-p_1(x)}2^{Jn\left(1-\frac{p_1(x)}{p_0(x)}\right)}\cdot 2^{Jn\cdot\frac{p_1(x)-p_0(x)}{p_0^{Jm}}} \le 1$$

or, equivalently,

$$2^{Jn[p_1(x)-p_0(x)]\cdot[\frac{1}{p_0^{Jm}}-\frac{1}{p_0(x)}]} \le (C/2)^{p_1(x)}.$$

Using Lemma 2.9 for $\frac{1}{p_0}$ (with constant $2^{c_{\log}}$), this follows from

$$2^{n[1-\frac{p_0(x)}{p_1(x)}]\cdot c_{\log}} \le C/2.$$

As $0 \le 1-\frac{p_0(x)}{p_1(x)} \le 1$, we may choose $C = 2^{nc_{\log}+1} \ge 2$.

Step 2. $0 < q(x) < \infty$ for all $x \in \mathbf{R}^n$. Let $\lambda > 0$ be a positive real number with

$$\int_{\mathbf{R}^n}\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n}2^{js_0(x)q(x)}|\gamma_{jm}|^{q(x)}\lambda^{-q(x)}\chi_{jm}(x)\Big)^{p_0(x)/q(x)}dx \le 1. \tag{3.12}$$

We have to show that there is a constant $C > 0$ independent of $\{\gamma_{jm}\}$, such that

$$\int_{\mathbf{R}^n}\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n}2^{js_1(x)q(x)}|\gamma_{jm}|^{q(x)}(C\lambda)^{-q(x)}\chi_{jm}(x)\Big)^{p_1(x)/q(x)}dx \le 1. \tag{3.13}$$

We show, that under (3.12) the following inequality holds for almost all $x \in \mathbf{R}^n$

$$\begin{aligned}&\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n}2^{js_1(x)q(x)}\frac{|\gamma_{jm}|^{q(x)}}{(C\lambda)^{q(x)}}\chi_{jm}(x)\Big)^{p_1(x)}\\ &\le \Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n}2^{js_0(x)q(x)}\frac{|\gamma_{jm}|^{q(x)}}{\lambda^{q(x)}}\chi_{jm}(x)\Big)^{p_0(x)}.\end{aligned} \tag{3.14}$$

Obviously, (3.14) implies (3.13).

For almost every $x \in \mathbf{R}^n$ and every $j \in \mathbf{N}_0$, there is exactly one $m = m(j) \in \mathbf{Z}^n$ such that $x \in Q_{j,m(j)}$. We fix one such an x. Then (3.14) reads like

$$\begin{aligned}&\sum_{j=0}^{\infty}2^{js_1(x)q(x)}|\gamma_{j,m(j)}|^{q(x)}(C\lambda)^{-q(x)}\\ &\le \Big(\sum_{j=0}^{\infty}2^{js_0(x)q(x)}|\gamma_{j,m(j)}|^{q(x)}\lambda^{-q(x)}\Big)^{p_0(x)/p_1(x)}.\end{aligned} \tag{3.15}$$

We set

$$\alpha_j := 2^{js_0(x)}\frac{|\gamma_{j,m(j)}|}{\lambda},\quad j\in\mathbf{N}_0$$

and rewrite (3.15) once again. It now becomes

$$\sum_{j=0}^\infty 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)}(\alpha_j/C)^{q(x)} \le \Big(\sum_{j=0}^\infty \alpha_j^{q(x)}\Big)^{p_0(x)/p_1(x)}. \tag{3.16}$$

Using (3.12) and Lemma 2.9 for s_0, we get

$$\begin{aligned}
1 &\ge \int_{Q_{j,m(j)}} \Big(2^{js_0(y)q(y)}|\gamma_{j,m(j)}|^{q(y)}\lambda^{-q(y)}\Big)^{p_0(y)/q(y)}dy\\
&= \int_{Q_{j,m(j)}} \Big(2^{js_0(y)}|\gamma_{j,m(j)}|\lambda^{-1}\Big)^{p_0(y)}dy\\
&= \int_{Q_{j,m(j)}} \Big(2^{j(s_0(y)-s_0(x))}2^{js_0(x)}|\gamma_{j,m(j)}|\lambda^{-1}\Big)^{p_0(y)}dy\\
&\ge \int_{Q_{j,m(j)}} \Big(c\,2^{js_0(x)}|\gamma_{j,m(j)}|\lambda^{-1}\Big)^{p_0(y)}dy\\
&= \int_{Q_{j,m(j)}} \big(c\,\alpha_j\big)^{p_0(y)}dy.
\end{aligned}$$

If $c\alpha_j > 1$, we may further estimate

$$1 \ge 2^{-jn}\big(c\,\alpha_j\big)^{\inf_{z\in Q_{j,m(j)}} p_0(z)},$$

or, equivalently,

$$c\,\alpha_j \le 2^{\frac{jn}{\inf_{z\in Q_{j,m(j)}} p_0(z)}} = 2^{\frac{jn}{p_0(x)}}2^{\frac{jn}{\inf_{z\in Q_{j,m(j)}} p_0(z)}-\frac{jn}{p_0(x)}} \le c'2^{\frac{jn}{p_0(x)}} \tag{3.17}$$

and this estimate holds true also if $c\,\alpha_j \le 1$.

If $\sum_{j=0}^\infty \alpha_j^{q(x)} \le 1$, then (3.16) follows by monotonicity and $p_0(x)\le p_1(x)$ for any $C\ge 1$. If $\sum_{j=0}^\infty \alpha_j^{q(x)} = \infty$, then there is nothing to prove. In the remaining case $1 < \sum_{j=0}^\infty \alpha_j^{q(x)} < \infty$ we find a non-negative integer $J\in\mathbf{N}_0$ such that

$$2^{\frac{Jnq(x)}{p_0(x)}} < \sum_{j=0}^\infty \alpha_j^{q(x)} \le 2^{\frac{(J+1)nq(x)}{p_0(x)}}. \tag{3.18}$$

We split the sum over $j \in \mathbf{N}_0$ into two parts, apply (3.17) in the first part and use the inequality $p_0(x) \le p_1(x)$ together with (3.18) in the second part.

$$\begin{aligned}
&\sum_{j=0}^{\infty} 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)}\alpha_j^{q(x)}\\
&= \sum_{j=0}^{J} 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)}\alpha_j^{q(x)} + \sum_{j=J+1}^{\infty} 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)}\alpha_j^{q(x)}\\
&\le c^{q(x)} \sum_{j=0}^{J} 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)} 2^{\frac{jnq(x)}{p_0(x)}} + 2^{(J+1)n\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q(x)} \sum_{j=J+1}^{\infty} \alpha_j^{q(x)}\\
&\le c^{q(x)} \sum_{j=0}^{J} 2^{\frac{jnq(x)}{p_1(x)}} + 2^{\frac{(J+1)nq(x)}{p_1(x)}} \le c_1^{q(x)}\, 2^{\frac{(J+1)nq(x)}{p_1(x)}}\\
&\le c_1^{q(x)}\, 2^{\frac{nq(x)}{p_1(x)}} \Big(\sum_{j=0}^{\infty} \alpha_j^{q(x)}\Big)^{\frac{p_0(x)}{p_1(x)}} \le C^{q(x)} \Big(\sum_{j=0}^{\infty} \alpha_j^{q(x)}\Big)^{\frac{p_0(x)}{p_1(x)}}.
\end{aligned}$$

In the last line, we used $0 < p_1^- \le p_1^+ < \infty$ and again (3.18). This finishes the proof of (3.16) and consequently of the whole Step 2. □

Theorem 3.2. *Let $-\infty < s_1(x) < s_0(x) < \infty$ and $0 < p_0(x) < p_1(x) < \infty$ for all $x \in \mathbf{R}^n$ with $0 < p_0^- < p_1^+ < \infty$,*

$$s_0(x) - \frac{n}{p_0(x)} = s_1(x) - \frac{n}{p_1(x)}, \quad x \in \mathbf{R}^n$$

and

$$\varepsilon := \inf_{x\in\mathbf{R}^n} (s_0(x) - s_1(x)) = n \inf_{x\in\mathbf{R}^n} \left(\frac{1}{p_0(x)} - \frac{1}{p_1(x)}\right) > 0. \tag{3.19}$$

Let $s_0, \frac{1}{p_0} \in C^{\log}_{1-\mathrm{loc}}(\mathbf{R}^n)$. Then, for every $0 < q \le \infty$,

$$f^{s_0(\cdot)}_{p_0(\cdot),\infty} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q}.$$

Proof. We use again the notation of (3.1)–(3.4).

$$\begin{aligned}
\|\gamma|f^{s_1(\cdot)}_{p_1(\cdot),q}\| &= \left\|\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{js_1(x)q}|\gamma_{j\,m}|^q\chi_{j\,m}(x)\Big)^{1/q}|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\
&\le \left\|\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{js_1(x)q}\Big(\inf_{y\in Q_{jm}} 2^{-js_0(y)}h(y)\Big)^q\chi_{jm}(x)\Big)^{1/q}|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\
&\le \left\|\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{j(s_1(x)-s_0(x))q}\Big(\inf_{y\in Q_{jm}} 2^{j(s_0(x)-s_0(y))}h(y)\Big)^q\chi_{jm}(x)\Big)^{1/q}|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|\\
&\le c\left\|\Big(\sum_{j=0}^{\infty}\sum_{m\in\mathbf{Z}^n} 2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q}\Big(\inf_{y\in Q_{jm}} h(y)\Big)^q\chi_{jm}(x)\Big)^{1/q}|L_{p_1(\cdot)}(\mathbf{R}^n)\right\|.
\end{aligned} \tag{3.20}$$

Let again $\lambda > 0$ be a positive real number, such that

$$\int_{\mathbf{R}^n} \Big(\frac{h(x)}{\lambda}\Big)^{p_0(x)} dx \le 1. \tag{3.21}$$

For almost every $x \in \mathbf{R}^n$ and every $j \in \mathbf{N}_0$ there is exactly one $m = m(j)$ such that $x \in Q_{j,m(j)}$. Fix one such $x \in \mathbf{R}^n$ and set

$$\alpha_j := \frac{\inf\limits_{y \in Q_{j,m(j)}} h(y)}{\lambda}.$$

Then $\{\alpha_j\}_{j=0}^\infty$ is a non-decreasing sequence of non-negative real numbers with $\alpha := \lim\limits_{j\to\infty} \alpha_j \le \dfrac{h(x)}{\lambda}$.

Let first $\alpha \le 1$. Then we use the monotonicity of $\{\alpha_j\}$, (3.19) and obtain for $C^q \ge (1-2^{-n\epsilon q})^{-1}$

$$\begin{aligned}\Bigg(\sum_{j=0}^\infty C^{-q} 2^{jn\big(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\big)q} \alpha_j^q\Bigg)^{p_1(x)/q} &\le \Bigg(\sum_{j=0}^\infty C^{-q} 2^{jn\big(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\big)q} \alpha^q\Bigg)^{p_1(x)/q} \\ = \Bigg(\sum_{j=0}^\infty C^{-q} 2^{jn\big(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\big)q}\Bigg)^{p_1(x)/q} &\cdot \alpha^{p_1(x)} \le \alpha^{p_0(x)} \le \Big(\frac{h(x)}{\lambda}\Big)^{p_0(x)}.\end{aligned} \tag{3.22}$$

Let us now consider the case $\alpha > 1$. By (3.21), we get

$$1 \ge \int_{\mathbf{R}^n} \Big(\frac{h(x)}{\lambda}\Big)^{p_0(x)} dx \ge \int_{Q_{j,m(j)}} \alpha_j^{p_0(x)} dx.$$

If $\alpha_j > 1$, we may further estimate

$$1 \ge 2^{-jn} \alpha_j^{\inf_{y\in Q_{j,m(j)}} p_0(y)}.$$

We apply Lemma 2.9 for $\frac{1}{p_0}$ to obtain an analogue of (3.17)

$$\alpha_j \le 2^{\frac{jn}{\inf_{y\in Q_{j,m(j)}} p_0(y)}} = 2^{\frac{jn}{p_0(x)}} \cdot 2^{\frac{jn}{\inf_{y\in Q_{j,m(j)}}} - \frac{jn}{p_0(x)}} \le c_{\log} 2^{\frac{jn}{p_0(x)}} \tag{3.23}$$

and this estimate holds true also for $\alpha_j \le 1$.

We show, that for $C > 0$ large enough (cf. (3.16))

$$\sum_{j=0}^\infty C^{-q} 2^{jn\big(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\big)q} \alpha_j^q \le \alpha^{\frac{qp_0(x)}{p_1(x)}}. \tag{3.24}$$

As $\alpha = \infty$ implies $h(x) = \infty$ and this happens only for a set of $x \in \mathbf{R}^n$ with measure zero, we may choose for almost every $x \in \mathbf{R}^n$ a non-negative integer $J \in \mathbf{N}_0$ such that

$$2^{\frac{Jn}{p_0(x)}} < \alpha \le 2^{\frac{(J+1)n}{p_0(x)}} \tag{3.25}$$

and split

$$\sum_{j=0}^\infty C^{-q} 2^{jn\big(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\big)q} \alpha_j^q = \underbrace{\sum_{j=0}^J \dots}_{I} + \underbrace{\sum_{j=J+1}^\infty \dots}_{II}.$$

By (3.23) and (3.25)

$$I=\sum_{j=0}^{J}C^{-q}2^{\frac{jnq}{p_1(x)}}\cdot 2^{-\frac{jnq}{p_0(x)}}\cdot\alpha_j^q\le\sum_{j=0}^{J}C^{-q}c_{\log}2^{\frac{jnq}{p_1(x)}}\le c^{-1}2^{\frac{(J+1)nq}{p_1(x)}}\le 2^{\frac{Jnq}{p_1(x)}}\le\alpha^{\frac{qp_0(x)}{p_1(x)}}.$$

The monotonicity of $\{\alpha_j\}$ and (3.25) lead to

$$\begin{aligned}II&\le\sum_{j=J+1}^{\infty}2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q}\alpha_j^qC^{-q}\le\alpha^qC^{-q}\sum_{j=J+1}^{\infty}2^{jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q}\\&\le\alpha^qC^{-q}2^{Jn\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q}\le\alpha^qC^{-q}\left(\alpha^{p_0(x)}2^{-n}\right)^{\left(\frac{1}{p_1(x)}-\frac{1}{p_0(x)}\right)q}\\&=\alpha^{\frac{qp_0(x)}{p_1(x)}}C^{-q}2^{n\left(\frac{1}{p_0(x)}-\frac{1}{p_1(x)}\right)q}\le\alpha^{\frac{qp_0(x)}{p_1(x)}}\end{aligned}$$

This finishes the proof of (3.24). Now (3.20), (3.22), (3.24) with (3.21) gives

$$||\gamma|f^{s_1(\cdot)}_{p_1(\cdot),q}||\le C||\gamma|f^{s_0(\cdot)}_{p_0(\cdot),\infty}||.$$ □

Remark 3.3. The original proof of Jawerth of Theorem 1.1 used the technique of a distribution function, which fails for $L_{p(\cdot)}(\mathbf{R}^n)$. Another proof was given by Johnsen and Sickel [22] and relied on an inequality of Plancherel–Pólya–Nikol'skij type. Its classical proof [34, Chapter 1.3] is based on dilation arguments and (at least to our knowledge) there is still no analogue of these inequalities for $L_{p(\cdot)}(\mathbf{R}^n)$ up to now.

Our proofs of Theorems 3.1 and 3.2 were motived by [38]. An essential technique used there was the concept of non-increasing rearrangement. Unfortunately, it fails completely in the case of variable integrability exponents $p_0(x)$ and $p_1(x)$. To avoid this obstacle, we had to employ the somehow artificial inequality (3.24)—or its analogue (3.16). To motivate this step, let us consider the interpolation inequality between Lorentz spaces

$$||f|L_{p_1,q}(0,1)||\le c\,||f|L_{p_0,\infty}(0,1)||^\theta\cdot||f|L_\infty(0,1)||^{1-\theta}\tag{3.26}$$

with

$$0<p_0<p_1<\infty,\quad\frac{1}{p_1}=\frac{\theta}{p_0}+\frac{1-\theta}{\infty},\quad 0<\theta<1$$

and its discrete version

$$\left(\sum_{j=0}^{\infty}2^{-jnq(\frac{1}{p_0}-\frac{1}{p_1})}f^*(2^{-jn})^q\right)^{1/q}\le c\left(\sup_{j\in\mathbf{N}_0}2^{-jn/p_0}f^*(2^{-jn})\right)^{1-\frac{p_0}{p_1}}\cdot\left(\sup_{j\in\mathbf{N}_0}f^*(2^{-jn})\right)^{\frac{p_0}{p_1}}.$$

We refer to [2, Chapter 2] as a standard reference for non-increasing rearrangements and to [2, Chapter 4.4] for the notation connected with Lorentz spaces. We leave the details to the reader. The reader may also observe some similarities between (3.26) and the inequality (4) of [22].

Using Theorem 2.12, we obtain immediately following

Theorem 3.4. *Let s_0,s_1,p_0,p_1 and q be continuous functions satisfying the standing assumptions of* [9]. *Let $s_0(x)\ge s_1(x)$ and $p_0(x)\le p_1(x)$ for all $x\in\mathbf{R}^n$ with*

$$s_0(x)-\frac{n}{p_0(x)}=s_1(x)-\frac{n}{p_1(x)},\quad x\in\mathbf{R}^n.$$

Then

$$F^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}(\mathbf{R}^n).$$

We denote by $W^k_{p(\cdot)}(\mathbf{R}^n)$ the Sobolev space of functions form $L_{p(\cdot)}(\mathbf{R}^n)$, such that all its distributional derivatives of order smaller or equal to k exist and belong to $L_{p(\cdot)}(\mathbf{R}^n)$. Furthermore, we introduce the Bessel potential spaces of variable integrability introduced by Almeida and Samko [1] and by Gurka, Harjulehto and Nekvinda [16]. Let $\sigma \in \mathbf{R}$ and let $B^\sigma = F^{-1}(1+|\xi|^2)^{-\sigma/2}F$ be the Bessel potential operator. We set

$$L^\sigma_{p(\cdot)}(\mathbf{R}^n) = \{B^\sigma f : f \in L_{p(\cdot)}(\mathbf{R}^n)\}$$

and equip this space with norm $||f|L^\sigma_{p(\cdot)}(\mathbf{R}^n)|| = ||B^{-\sigma}f|L_{p(\cdot)}(\mathbf{R}^n)||$.

Let $p \in C^{\log}(\mathbf{R}^n)$ with $1 < p^- \le p^+ < \infty$ and $\sigma \in [0,\infty)$. It was shown in [9, Theorem 4.5] that $F^\sigma_{p(\cdot),2}(\mathbf{R}^n) \cong L^\sigma_{p(\cdot)}(\mathbf{R}^n)$ in the sense of equivalent norms. If moreover $\sigma \in \mathbf{N}_0$, then $F^\sigma_{p(\cdot),2}(\mathbf{R}^n) \cong W^\sigma_{p(\cdot)}(\mathbf{R}^n)$.

Hence setting $q = 2$ implies embeddings of Bessel potential spaces.

Theorem 3.5. *Let $0 \le s_1 \le s_0 < \infty$ and $p_0, p_1 \in C^{\log}(\mathbf{R}^n)$ with $1 < p_0^- \le p_0(x) \le p_1(x) \le p_1^+ < \infty$ for all $x \in \mathbf{R}^n$. If*

$$s_0 - \frac{n}{p_0(x)} = s_1 - \frac{n}{p_1(x)}, \quad x \in \mathbf{R}^n,$$

then

$$L^{s_0}_{p_0(\cdot)}(\mathbf{R}^n) \hookrightarrow L^{s_1}_{p_1(\cdot)}(\mathbf{R}^n).$$

If $s_1 \in \mathbf{N}_0$, then $L^{s_1}_{p_1(\cdot)}(\mathbf{R}^n)$ may be replaced by $W^{s_1}_{p_1(\cdot)}(\mathbf{R}^n)$ and similarly for s_0.

Remark 3.6. Let us only mention, that if $1 < p^- \le p^+ < \infty$, then $p \in C^{\log}(\mathbf{R}^n)$ if, and only if, $\frac{1}{p} \in C^{\log}(\mathbf{R}^n)$. So the *standing assumptions* on p_0 and p_1 are satisfied and the proof becomes trivial.

Theorem 3.7. *Let $s_0, s_1, p_0, p_1, q_0, q_1$ be continuous functions satisfying the standing assumptions of* [9] *with*

$$s_0(x) - \frac{n}{p_0(x)} = s_1(x) - \frac{n}{p_1(x)}, \quad x \in \mathbf{R}^n$$

and

$$\inf_{x\in\mathbf{R}^n}(s_0(x) - s_1(x)) = n \inf_{x\in\mathbf{R}^n}\Big(\frac{1}{p_0(x)} - \frac{1}{p_1(x)}\Big) > 0.$$

Then

$$F^{s_0(\cdot)}_{p_0(\cdot),q_0(\cdot)}(\mathbf{R}^n) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q_1(\cdot)}(\mathbf{R}^n).$$

Proof. By monotonicity and using Theorem 3.2, we obtain

$$f^{s_0(\cdot)}_{p_0(\cdot),q_0(\cdot)} \hookrightarrow f^{s_0(\cdot)}_{p_0(\cdot),\infty} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q_1^-} \hookrightarrow f^{s_1(\cdot)}_{p_1(\cdot),q_1(\cdot)}$$

and Theorem 2.12 finishes the proof. □

Finally, we may combine our embedding results with the trace results of [9] and obtain the following Sobolev embeddings for traces. We state it for Sobolev spaces, but a similar assertion holds also for Bessel potential spaces and Triebel–Lizorkin spaces.

Theorem 3.8. *Let $k \in \mathbf{N}$ and $1 < p^- \le p^+ < \frac{n}{k}$ with $\frac{1}{p} \in C^{\log}(\mathbf{R}^n)$. Then*

$$W^k_{p(\cdot)}(\mathbf{R}^n) \hookrightarrow L_{\frac{(n-1)p(\cdot)}{n-kp(\cdot)}}(\mathbf{R}^{n-1}).$$

Proof. By Theorem 3.13. of [9], we have

$$\operatorname{tr} W^k_{p(\cdot)}(\mathbf{R}^n) \to F^{k-\frac{1}{p(\cdot)}}_{p(\cdot),p(\cdot)}(\mathbf{R}^{n-1}),$$

which may be combined with Theorem 3.7

$$F^{k-\frac{1}{p(\cdot)}}_{p(\cdot),p(\cdot)}(\mathbf{R}^{n-1}) \hookrightarrow F^0_{\tilde{p}(\cdot),2}(\mathbf{R}^{n-1}) = L_{\tilde{p}(\cdot)}(\mathbf{R}^{n-1})$$

for $\tilde{p}(\cdot)$ given by

$$k - \frac{1}{p(\cdot)} - \frac{n-1}{p(\cdot)} = -\frac{n-1}{\tilde{p}(\cdot)}.$$

This finishes the proof. □

Acknowledgement. I would like to thank Lars Diening, Peter Hästö and the anonymous referee for their comments, which helped to improve the paper.

References

[1] Almeida, A., and S. Samko: Characterization of Riesz and Bessel potentials on variable Lebesgue spaces. - J. Funct. Spaces Appl. 4:2, 2006, 113–144.

[2] Bennett, C., and R. Sharpley: Interpolation of operators. - Academic Press, San Diego, 1988.

[3] Besov, O. V.: Embeddings of spaces of differentiable functions of variable smoothness. - Tr. Mat. Inst. Steklova 214:17, 1997, 25–58; Engl. transl.: Proc. Steklov Inst. Math. 214:3, 1996, 19–53.

[4] Besov, O. V.: Interpolation, embedding, and extension of spaces of functions of variable smoothness. - Tr. Mat. Inst. Steklova 248, 2005, 52–63; Engl. transl.: Proc. Steklov Inst. Math. 248:1, 2005, 47–58.

[5] Bony, J.-M.: Second microlocalization and propagation of singularities for semilinear hyperbolic equations. - In: Hyperbolic equations and related topics (Katata/Kyoto, 1984), Academic Press, Boston, MA, 1986, 11–49.

[6] Diening, L.: Maximal function on generalized Lebesgue spaces $L^{p(\cdot)}$. - Math. Inequal. Appl. 7:2, 2004, 245–253.

[7] Diening, L.: Riesz potential and Sobolev embeddings on generalized Lebesgue and Sobolev spaces $L^{p(\cdot)}$ and $W^{k,p(\cdot)}$. - Math. Nachr. 268, 2004, 31–43.

[8] Diening, L., P. Hästö, and A. Nekvinda: Open problems in variable exponent Lebesgue and Sobolev spaces. - In: FSDONA04 Proceedings, edited by Drábek and Rákosník, Milovy, Czech Republic, 2004, 38–58.

[9] Diening, L., P. Hästö, and S. Roudenko: Function spaces of variable smoothness and integrability. - J. Funct. Anal. 256:6, 2009, 1731–1768.

[10] Edmunds, D. E., and J. Rákosník: Sobolev embeddings with variable exponent. - Studia Math. 143:3, 2000, 267–293.

[11] Edmunds, D. E., and J. Rákosník: Sobolev embeddings with variable exponent II. - Math. Nachr. 246/247, 2002, 53–67.

[12] Fan, X.-L.: Boundary trace embedding theorems for variable exponent Sobolev spaces. - J. Math. Anal. Appl. 339:2, 2008, 1395–1412.

[13] Fan, X.-L., J. Shen, and D. Zhao: Sobolev embedding theorems for spaces $W^{k,p(x)}$. -J. Math. Anal. Appl. 262, 2001, 749–760.

[14] Frazier, M., and B. Jawerth: Decomposition of Besov spaces. - Indiana Univ. Math. J. 34, 1985, 777–799.

[15] Frazier, M., and B. Jawerth: A discrete transform and decompositions of distribution spaces. - J. Funct. Anal. 93, 1990, 34–170.

[16] Gurka, P., P. Harjulehto, and A. Nekvinda: Bessel potential spaces with variable exponent. - Math. Inequal. Appl. 10:3, 2007, 661–676.

[17] Hardy, G. H., and J. E. Littlewood: Some properties of fractional integrals I. - Math. Z. 27, 1928, 565–606.

[18] Hardy, G. H., and J. E. Littlewood: Some properties of fractional integrals II. - Math. Z. 34, 1932, 403–439.

[19] Hästö, P.: Local-to-global results in variable exponent spaces. - Math. Res. Letters (to appear).

[20] Jaffard, S.: Pointwise smoothness, two-microlocalization and wavelet coefficients. - In: Conference on Mathematical Analysis (El Escorial, 1989), Publ. Mat. 35:1, 1991, 155–168.

[21] Jawerth, B.: Some observations on Besov and Lizorkin–Triebel spaces. - Math. Scand. 40, 1977, 94–104.

[22] Johnsen, J., and W. Sickel: A direct proof of Sobolev embeddings for quasi-homogeneous Lizorkin–Triebel spaces with mixed norms. - J. Funct. Spaces Appl. 5, 2007, 183–198.

[23] Kempka, H.: 2-Microlocal Besov and Triebel–Lizorkin spaces of variable integrability. - Rev. Mat. Complut. (to appear).

[24] Kováčik, O., and J. Rákosník: On spaces $L^{p(x)}$ and $W^{1,p(x)}$. - Czechoslovak Math. J. 41:116, 1991, 592–618.

[25] Leopold, H.-G.: On Besov spaces of variable order of differentiation. - Z. Anal. Anwendungen 8:1, 1989, 69–82.

[26] Leopold, H.-G.: Embedding of function spaces of variable order of differentiation in function spaces of variable order of integration. - Czechoslovak Math. J. 49:124, 1999, 633–644.

[27] Orlicz, W.: Über konjugierte Exponentenfolgen. - Studia Math. 3, 1931, 200–212.

[28] Peetre, J.: New thoughts on Besov spaces. - Duke Univ. Math. Series, Durham, 1976.

[29] Ružička, M.: Electrorheological fluids: mathematical modelling and existence theory. - Habilitationsschrift, Universität Bonn, 1998.

[30] Ružička, M.: Flow of shear dependent electrorheological fluids. - C. R. Acad. Sci. Paris S. I Math. 329:5, 1999, 393–398.

[31] Ružička, M.: Electrorheological fluids: modeling and mathematical theory. - Lecture Notes in Math. 1748, Springer-Verlag, Berlin, 2000.

[32] Samko, S.: Convolution and potential type operators in $L^{p(x)}(\mathbf{R}^n)$. - Integral Transforms Spec. Funct. 7:3-4, 1998, 261–284.

[33] Sobolev, S. L.: On a theorem of functional analysis. - Mat. Sbornik 4, 1938; Engl. transl.: Amer. Math. Soc. Transl. (2) 34, 1963, 39–68.

[34] Triebel, H.: Theory of function spaces. - Birkhäuser, Basel, 1983.

[35] Triebel, H.: Theory of function spaces II. - Birkhäuser, Basel, 1992.

[36] Triebel, H.: The structure of functions. - Birkhäuser, Basel, 2001.

[37] Triebel, H.: Theory of function spaces III. - Birkhäuser, Basel, 2006.

[38] Vybíral, J.: A new proof of the Jawerth–Franke embedding. - Rev. Mat. Complut. 21, 2008, 75–82.

[39] Xu, J.: Variable Besov and Triebel–Lizorkin spaces. - Ann. Acad. Sci. Fenn. Math. 33:2, 2008, 511–522.

Received 13 October 2008

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Functional Analysis 264 (2013) 1197–1237

JOURNAL OF
Functional
Analysis

www.elsevier.com/locate/jfa

Non-smooth atomic decompositions, traces on Lipschitz domains, and pointwise multipliers in function spaces

Cornelia Schneider [a], Jan Vybíral [b,*]

[a] Applied Mathematics III, University Erlangen–Nuremberg, Cauerstraße 11, D-91058 Erlangen, Germany
[b] Department of Mathematics, Technical University Berlin, Street of 17. June 136, D-10623 Berlin, Germany

Received 11 January 2012; accepted 18 December 2012

Available online 7 January 2013

Communicated by F. Otto

Abstract

We provide non-smooth atomic decompositions for Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$, $s > 0$, $0 < p, q \leqslant \infty$, defined via differences. The results are used to compute the trace of Besov spaces on the boundary Γ of bounded Lipschitz domains Ω with smoothness s restricted to $0 < s < 1$ and no further restrictions on the parameters p, q. We conclude with some more applications in terms of pointwise multipliers.

Keywords: Lipschitz domains; Besov spaces; Differences; Real interpolation; Atoms; Traces; Pointwise multipliers

0. Introduction

Besov spaces – sometimes briefly denoted as B-spaces in the sequel – of positive smoothness, have been investigated for many decades already, resulting, for instance, from the study of partial differential equations, interpolation theory, approximation theory, harmonic analysis.

There are several definitions of Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ to be found in the literature. Two of the most prominent approaches are the *Fourier-analytic approach* using Fourier transforms on the one hand and the *classical approach* via higher order differences involving the modulus of smoothness on the other. These two definitions are equivalent only with certain restrictions on the parameters, in particular, they differ for $0 < p < 1$ and $0 < s \leqslant n(\frac{1}{p} - 1)$, but may otherwise share similar properties.

[*] Corresponding author.
E-mail addresses: schneider@am.uni-erlangen.de (C. Schneider), vybiral@math.tu-berlin.de (J. Vybíral).

http://dx.doi.org/10.1016/j.jfa.2012.12.005

In the present paper we focus on the *classical approach*, which introduces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ as those subspaces of $L_p(\mathbb{R}^n)$ such that

$$\left\| f \middle| \mathbf{B}^s_{p,q}(\mathbb{R}^n) \right\|_r = \left\| f \middle| L_p(\mathbb{R}^n) \right\| + \left(\int_0^1 t^{-sq} \omega_r(f,t)_p^q \frac{\mathrm{d}t}{t} \right)^{1/q}$$

is finite, where $0 < p, q \leqslant \infty$, $s > 0$, $r \in \mathbb{N}$ with $r > s$, and $\omega_r(f,t)_p$ is the usual r-th modulus of smoothness of $f \in L_p(\mathbb{R}^n)$.

These spaces occur naturally in non-linear approximation theory. Especially important is the case $p < 1$, which is needed for the description of approximation classes of classical methods such as rational approximation and approximation by splines with free knots. For more details we refer to the introduction of [7].

For our purposes it will be convenient to use an equivalent characterization for the classical Besov spaces, cf. [16], [43, Sect. 9.2], and also [33, Th. 2.11], relying on *smooth atomic decompositions*. They allow us to characterize $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ as the space of those $f \in L_p(\mathbb{R}^n)$ which can be represented as

$$f(x) = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} a_{j,m}(x), \quad x \in \mathbb{R}^n, \tag{0.1}$$

with the sequence of coefficients $\lambda = \{\lambda_{j,m} \in \mathbb{C}\colon j \in \mathbb{N}_0,\ m \in \mathbb{Z}^n\}$ belonging to some appropriate sequence space $b^s_{p,q}$, where $s > 0$, $0 < p, q \leqslant \infty$, and with smooth atoms $a_{j,m}(x)$.

It is one of the aims of the present paper to develop non-smooth atomic decompositions for Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$, cf. Theorem 2.6 and Corollary 2.8. We will show that one can relax the assumptions on the smoothness of the atoms $a_{j,m}$ used in the representation (0.1) and, thus, replace these atoms with more general ones without loosing any crucial information compared smooth atomic decompositions for functions $f \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$.

There are only few forerunners dealing with non-smooth atomic decompositions in function spaces so far. We refer to the papers [42,25,4], all mainly considering the different Fourier-analytic approach for Besov spaces and having in common that they restrict themselves to the technically simpler case when $p = q$. Our approach generalizes and extends these results and seems to be the first one covering the full range of indices $0 < p, q \leqslant \infty$. The reader may also consult [30] for another generalization of the classical atomic decomposition technique using building blocks of limited smoothness.

The additional freedom we gain in the choice of suitable non-smooth atoms $a_{j,m}$ for the atomic decompositions of $f \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ makes this approach well suited to further investigate Besov spaces $\mathbf{B}^s_{p,q}(\Omega)$ on non-smooth domains Ω and their boundaries Γ. In particular, we shall focus on bounded Lipschitz domains and start by obtaining some interesting new properties concerning interpolation and equivalent quasi-norms for these spaces as well as an atomic decomposition for Besov spaces $\mathbf{B}^s_{p,q}(\Gamma)$, defined on the boundary $\Gamma = \partial\Omega$ of a Lipschitz domain.

But the main goal of this article is to demonstrate the strength of the newly developed non-smooth atomic decompositions in view of trace results. The trace is taken with respect to the boundary Γ of bounded Lipschitz domains Ω. Our main result reads as

$$\operatorname{Tr} \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega) = \mathbf{B}^s_{p,q}(\Gamma),$$

where $n \geqslant 2$, $0 < s < 1$, and $0 < p, q \leqslant \infty$, cf. Theorem 4.11. Its proof reveals how well suited non-smooth atoms are in order to tackle this problem. The limiting case $s = 0$ is also considered in Corollary 4.13.

In the range $0 < s < 1$, our results are optimal in the sense that there are no further restrictions on the parameters p, q. The fact that we now also cover traces in Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with $p < 1$ could be of particular interest in non-linear approximation theory.

Moreover, as a by-product we obtain corresponding trace results on Lipschitz domains for Triebel–Lizorkin spaces, defined via atomic decompositions.

The papers [32] and [33], dealing with traces on hyperplanes and smooth domains, respectively, might be considered as forerunners of the trace results established in this paper. Nevertheless, the methods we use now are completely different.

The same question for $s \geqslant 1$ was studied in [19]. It turns out that in this case the function spaces on the boundary look very different and also the extension operator must be changed. Moreover, based on the seminal work [18], traces on Lipschitz domains were studied in [21, Th. 1.1.3] for the Fourier-analytic Besov spaces with the natural restrictions

$$(n-1)\max\left(\frac{1}{p}-1,0\right) < s < 1 \quad \text{and} \quad \frac{n-1}{n} < p. \tag{0.2}$$

Our Theorem 4.11 actually covers and extends [21, Th. 1.1.3], as for the parameters restricted by (0.2) the Besov spaces defined by differences coincide with the Fourier-analytic Besov spaces.

In contrast to Mayboroda we make use of the classical Whitney extension operator and the cone property of Lipschitz domains in order to establish our results instead of potential layers and interpolation. Moreover, the extension operator we construct is not linear – and in fact cannot be whenever $0 < s < (n-1)\max(\frac{1}{p}-1,0)$ – compared to the extension operator in [21, Th. 1.1.3]. Let us recall that the importance of non-linear extension operators is known in the theory of differentiable spaces since the pioneering work of Gagliardo [13], cf. also [2, Chapter 5].

Finally, we shall use the non-smooth atomic decompositions again to deal with pointwise multipliers in the respective function spaces. Let $\mathbf{B}^s_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$ denote the self-similar spaces introduced in Definition 5.1 and $M(\mathbf{B}^s_{p,q}(\mathbb{R}^n))$ the set of all pointwise multipliers of $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$. We prove for $s > 0$, $0 < p, q \leqslant \infty$ in Theorem 5.4 the relationship

$$\bigcup_{\sigma > s} \mathbf{B}^\sigma_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right) \subset M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right) \hookrightarrow \mathbf{B}^s_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right). \tag{0.3}$$

Additionally, if $0 < p \leqslant 1$, one even has a coincidence in terms of $M(\mathbf{B}^s_{p,p}(\mathbb{R}^n)) = \mathbf{B}^s_{p,p,\mathrm{selfs}}(\mathbb{R}^n)$. Our results generalize the multiplier assertions from [42] to the case when $p \neq q$. Moreover, they extend previous results to classical Besov spaces with small parameters s and p. In this context we refer to [22–24], where pointwise multipliers in Besov spaces with $p, q \geqslant 1$ and $p = q$ were studied in detail.

We conclude using (0.3) in order to discuss under which circumstances the characteristic function χ_Ω of a bounded domain Ω in $\mathbb{R}^n$ is a pointwise multiplier in $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ – establishing a connection between pointwise multipliers and certain fundamental notion of fractal geometry, so-called h-sets, cf. Definition 5.6. In particular, if a boundary $\Gamma = \partial\Omega$ is an h-set satisfying

$$\sup_{j \in \mathbb{N}_0} \sum_{k=0}^{\infty} 2^{k\sigma q} \left(\frac{h(2^{-j})}{h(2^{-j-k})} 2^{-kn}\right)^{q/p} < \infty,$$

where $\sigma > 0$, $0 < p < \infty$, and $0 < q \leqslant \infty$, then Theorem 5.8 shows that

$$\chi_\Omega \in \mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\big(\mathbb{R}^n\big).$$

The present paper is organized as follows: Section 1 contains notation, definitions, and preliminary assertions on smooth atomic decompositions. The main investigation starts in Section 2, where we construct non-smooth atomic decompositions for the spaces under focus. Afterwards Section 3 provides new insights (and helpful results) concerning function spaces on Lipschitz domains and their boundaries. These powerful techniques are then used in Section 4 in order to compute traces on Lipschitz domains – the heart of this article. Finally, we conclude with some further applications of non-smooth atomic decompositions in terms of pointwise multipliers in Section 5.

1. Preliminaries

We use standard notation. Let $\mathbb{N}$ be the collection of all natural numbers and let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let $\mathbb{R}^n$ be euclidean n-space, $n \in \mathbb{N}$, $\mathbb{C}$ the complex plane. The set of multi-indices $\beta = (\beta_1, \ldots, \beta_n)$, $\beta_i \in \mathbb{N}_0$, $i = 1, \ldots, n$, is denoted by $\mathbb{N}_0^n$, with $|\beta| = \beta_1 + \cdots + \beta_n$, as usual. Moreover, if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}_0^n$ we put $x^\beta = x_1^{\beta_1} \cdots x_n^{\beta_n}$.

We use the symbol '$\lesssim$' in

$$a_k \lesssim b_k \quad \text{or} \quad \varphi(x) \lesssim \psi(x)$$

always to mean that there is a positive number c_1 such that

$$a_k \leqslant c_1 b_k \quad \text{or} \quad \varphi(x) \leqslant c_1 \psi(x)$$

for all admitted values of the discrete variable k or the continuous variable x, where $\{a_k\}_k$, $\{b_k\}_k$ are non-negative sequences and φ, ψ are non-negative functions. We use the equivalence '$\sim$' in

$$a_k \sim b_k \quad \text{or} \quad \varphi(x) \sim \psi(x)$$

for

$$a_k \lesssim b_k \quad \text{and} \quad b_k \lesssim a_k \quad \text{or} \quad \varphi(x) \lesssim \psi(x) \quad \text{and} \quad \psi(x) \lesssim \varphi(x).$$

If $a \in \mathbb{R}$, then $a_+ := \max(a, 0)$ and $[a]$ denotes the integer part of a.

Given two (quasi-) Banach spaces X and Y, we write $X \hookrightarrow Y$ if $X \subset Y$ and the natural embedding of X into Y is continuous. All unimportant positive constants will be denoted by c, occasionally with subscripts. For convenience, let both $\mathrm{d}x$ and $|\cdot|$ stand for the (n-dimensional) Lebesgue measure in the sequel. $L_p(\mathbb{R}^n)$, with $0 < p \leqslant \infty$, stands for the usual quasi-Banach space with respect to the Lebesgue measure, quasi-normed by

$$\big\| f \big| L_p\big(\mathbb{R}^n\big) \big\| := \bigg(\int_{\mathbb{R}^n} \big| f(x) \big|^p \,\mathrm{d}x \bigg)^{\frac{1}{p}}$$

with the appropriate modification if $p = \infty$. Throughout the paper Ω will denote a domain in $\mathbb{R}^n$ and the Lebesgue space $L_p(\Omega)$ is defined in the usual way.

We denote by $C^K(\mathbb{R}^n)$ the space of all K-times continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ equipped with the norm

$$\big\| f \big| C^K\big(\mathbb{R}^n\big)\big\| = \max_{|\alpha| \leqslant K} \sup_{x \in \mathbb{R}^n} \big| D^\alpha f(x) \big|.$$

Additionally, $C^\infty(\mathbb{R}^n)$ contains the set of smooth and bounded functions on $\mathbb{R}^n$, i.e.,

$$C^\infty\big(\mathbb{R}^n\big) := \bigcap_{K \in \mathbb{N}} C^K\big(\mathbb{R}^n\big),$$

whereas $C_0^\infty(\mathbb{R}^n)$ denotes the space of smooth functions with compact support.

Furthermore, $B(x_0, R)$ stands for an open ball with radius $R > 0$ around $x_0 \in \mathbb{R}^n$,

$$B(x_0, R) = \big\{ x \in \mathbb{R}^n \colon |x - x_0| < R \big\}. \tag{1.1}$$

Let $Q_{j,m}$ with $j \in \mathbb{N}_0$ and $m \in \mathbb{Z}^n$ denote a cube in $\mathbb{R}^n$ with sides parallel to the axes of coordinates, centered at $2^{-j}m$, and with side length 2^{-j+1}. For a cube Q in $\mathbb{R}^n$ and $r > 0$, we denote by rQ the cube in $\mathbb{R}^n$ concentric with Q and with side length r times the side length of Q. Furthermore, $\chi_{j,m}$ stands for the characteristic function of $Q_{j,m}$.

Let $G \subset \mathbb{R}^n$ and $j \in \mathbb{N}_0$. We use the abbreviation

$$\sum_{m \in \mathbb{Z}^n}^{G,j} = \sum_{m \in \mathbb{Z}^n,\, Q_{j,m} \cap G \neq \emptyset}, \tag{1.2}$$

where G will usually denote either a domain Ω in $\mathbb{R}^n$ or its boundary Γ.

1.1. Smooth atomic decompositions in function spaces

We introduce the Besov spaces $\mathbf{B}^s_{p,q}(\Omega)$ through their decomposition properties. This provides a constructive definition expanding functions f via smooth atoms (excluding any moment conditions) and suitable coefficients, where the latter belong to certain sequence spaces denoted by $b^s_{p,q}(\Omega)$ defined below.

Definition 1.1. Let $0 < p, q \leqslant \infty$, $s \in \mathbb{R}$. Furthermore, let $\Omega \subset \mathbb{R}^n$ and $\lambda = \{\lambda_{j,m} \in \mathbb{C} \colon j \in \mathbb{N}_0, m \in \mathbb{Z}^n\}$. Then

$$b^s_{p,q}(\Omega) = \left\{ \lambda \colon \big\| \lambda \big| b^s_{p,q}(\Omega) \big\| = \left(\sum_{j=0}^{\infty} 2^{j(s - \frac{n}{p})q} \left(\sum_{m \in \mathbb{Z}^n}^{\Omega, j} |\lambda_{j,m}|^p \right)^{q/p} \right)^{1/q} < \infty \right\}$$

(with the usual modification if $p = \infty$ and/or $q = \infty$).

Remark 1.2. If $\Omega = \mathbb{R}^n$, we simply write $b^s_{p,q}$ and $\sum_m$ instead of $b^s_{p,q}(\Omega)$ and $\sum_m^{\Omega,j}$, respectively.

Now we define the smooth atoms.

Definition 1.3. Let $K \in \mathbb{N}_0$ and $d > 1$. A K-times continuously differentiable complex-valued function a on $\mathbb{R}^n$ (continuous if $K = 0$) is called a K-atom if for some $j \in \mathbb{N}_0$

$$\operatorname{supp} a \subset dQ_{j,m} \quad \text{for some } m \in \mathbb{Z}^n, \tag{1.3}$$

and

$$\left|D^\alpha a(x)\right| \leqslant 2^{|\alpha| j} \quad \text{for } |\alpha| \leqslant K. \tag{1.4}$$

It is convenient to write $a_{j,m}(x)$ instead of $a(x)$ if this atom is located at $Q_{j,m}$ according to (1.3). Furthermore, K denotes the smoothness of the atom, cf. (1.4).

We define Besov spaces $\mathbf{B}^s_{p,q}(\Omega)$ using the *atomic approach*.

Definition 1.4. Let $s > 0$ and $0 < p, q \leqslant \infty$. Let $d > 1$ and $K \in \mathbb{N}_0$ with

$$K \geqslant \bigl(1 + [s]\bigr)$$

be fixed. Then $f \in L_p(\Omega)$ belongs to $\mathbf{B}^s_{p,q}(\Omega)$ if, and only if, it can be represented as

$$f(x) = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n}^{\Omega, j} \lambda_{j,m} a_{j,m}(x), \tag{1.5}$$

where the $a_{j,m}$ are K-atoms ($j \in \mathbb{N}_0$) with

$$\operatorname{supp} a_{j,m} \subset dQ_{j,m}, \quad j \in \mathbb{N}_0,\ m \in \mathbb{Z}^n,$$

and $\lambda \in b^s_{p,q}(\Omega)$, convergence being in $L_p(\Omega)$. Furthermore,

$$\bigl\| f \big| \mathbf{B}^s_{p,q}(\Omega) \bigr\| := \inf \bigl\| \lambda \big| b^s_{p,q}(\Omega) \bigr\|, \tag{1.6}$$

where the infimum is taken over all admissible representations (1.5).

Remark 1.5. According to [43], based on [16], the above defined spaces are independent of d and K. This may justify our omission of K and d in (1.6).

Since the atoms $a_{j,m}$ used in Definition 1.4 are defined also outside of Ω, the spaces $\mathbf{B}^s_{p,q}(\Omega)$ can as well be regarded as restrictions of the corresponding spaces on $\mathbb{R}^n$ in the usual interpretation, i.e.,

$$\mathbf{B}^s_{p,q}(\Omega) = \bigl\{ f \in L_p(\Omega):\ \text{there exists } g \in \mathbf{B}^s_{p,q}\bigl(\mathbb{R}^n\bigr) \text{ with } g|_\Omega = f \bigr\},$$

furnished with the norm

$$\bigl\| f \big| \mathbf{B}^s_{p,q}(\Omega) \bigr\| = \inf \bigl\{ \bigl\| g \big| \mathbf{B}^s_{p,q}\bigl(\mathbb{R}^n\bigr) \bigr\| \text{ with } g|_\Omega = f \bigr\},$$

where $g|_\Omega = f$ denotes the restriction of g to Ω. Therefore, well-known embedding results for B-spaces defined on $\mathbb{R}^n$ carry over to those defined on domains Ω. Let $s > 0$, $\varepsilon > 0$, $0 < q, u \leqslant \infty$, and $q \leqslant v \leqslant \infty$. Then we have

$$\mathbf{B}^{s+\varepsilon}_{p,u}(\Omega) \hookrightarrow \mathbf{B}^s_{p,q}(\Omega) \quad \text{and} \quad \mathbf{B}^s_{p,q}(\Omega) \hookrightarrow \mathbf{B}^s_{p,v}(\Omega),$$

cf. [17, Th. 1.15], where also further embeddings for Besov spaces may be found.

Classical approach Originally Besov spaces were defined merely using higher order differences instead of atomic decompositions. The question arises whether this *classical approach* coincides with our *atomic approach*. This might not always be the case but is true for spaces defined on $\mathbb{R}^n$ and on so-called (ε, δ)-domains which we introduce next.

Recall that domain always stands for open set. The boundary of Ω is denoted by $\Gamma = \partial\Omega$.

Definition 1.6. Let Ω be a domain in $\mathbb{R}^n$ with $\Omega \neq \mathbb{R}^n$. Then Ω is said to be an (ε, δ)-domain, where $0 < \varepsilon < \infty$ and $0 < \delta < \infty$, if it is connected and if for any $x \in \Omega$, $y \in \Omega$ with $|x - y| < \delta$ there is a curve $L \subset \Omega$, connecting x and y such that $|L| \leqslant \varepsilon^{-1}|x - y|$ and

$$\operatorname{dist}(z, \Gamma) \geqslant \varepsilon \min\bigl(|x - z|, |y - z|\bigr), \quad z \in L. \tag{1.7}$$

Remark 1.7. All domains we will be concerned with in the sequel are (ε, δ)-domains. In particular, the definition includes *minimally smooth* domains in the sense of Stein, cf. [37, p. 189], and therefore bounded Lipschitz domains (as will be considered in Section 3).

Furthermore, the half-space $\mathbb{R}^n_+ := \{x\colon\ x = (x', x_n) \in \mathbb{R}^n,\ x' \in \mathbb{R}^{n-1},\ x_n > 0\}$ is another example.

It is well-known that (ε, δ)-domains play a crucial role concerning questions of extendability. It is precisely this property which was used in [33, Th. 2.10] to show that for (ε, δ)-domains the atomic approach for B-spaces is equivalent to the *classical approach* (in terms of equivalent quasi-norms), which introduces $\mathbf{B}^s_{p,q}(\Omega)$ as the subspace of $L_p(\Omega)$ such that

$$\bigl\| f \big| \mathbf{B}^s_{p,q}(\Omega) \bigr\|_r = \bigl\| f | L_p(\Omega) \bigr\| + \left(\int_0^1 t^{-sq} \omega_r(f, t, \Omega)_p^q \, \frac{dt}{t} \right)^{1/q} \tag{1.8}$$

is finite, where $0 < p, q \leqslant \infty$ (with the usual modification if $q = \infty$), $s > 0$, $r \in \mathbb{N}$ with $r > s$. Here $\omega_r(f, t, \Omega)_p$ stands for the usual r-th modulus of smoothness of a function $f \in L_p(\Omega)$,

$$\omega_r(f, t, \Omega)_p = \sup_{|h| \leqslant t} \bigl\| \Delta^r_h f(\cdot, \Omega) \bigm| L_p(\Omega) \bigr\|, \quad t > 0, \tag{1.9}$$

where

$$\Delta^r_h f(x, \Omega) := \begin{cases} \Delta^r_h f(x), & x, x+h, \ldots, x+rh \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{1.10}$$

This approach for the spaces $\mathbf{B}^s_{p,q}(\Omega)$ was used in [8]. The proof of the coincidence uses the fact that the classical and atomic approach can be identified for spaces defined on $\mathbb{R}^n$, which follows from results by Hedberg and Netrusov [16] on atomic decompositions and by Triebel [43, Section 9.2] on the reproducing formula.

The classical scale of Besov spaces contains many well-known function spaces. For example, if $p = q = \infty$, one recovers the Hölder–Zygmund spaces $\mathcal{C}^s(\mathbb{R}^n)$, i.e.,

$$\mathbf{B}^s_{\infty,\infty}\big(\mathbb{R}^n\big) = \mathcal{C}^s\big(\mathbb{R}^n\big), \quad s > 0. \tag{1.11}$$

Later on we will need the following homogeneity estimate proved recently in [35, Th. 2] based on [3].

Theorem 1.8. *Let* $0 < \lambda \leqslant 1$ *and* $f \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ *with* $\operatorname{supp} f \subset B(0,\lambda)$. *Then*

$$\big\| f(\lambda\cdot)\big|\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)\big\| \sim \lambda^{s-n/p}\big\| f\big|\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)\big\|. \tag{1.12}$$

2. Non-smooth atomic decompositions

Our aim is to provide a non-smooth atomic characterization of Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$, i.e., relaxing the assumptions about the smoothness of the atoms $a_{j,m}$ in Definition 1.3. Note that condition (1.4) is equivalent to

$$\big\|a\big(2^{-j}\cdot\big)\big|C^K\big(\mathbb{R}^n\big)\big\| \leqslant 1. \tag{2.1}$$

We replace the C^K-norm with $K > s$ by a Besov quasi-norm $\mathbf{B}^\sigma_{p,p}(\mathbb{R}^n)$ with $\sigma > s$ or in case of $0 < s < 1$ by a norm in the space of Lipschitz functions $\operatorname{Lip}(\mathbb{R}^n)$.

The following non-smooth atoms were introduced in [41]. They will be very adequate when considering (non-smooth) atomic decompositions of spaces defined on Lipschitz domains (or on the boundary of a Lipschitz domain, respectively).

Definition 2.1.

(i) The space of Lipschitz functions $\operatorname{Lip}(\mathbb{R}^n)$ is defined as the collection of all real-valued functions $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\big\| f\big|\operatorname{Lip}\big(\mathbb{R}^n\big)\big\| = \max\left\{\sup_x\big|f(x)\big|, \sup_{x\neq y}\frac{|f(x)-f(y)|}{|x-y|}\right\} < \infty.$$

(ii) We say that $a \in \operatorname{Lip}(\mathbb{R}^n)$ is a Lip-atom, if for some $j \in \mathbb{N}_0$

$$\operatorname{supp} a \subset dQ_{j,m}, \quad m \in \mathbb{Z}^n,\ d > 1, \tag{2.2}$$

and

$$\big|a(x)\big| \leqslant 1, \qquad \big|a(x)-a(y)\big| \leqslant 2^j|x-y|. \tag{2.3}$$

Remark 2.2. One might use alternatively in (2.3) that

$$\| a(2^{-j}\cdot) | \mathrm{Lip}(\mathbb{R}^n) \| \leqslant 1. \tag{2.4}$$

We use the abbreviation

$$\mathbf{B}^s_p(\mathbb{R}^n) = \mathbf{B}^s_{p,p}(\mathbb{R}^n) \quad \text{with } 0 < p \leqslant \infty,\ s > 0.$$

In particular, in view of (1.11),

$$\mathcal{C}^s(\mathbb{R}^n) = \mathbf{B}^s_\infty(\mathbb{R}^n), \quad s > 0,$$

are the Hölder–Zygmund spaces.

Definition 2.3. Let $0 < p \leqslant \infty$, $\sigma > 0$ and $d > 1$. Then $a \in \mathbf{B}^\sigma_p(\mathbb{R}^n)$ is called a (σ, p)-atom if for some $j \in \mathbb{N}_0$

$$\operatorname{supp} a \subset dQ_{j,m} \quad \text{for some } m \in \mathbb{Z}^n, \tag{2.5}$$

and

$$\| a(2^{-j}\cdot) | \mathbf{B}^\sigma_p(\mathbb{R}^n) \| \leqslant 1. \tag{2.6}$$

Remark 2.4. Note that if $\sigma < \frac{n}{p}$ then (σ, p)-atoms might be unbounded. Roughly speaking, they arise by dilating $\mathbf{B}^\sigma_p$-normalized functions. Obviously, the condition (2.6) is a straightforward modification of (2.1) and (2.4).

In general, it is convenient to write $a_{j,m}(x)$ instead of $a(x)$ if the atoms are located at $Q_{j,m}$ according to (2.2) and (2.5), respectively. Furthermore, σ denotes the 'non-smoothness' of the atom, cf. (1.4).

The non-smooth atoms we consider in Definition 2.3, are renormalized versions of the non-smooth $(s, p)^\sigma$-atoms considered in [42] and [46], where (2.6) is replaced by

$$a \in B^\sigma_p(\mathbb{R}^n) \quad \text{with } \| a(2^{-j}\cdot) | B^\sigma_p(\mathbb{R}^n) \| \leqslant 2^{j(\sigma - s)},$$

resulting in corresponding changes concerning the definition of the sequence spaces $b^s_{p,q}$ used for the atomic decomposition.

However, the function spaces we consider are different from the ones considered there. Furthermore, for our purposes (studying traces later on) it is convenient to shift the factors $2^{j(s-\frac{n}{p})}$ to the sequence spaces.

We wish to compare these atoms with the smooth atoms in Definition 1.3.

Proposition 2.5. *Let $0 < p \leqslant \infty$ and $0 < \sigma < K$. Furthermore, let $d > 1$, $j \in \mathbb{N}_0$, and $m \in \mathbb{Z}^n$. Then any K-atom $a_{j,m}$ is a (σ, p)-atom.*

Proof. Since the functions $a_{j,m}(2^{-j}\cdot)$ have compact support, we obtain

$$\| a_{j,m}(2^{-j}\cdot) | \mathbf{B}^\sigma_p(\mathbb{R}^n) \| \lesssim \| a_{j,m}(2^{-j}\cdot) | C^K(\mathbb{R}^n) \| \leqslant 1,$$

with constants independent of j, giving the desired result for non-smooth atoms from Definition 2.3. □

The use of atoms with limited smoothness (i.e. finite element functions or splines) was studied already in [26], where the author deals with spline approximation (and traces) in Besov spaces.

The following theorem contains the main result of this section. It gives the counterpart of Definition 1.4 and provides a non-smooth atomic decomposition of the spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$.

Theorem 2.6. *Let $0 < p, q \leqslant \infty$, $0 < s < \sigma$, and $d > 1$. Then $f \in L_p(\mathbb{R}^n)$ belongs to $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ if, and only if, it can be represented as*

$$f = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} a_{j,m}, \tag{2.7}$$

where the $a_{j,m}$ are (σ, p)-atoms $(j \in \mathbb{N}_0)$ with $\operatorname{supp} a_{j,m} \subset dQ_{j,m}$, $j \in \mathbb{N}_0$, $m \in \mathbb{Z}^n$, and $\lambda \in b^s_{p,q}$, convergence being in $L_p(\mathbb{R}^n)$. Furthermore,

$$\big\| f \big| \mathbf{B}^s_{p,q}(\mathbb{R}^n) \big\| = \inf \big\| \lambda \big| b^s_{p,q} \big\|, \tag{2.8}$$

where the infimum is taken over all admissible representations (2.7).

Proof. We have the atomic decomposition based on smooth K-atoms according to Definition 1.4. By Proposition 2.5 classical K-atoms are special (σ, p)-atoms. Hence, it is enough to prove that

$$\big\| f \big| \mathbf{B}^s_{p,q}(\mathbb{R}^n) \big\| \lesssim \left(\sum_{k=0}^{\infty} 2^{k(s-\frac{n}{p})q} \left(\sum_{l \in \mathbb{Z}^n} |\lambda_{k,l}|^p \right)^{q/p} \right)^{1/q} \tag{2.9}$$

for any atomic decomposition

$$f = \sum_{k=0}^{\infty} \sum_{l \in \mathbb{Z}^n} \lambda_{k,l} a^{k,l}, \tag{2.10}$$

where $a^{k,l}$ are (σ, p)-atoms according to Definition 2.3.

For this purpose we expand each function $a^{k,l}(2^{-k}\cdot)$ optimally in $\mathbf{B}^\sigma_p(\mathbb{R}^n)$ with respect to classical K-atoms $b^{j,w}_{k,l}$ where $\sigma < K$,

$$a^{k,l}\big(2^{-k}x\big) = \sum_{j=0}^{\infty} \sum_{w \in \mathbb{Z}^n} \eta^{k,l}_{j,w} b^{j,w}_{k,l}(x), \quad x \in \mathbb{R}^n, \tag{2.11}$$

with

$$\operatorname{supp} b^{j,w}_{k,l} \subset Q_{j,w}, \qquad \big| D^\alpha b^{j,w}_{k,l}(x) \big| \leqslant 2^{|\alpha| j}, \quad |\alpha| \leqslant K, \tag{2.12}$$

and

$$\left(\sum_{j=0}^{\infty}2^{j(\sigma-\frac{n}{p})p}\sum_{w\in\mathbb{Z}^n}\left|\eta_{j,w}^{k,l}\right|^p\right)^{\frac{1}{p}}=\left\|\eta^{k,l}\middle|b_{p,p}^{\sigma}\right\|\sim\left\|a^{k,l}\left(2^{-k}\cdot\right)\middle|\mathbf{B}_p^{\sigma}\left(\mathbb{R}^n\right)\right\|\lesssim 1. \tag{2.13}$$

Hence,

$$a^{k,l}(x)=\sum_{j=0}^{\infty}\sum_{w\in\mathbb{Z}^n}\eta_{j,w}^{k,l}b_{k,l}^{j,w}\left(2^kx\right),$$

where the functions $b_{k,l}^{j,w}(2^k\cdot)$ are supported by cubes with side lengths $\sim 2^{-k-j}$. By (2.12) we have

$$\left|D^{\alpha}b_{k,l}^{j,w}\left(2^kx\right)\right|=2^{k|\alpha|}\left|\left(D^{\alpha}b_{k,l}^{j,w}\right)\left(2^kx\right)\right|\leqslant 2^{(j+k)|\alpha|}.$$

Replacing $j+k$ by j and putting $d_{k,l}^{j,w}(x):=b_{k,l}^{j-k,w}(2^kx)$, we obtain that

$$a^{k,l}(x)=\sum_{j=k}^{\infty}\sum_{w\in\mathbb{Z}^n}\eta_{j-k,w}^{k,l}d_{k,l}^{j,w}(x), \tag{2.14}$$

where $d_{k,l}^{j,w}$ are classical K-atoms supported by cubes with side lengths $\sim 2^{-j}$. We insert (2.14) into the expansion (2.10). We fix $j\in\mathbb{N}_0$ and $w\in\mathbb{Z}^n$, and collect all non-vanishing terms $d_{k,l}^{j,w}$ in the expansions (2.14). We have $k\leqslant j$. Furthermore, multiplying (2.11) if necessary with suitable cut-off functions it follows that there is a natural number N such that for fixed k only at most N points $l\in\mathbb{Z}^n$ contribute to $d_{k,l}^{j,w}$. We denote this set by (j,w,k). Hence its cardinality is at most N, where N is independent of j,w,k. Then

$$d^{j,w}(x)=\frac{\sum_{k\leqslant j}\sum_{l\in(j,w,k)}\eta_{j-k,w}^{k,l}\cdot\lambda_{k,l}\cdot d_{k,l}^{j,w}(x)}{\sum_{k\leqslant j}\sum_{l\in(j,w,k)}|\eta_{j-k,w}^{k,l}|\cdot|\lambda_{k,l}|}$$

are correctly normalized smooth K-atoms located in cubes with side lengths $\sim 2^{-j}$ and centered at $2^{-j}w$. Let

$$\nu_{j,w}=\sum_{k\leqslant j}\sum_{l\in(j,w,k)}\left|\eta_{j-k,w}^{k,l}\right|\cdot|\lambda_{k,l}|. \tag{2.15}$$

Then we obtain a classical atomic decomposition in the sense of Definition 1.4

$$f=\sum_j\sum_w\nu_{j,w}d^{j,w}(x),$$

where $d^{j,w}$ are K-atoms and

$$\left\|f\middle|\mathbf{B}_{p,q}^s\left(\mathbb{R}^n\right)\right\|\lesssim\left\|\nu\middle|b_{p,q}^s\right\|.$$

Therefore, in order to prove (2.9), it is enough to show, that

$$\|\nu|b^s_{p,q}\| \lesssim \|\lambda|b^s_{p,q}\| \tag{2.16}$$

if (2.13) holds.

Let $0<\varepsilon<\sigma-s$. Then we obtain by (2.15) that (assuming $p<\infty$)

$$|\nu_{j,w}|^p \lesssim \sum_{k\leqslant j}\sum_{l\in(j,w,k)} 2^{(j-k)p\varepsilon}\big|\eta^{k,l}_{j-k,w}\big|^p|\lambda_{k,l}|^p, \tag{2.17}$$

where we used the bounded cardinality of the sets (j,w,k).

This gives for $q/p\leqslant 1$

$$\begin{aligned}
\|\nu|b^s_{p,q}\|^q &= \sum_{j=0}^\infty 2^{j(s-n/p)q}\Big(\sum_{w\in\mathbb{Z}^n}|\nu_{j,w}|^p\Big)^{q/p}\\
&\lesssim \sum_{j=0}^\infty 2^{j(s-n/p)q}\Big(\sum_{w\in\mathbb{Z}^n}\sum_{k=0}^{j}\sum_{l\in(j,w,k)} 2^{(j-k)p\varepsilon}\big|\eta^{k,l}_{j-k,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&\leqslant \sum_{j=0}^\infty 2^{j(s-n/p)q}\sum_{k=0}^{j}\Big(\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j,w,k)} 2^{(j-k)p\varepsilon}\big|\eta^{k,l}_{j-k,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&= \sum_{k=0}^\infty\sum_{j=k}^\infty 2^{j(s-n/p)q}\Big(\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j,w,k)} 2^{(j-k)p\varepsilon}\big|\eta^{k,l}_{j-k,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&= \sum_{k=0}^\infty\sum_{j=0}^\infty 2^{(j+k)(s-n/p)q}\Big(\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j+k,w,k)} 2^{jp\varepsilon}\big|\eta^{k,l}_{j,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&= \sum_{k=0}^\infty 2^{k(s-n/p)q}\sum_{j=0}^\infty 2^{j(s-\sigma+\varepsilon)q}\Big(\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j+k,w,k)} 2^{j(\sigma-n/p)p}\big|\eta^{k,l}_{j,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&\lesssim \sum_{k=0}^\infty 2^{k(s-n/p)q}\Big(\sum_{j=0}^\infty\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j+k,w,k)} 2^{j(\sigma-n/p)p}\big|\eta^{k,l}_{j,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&\leqslant \sum_{k=0}^\infty 2^{k(s-n/p)q}\Big(\sum_{j=0}^\infty\sum_{w\in\mathbb{Z}^n}\sum_{l\in\mathbb{Z}^n} 2^{j(\sigma-n/p)p}\big|\eta^{k,l}_{j,w}\big|^p|\lambda_{k,l}|^p\Big)^{q/p}\\
&= \sum_{k=0}^\infty 2^{k(s-n/p)q}\Big(\sum_{l\in\mathbb{Z}^n}|\lambda_{k,l}|^p\sum_{j=0}^\infty\sum_{w\in\mathbb{Z}^n} 2^{j(\sigma-n/p)p}\big|\eta^{k,l}_{j,w}\big|^p\Big)^{q/p}\\
&\lesssim \sum_{k=0}^\infty 2^{k(s-n/p)q}\Big(\sum_{l\in\mathbb{Z}^n}|\lambda_{k,l}|^p\Big)^{q/p} = \|\lambda|b^s_{p,q}\|^q.
\end{aligned}$$

We have used (2.13) in the last inequality.

If $q/p > 1$, we shall use the following inequality, which holds for every non-negative sequence $\{\gamma_{j,k}\}_{0\leqslant k\leqslant j<\infty}$, every $\alpha \geqslant 1$ and every $\varepsilon > 0$

$$\sum_{j=0}^{\infty}\left(\sum_{k=0}^{j} 2^{-(j-k)\varepsilon}\gamma_{j,k}\right)^{\alpha} \leqslant c_{\alpha,\varepsilon}\sum_{k=0}^{\infty}\left(\sum_{j=k}^{\infty}\gamma_{j,k}\right)^{\alpha}. \tag{2.18}$$

If $\alpha = \infty$, (2.18) has to be modified appropriately. To prove (2.18) for $\alpha < \infty$, we use Hölder's inequality and the embedding $\ell_1 \hookrightarrow \ell_\alpha$

$$\begin{aligned}\sum_{j=0}^{\infty}\left(\sum_{k=0}^{j} 2^{-(j-k)\varepsilon}\gamma_{j,k}\right)^{\alpha} &\leqslant \sum_{j=0}^{\infty}\left(\sum_{k=0}^{j} 2^{-(j-k)\varepsilon\alpha'}\right)^{\alpha/\alpha'}\left(\sum_{k=0}^{j}\gamma_{j,k}^{\alpha}\right)^{\alpha/\alpha}\\ &\lesssim \sum_{j=0}^{\infty}\sum_{k=0}^{j}\gamma_{j,k}^{\alpha} = \sum_{k=0}^{\infty}\sum_{j=k}^{\infty}\gamma_{j,k}^{\alpha} \leqslant \sum_{k=0}^{\infty}\left(\sum_{j=k}^{\infty}\gamma_{j,k}\right)^{\alpha}.\end{aligned}$$

We use (2.17) and (2.18) with $p(\sigma - s - \varepsilon)$ instead of ε and $\alpha = q/p > 1$,

$$\begin{aligned}&\left\|\nu\middle|b^{s}_{p,q}\right\|^{q}\\ &\quad\lesssim \sum_{j=0}^{\infty}2^{j(\sigma-\frac{n}{p})q}\left(\sum_{w\in\mathbb{Z}^n}\sum_{k=0}^{j}\sum_{l\in(j,w,k)}2^{(j-k)p\varepsilon}\left|\eta^{k,l}_{j-k,w}\right|^{p}|\lambda_{k,l}|^{p}\right)^{q/p}\\ &\quad=\sum_{j=0}^{\infty}\left(\sum_{k=0}^{j}2^{-(j-k)p(\sigma-s-\varepsilon)}\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j,w,k)}2^{k(s-n/p)p}2^{(j-k)(\sigma-\frac{n}{p})p}\left|\eta^{k,l}_{j-k,w}\right|^{p}|\lambda_{k,l}|^{p}\right)^{q/p}\\ &\quad\lesssim\sum_{k=0}^{\infty}\left(\sum_{j=k}^{\infty}\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j,w,k)}2^{k(s-n/p)p}2^{(j-k)(\sigma-\frac{n}{p})p}\left|\eta^{k,l}_{j-k,w}\right|^{p}|\lambda_{k,l}|^{p}\right)^{q/p}\\ &\quad=\sum_{k=0}^{\infty}2^{k(s-n/p)q}\left(\sum_{j=0}^{\infty}\sum_{w\in\mathbb{Z}^n}\sum_{l\in(j+k,w,k)}2^{j(\sigma-\frac{n}{p})p}\left|\eta^{k,l}_{j,w}\right|^{p}|\lambda_{k,l}|^{p}\right)^{q/p}\\ &\quad=\sum_{k=0}^{\infty}2^{k(s-n/p)q}\left(\sum_{l\in\mathbb{Z}^n}\sum_{j=0}^{\infty}\sum_{w\in\mathbb{Z}^n:l\in(j+k,w,k)}2^{j(\sigma-\frac{n}{p})p}\left|\eta^{k,l}_{j,w}\right|^{p}|\lambda_{k,l}|^{p}\right)^{q/p}\\ &\quad\lesssim\sum_{k=0}^{\infty}2^{k(s-n/p)q}\left(\sum_{l\in\mathbb{Z}^n}|\lambda_{k,l}|^{p}\sum_{j=0}^{\infty}\sum_{w\in\mathbb{Z}^n}2^{j(\sigma-\frac{n}{p})p}\left|\eta^{k,l}_{j,w}\right|^{p}\right)^{q/p}\\ &\quad\leqslant\sum_{k=0}^{\infty}2^{k(s-n/p)q}\left(\sum_{l\in\mathbb{Z}^n}|\lambda_{k,l}|^{p}\right)^{q/p}=\left\|\lambda\middle|b^{s}_{p,q}\right\|^{q}.\end{aligned}$$

The proof of (2.16) is finished. We again used (2.13) in the last inequality. If p and/or q are equal to infinity, only notational changes are necessary. □

Remark 2.7. Our results generalize [42, Th. 2] and [46, Th. 2.3], where non-smooth atomic decompositions for spaces $\mathbf{B}^s_{p,p}(\mathbb{R}^n)$ with $s > \max(n(1/p - 1), 0)$ can be found, to $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with no restrictions on the parameters. In particular, the case when $p \neq q$ is completely new.

Using the Lip-atoms from Definition 2.1 and the embedding

$$\mathrm{Lip}\big(\mathbb{R}^n\big) \hookrightarrow B^1_\infty\big(\mathbb{R}^n\big),$$

cf. [40, pp. 89, 90], as a corollary we now obtain the following non-smooth atomic decomposition for Besov spaces with smoothness $0 < s < 1$.

Corollary 2.8. *Let $0 < p, q \leqslant \infty$, $0 < s < 1$, and $d > 1$. Then $f \in L_p(\mathbb{R}^n)$ belongs to $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ if, and only if, it can be represented as*

$$f = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} a_{j,m}, \tag{2.19}$$

where the $a_{j,m}$ are Lip-*atoms ($j \in \mathbb{N}_0$) with* $\operatorname{supp} a_{j,m} \subset dQ_{j,m}$, *$j \in \mathbb{N}_0$, $m \in \mathbb{Z}^n$, and $\lambda \in b^s_{p,q}$, convergence being in $L_p(\mathbb{R}^n)$. Furthermore,*

$$\big\| f \big| \mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big) \big\| = \inf \big\| \lambda \big| b^s_{p,q} \big\|, \tag{2.20}$$

where the infimum is taken over all admissible representations (2.19).

3. Spaces on Lipschitz domains and their boundaries

We call a one-to-one mapping $\Phi : \mathbb{R}^n \mapsto \mathbb{R}^n$, a *Lipschitz diffeomorphism*, if the components $\Phi_k(x)$ of $\Phi(x) = (\Phi_1(x), \ldots, \Phi_n(x))$ are Lipschitz functions on $\mathbb{R}^n$ and

$$\big|\Phi(x) - \Phi(y)\big| \sim |x - y|, \quad x, y \in \mathbb{R}^n,\ |x - y| \leqslant 1,$$

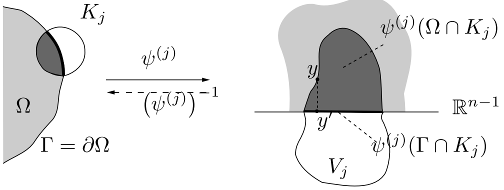
where the equivalence constants are independent of x and y. Of course the inverse of Φ^{-1} is also a Lipschitz diffeomorphism on $\mathbb{R}^n$.

Definition 3.1. Let Ω be a bounded domain in $\mathbb{R}^n$. Then Ω is said to be a Lipschitz domain, if there exist N open balls $K_1, \ldots, K_N$ such that $\bigcup_{j=1}^N K_j \supset \Gamma$ and $K_j \cap \Gamma \neq \emptyset$ if $j = 1, \ldots, N$, with the following property: for every ball K_j there are Lipschitz diffeomorphisms $\psi^{(j)}$ such that

$$\psi^{(j)} : K_j \longrightarrow V_j, \quad j = 1, \ldots, N,$$

where $V_j := \psi^{(j)}(K_j)$ and

$$\psi^{(j)}(K_j \cap \Omega) \subset \mathbb{R}^n_+, \qquad \psi^{(j)}(K_j \cap \Gamma) \subset \mathbb{R}^{n-1}.$$


K_j
$\psi^{(j)}$
$(\psi^{(j)})^{-1}$
Ω
$\Gamma = \partial\Omega$
$\psi^{(j)}(\Omega \cap K_j)$
y
$\mathbb{R}^{n-1}$
y'
$\psi^{(j)}(\Gamma \cap K_j)$
V_j

Remark 3.2. The maps $\psi^{(j)}$ can be extended outside K_j in such a way that the extended vector functions (denoted by $\psi^{(j)}$ as well) yield diffeomorphic mappings from $\mathbb{R}^n$ onto itself (Lipschitz diffeomorphisms).

There are several equivalent definitions of Lipschitz domains in the literature. Our approach follows [5]. Another version as can be found in [37], which defines first a *special* (*unbounded*) *Lipschitz domain* Ω in $\mathbb{R}^n$ as simply the domain above the graph of a Lipschitz function $h : \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$, i.e.,

$$\Omega = \{(x', x_n) \colon h(x') < x_n\}.$$

Then a *bounded Lipschitz domain* Ω in $\mathbb{R}^n$ is defined as a bounded domain where the boundary $\Gamma = \partial\Omega$ can be covered by finitely many open balls B_j in $\mathbb{R}^n$ with $j = 1, \ldots, J$, centered at Γ such that

$$B_j \cap \Omega = B_j \cap \Omega_j \quad \text{for } j = 1, \ldots, J,$$

where Ω_j are rotations of suitable special Lipschitz domains in $\mathbb{R}^n$.

We shall occasionally use this alternative definition, in particular, since it usually suffices to consider special Lipschitz domains in our proofs (the related covering involves only finitely many balls), simplifying the notation considerably.

Consider a covering $\Omega \subset K_0 \cup (\bigcup_{j=1}^{N} K_j)$, where K_0 is an inner domain with $\overline{K}_0 \subset \Omega$. Let $\{\varphi_j\}_{j=0}^{N}$ be a related *resolution of unity* of $\overline{\Omega}$, i.e., φ_j are smooth non-negative functions with support in K_j additionally satisfying

$$\sum_{j=0}^{N} \varphi_j(x) = 1 \quad \text{if } x \in \overline{\Omega}. \tag{3.1}$$

Obviously, the restriction of φ_j to Γ is a resolution of unity with respect to Γ.

3.1. Atomic decompositions for Besov spaces on boundaries

The boundary $\partial\Omega = \Gamma$ of a bounded Lipschitz domain Ω will be furnished in the usual way with a surface measure $\mathrm{d}\sigma$. The corresponding complex-valued Lebesgue spaces $L_p(\Gamma)$, $0 < p \leqslant \infty$, are normed by

$$\|g|L_p(\Gamma)\| = \Bigl(\int_{\Gamma} |g(\gamma)|^p \,\mathrm{d}\sigma(\gamma)\Bigr)^{1/p}$$

(with obvious modifications if $p=\infty$). We require the introduction of Besov spaces on Γ. We rely on the resolution of unity according to (3.1) and the local Lipschitz diffeomorphisms $\psi^{(j)}$ mapping $\Gamma_j = \Gamma \cap K_j$ onto $W_j = \psi^{(j)}(\Gamma_j)$, recall Definition 3.1. We define

$$g_j(y) := (\varphi_j f) \circ \bigl(\psi^{(j)}\bigr)^{-1}(y), \quad j=1,\dots,N,$$

which restricted to $y=(y',0) \in W_j$,

$$g_j\bigl(y'\bigr) = (\varphi_j f) \circ \bigl(\psi^{(j)}\bigr)^{-1}\bigl(y'\bigr), \quad j=1,\dots,N,\ f \in L_p(\Gamma),$$

makes sense. This results in functions $g_j \in L_p(W_j)$ with compact supports in the $(n-1)$-dimensional Lipschitz domain W_j. We do not distinguish notationally between g_j and $(\psi^{(j)})^{-1}$ as functions of $(y',0)$ and of y'.

Our constructions enable us to transport Besov spaces naturally from $\mathbb{R}^{n-1}$ to the boundary Γ of a (bounded) Lipschitz domain via pull-back and a partition of unity.

Definition 3.3. Let $n \geqslant 2$, and let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ with boundary Γ, and φ_j, $\psi^{(j)}$, W_j be as above. Assume $0<s<1$ and $0<p,q\leqslant\infty$. Then we introduce

$$\mathbf{B}^s_{p,q}(\Gamma) = \bigl\{ f \in L_p(\Gamma)\colon\ g_j \in \mathbf{B}^s_{p,q}(W_j),\ j=1,\dots,N \bigr\},$$

equipped with the quasi-norm $\|f|\mathbf{B}^s_{p,q}(\Gamma)\| := \sum_{j=1}^N \|g_j|\mathbf{B}^s_{p,q}(W_j)\|$.

Remark 3.4. The spaces $\mathbf{B}^s_{p,q}(\Gamma)$ turn out to be independent of the particular choice of the resolution of unity $\{\varphi_j\}_{j=1}^N$ and the local diffeomorphisms $\psi^{(j)}$ (the proof is similar to the proof of [40, Prop. 3.2.3(ii)], making use of Propositions 3.11 and 3.12 below). We furnish $\mathbf{B}^s_{p,q}(W_j)$ with the intrinsic $(n-1)$-dimensional norms according to Definition 1.4. Note that we could furthermore replace W_j in the definition of the norm above by $\mathbb{R}^{n-1}$ if we extend g_j outside W_j with zero, i.e.,

$$\bigl\|f\big|\mathbf{B}^s_{p,q}(\Gamma)\bigr\| \sim \sum_{j=1}^N \bigl\|g_j\big|\mathbf{B}^s_{p,q}\bigl(\mathbb{R}^{n-1}\bigr)\bigr\|. \tag{3.2}$$

In particular, the equivalence (3.2) yields that characterizations for B-spaces defined on $\mathbb{R}^{n-1}$ can be generalized to B-spaces defined on Γ. This will be done in Theorem 3.8 for non-smooth atomic decompositions and is very likely to work as well for characterizations in terms of differences.

Atomic decompositions for $\mathbf{B}^s_{p,q}(\Gamma)$ Similarly to the non-smooth atomic decompositions constructed in Section 2 we now establish corresponding atomic decompositions for Besov spaces defined on Lipschitz boundaries. They will be very useful when investigating traces on Lipschitz domains in Section 3.

The relevant sequence spaces and Lipschitz-atoms on the boundary Γ we shall define next are closely related to the sequence spaces $b^s_{p,q}(\Omega)$ and Lip-atoms used for the non-smooth atomic decompositions as used in Corollary 2.8.

Definition 3.5. Let $0 < p, q \leqslant \infty$, $s \in \mathbb{R}$. Furthermore, let Γ be the boundary of a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$, and $\lambda = \{\lambda_{j,m} \in \mathbb{C}\colon\ j \in \mathbb{N}_0,\ m \in \mathbb{Z}^n\}$. Then

$$b^s_{p,q}(\Gamma) = \left\{\lambda\colon\ \big\|\lambda\big|b^s_{p,q}(\Gamma)\big\| = \left(\sum_{j=0}^{\infty} 2^{j(s-\frac{n-1}{p})q}\left(\sum_{m\in\mathbb{Z}^n}^{\Gamma,j} |\lambda_{j,m}|^p\right)^{q/p}\right)^{1/q} < \infty\right\}$$

(with the usual modification if $p = \infty$ and/or $q = \infty$).

Definition 3.6. Let $j \in \mathbb{N}_0$, $m \in \mathbb{Z}^n$, $d > 1$, and let Γ be the boundary of a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$. Put $Q^\Gamma_{j,m} := dQ_{j,m} \cap \Gamma \neq \emptyset$. A function $a \in \mathrm{Lip}(\Gamma)$ is a Lip^Γ-atom, if

$$\operatorname{supp} a \subset Q^\Gamma_{j,m}, \quad d > 1,$$
$$\big\|a|L_\infty(\Gamma)\big\| \leqslant 1 \quad \text{and} \quad \sup_{\substack{x,y\in\Gamma,\\ x\neq y}} \frac{|a(x)-a(y)|}{|x-y|} \leqslant 2^j. \tag{3.3}$$

Remark 3.7. Note that if we put $2^j\Gamma := \{2^j x\colon\ x \in \Gamma\}$, we can state (3.3) like $\|a(2^{-j}\cdot)|\,\mathrm{Lip}(2^j\Gamma)\| \leqslant 1$.

The theorem below provides atomic decompositions for the spaces $\mathbf{B}^s_{p,q}(\Gamma)$.

Theorem 3.8. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and let $0 < s < 1$, $0 < p, q \leqslant \infty$. Then $f \in L_p(\Gamma)$ belongs to $\mathbf{B}^s_{p,q}(\Gamma)$ if, and only if,*

$$f = \sum_{j,m} \lambda_{j,m} a_{j,m},$$

where $a_{j,m}$ are Lip^Γ-atoms with $\operatorname{supp} a_{j,m} \subset Q^\Gamma_{j,m}$ and $\lambda \in b^s_{p,q}(\Gamma)$, convergence being in $L_p(\Gamma)$. Furthermore,

$$\big\|f\big|\mathbf{B}^s_{p,q}(\Gamma)\big\| = \inf\big\|\lambda\big|b^s_{p,q}(\Gamma)\big\|,$$

where the infimum is taken over all possible representations.

Proof. *Step 1*: Fix $f \in \mathbf{B}^s_{p,q}(\Gamma)$. For simplicity, we suppose that $\operatorname{supp} f \subset \{x \in \Gamma\colon\ \varphi_l(x) = 1\}$ for some $l \in \{1, 2, \ldots, N\}$. If this is not the case the arguments have to be slightly modified to incorporate the decomposition of unity (3.1). To simplify the notation we write φ instead of φ_l and ψ instead of $\psi^{(l)}$.

Then we obtain

$$\big\|f\big|\mathbf{B}^s_{p,q}(\Gamma)\big\| = \big\|f \circ \psi^{-1}\big|\mathbf{B}^s_{p,q}\big(\mathbb{R}^{n-1}\big)\big\|.$$

We use Corollary 2.8 with n replaced by $n-1$ to obtain an optimal atomic decomposition

$$f \circ \psi^{-1} = \sum_{j,m} \lambda_{j,m} a_{j,m} \quad \text{where } \big\|f \circ \psi^{-1}\big|\mathbf{B}^s_{p,q}\big(\mathbb{R}^{n-1}\big)\big\| \sim \big\|\lambda\big|b^s_{p,q}\big(\mathbb{R}^{n-1}\big)\big\|. \tag{3.4}$$

For $j \in \mathbb{N}_0$ and $m \in \mathbb{Z}^{n-1}$ fixed, we consider the function $a_{j,m}(\psi(x))$. Due to the Lipschitz properties of ψ, this function is supported in $Q^{\Gamma}_{j,l}$ for some $l \in \mathbb{Z}^n$ and we denote it by $a^{\Gamma}_{j,l}(x)$. Furthermore, we set $\lambda'_{j,l} = \lambda_{j,m}$. This leads to the decomposition

$$f = \sum_{j,l} \lambda'_{j,l} a^{\Gamma}_{j,l}. \tag{3.5}$$

It is straightforward to verify that $a^{\Gamma}_{j,l}$ are Lip^{Γ}-atoms since $\|a^{\Gamma}_{j,l}|L_\infty(\Gamma)\| \lesssim \|a_{j,m}|L_\infty(W_l)\| \lesssim 1$ and

$$\frac{|a^{\Gamma}_{j,l}(x) - a^{\Gamma}_{j,l}(y)|}{|x-y|} = \frac{|a_{j,m}(x') - a_{j,m}(y')|}{|\psi^{-1}(x') - \psi^{-1}(y')|} \sim \frac{|a_{j,m}(x') - a_{j,m}(y')|}{|x'-y'|} \lesssim 2^j, \quad x, y \in \Gamma.$$

Furthermore, we have the estimate

$$\|f|\mathbf{B}^s_{p,q}(\Gamma)\| = \|f \circ \psi^{-1}|\mathbf{B}^s_{p,q}(\mathbb{R}^{n-1})\| \sim \|\lambda|b^s_{p,q}(\mathbb{R}^{n-1})\| = \|\lambda'|b^s_{p,q}(\Gamma)\|.$$

Step 2: The proof of the opposite direction follows along the same lines. If f on Γ is given by

$$f = \sum_{j,l} \lambda'_{j,l} a^{\Gamma}_{j,l},$$

then $f \circ \psi^{-1} = \sum_{j,m} \lambda_{j,m} a_{j,m}$, where $a_{j,m}(x) = a^{\Gamma}_{j,l}(\psi^{-1}(x))$ and $\lambda_{j,m} = \lambda'_{j,l}$ for suitable $m \in \mathbb{Z}^{n-1}$. Again it follows that $a_{j,m}$ are Lip-atoms on $\mathbb{R}^{n-1}$ and

$$\|f|\mathbf{B}^s_{p,q}(\Gamma)\| = \|f \circ \psi^{-1}|\mathbf{B}^s_{p,q}(\mathbb{R}^{n-1})\| \lesssim \|\lambda|b^s_{p,q}(\mathbb{R}^{n-1})\| = \|\lambda'|b^s_{p,q}(\Gamma)\|.$$

Step 3: The convergence in $L_p(\Gamma)$ of the representation $f = \sum_{j,m}^{j,\Gamma} \lambda_{j,m} a^{\Gamma}_{j,m}$, follows for $p \leqslant 1$ by

$$\begin{aligned}\Big\|\sum_{j,m}^{j,\Gamma} \lambda_{j,m} a^{\Gamma}_{j,m}\Big|L_p(\Gamma)\Big\|^p &\leqslant \sum_{j,m}^{j,\Gamma} |\lambda_{j,m}|^p \big\|a^{\Gamma}_{j,m}\big|L_p(\Gamma)\big\|^p \\ &\lesssim \sum_j 2^{-j(n-1)} \sum_m^{j,\Gamma} |\lambda_{j,m}|^p = \big\|\lambda\big|b^0_{p,p}(\Gamma)\big\|^p \\ &\lesssim \big\|\lambda\big|b^s_{p,q}(\Gamma)\big\|^p \end{aligned} \tag{3.6}$$

and using

$$\begin{aligned}\Big\|\sum_{j,m}^{j,\Gamma} \lambda_{j,m} a^{\Gamma}_{j,m}\Big|L_p(\Gamma)\Big\| &\leqslant \sum_j \Big\|\sum_m^{j,\Gamma} \lambda_{j,m} a^{\Gamma}_{j,m}\Big|L_p(\Gamma)\Big\| \\ &\lesssim \sum_j 2^{-j(n-1)/p} \Big(\sum_m^{j,\Gamma} |\lambda_{j,m}|^p\Big)^{1/p} \\ &= \big\|\lambda\big|b^0_{p,1}(\Gamma)\big\| \lesssim \big\|\lambda\big|b^s_{p,q}(\Gamma)\big\| \end{aligned} \tag{3.7}$$

for $p > 1$. □

3.2. Interpolation results

Interpolation results for $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ as obtained in [7, Cor. 6.2, 6.3] carry over to the spaces $\mathbf{B}^s_{p,q}(\Gamma)$, which follows immediately from their definition and properties of real interpolation.

Theorem 3.9. *Let Ω be a bounded Lipschitz domain with boundary Γ.*

(i) *Let $0 < p, q, q_0, q_1 \leqslant \infty$, $s_0 \neq s_1$, and $0 < s_i < 1$. Then*

$$\big(\mathbf{B}^{s_0}_{p,q_0}(\Gamma), \mathbf{B}^{s_1}_{p,q_1}(\Gamma)\big)_{\theta,q} = \mathbf{B}^s_{p,q}(\Gamma),$$

where $0 < \theta < 1$ and $s = (1-\theta)s_0 + \theta s_1$.

(ii) *Let $0 < p_i, q_i \leqslant \infty$, $s_0 \neq s_1$ and $0 < s_i < 1$. Then for each $0 < \theta < 1$, $s = (1-\theta)s_0 + \theta s_1$, $\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$, and for $\frac{1}{q} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1}$ we have*

$$\big(\mathbf{B}^{s_0}_{p_0,q_0}(\Gamma), \mathbf{B}^{s_1}_{p_1,q_1}(\Gamma)\big)_{\theta,q} = \mathbf{B}^s_{p,q}(\Gamma),$$

provided $p = q$.

Proof. By definition of the spaces $\mathbf{B}^s_{p,q}(\Gamma)$ we can construct a well-defined and bounded linear operator

$$E : \mathbf{B}^s_{p,q}(\Gamma) \longrightarrow \bigoplus_{1\leqslant j\leqslant N} \mathbf{B}^s_{p,q}\big(\mathbb{R}^{n-1}\big),$$

$$(Ef)_j := (\varphi_j f) \circ \psi^{(j)^{-1}} \quad \text{on } \mathbb{R}^{n-1},\ 1 \leqslant j \leqslant N,$$

which has a bounded and linear left inverse given by

$$R : \bigoplus_{1\leqslant j\leqslant N} \mathbf{B}^s_{p,q}\big(\mathbb{R}^{n-1}\big) \longrightarrow \mathbf{B}^s_{p,q}(\Gamma),$$

$$R\big((g_j)_{1\leqslant j\leqslant N}\big) := \sum_{j=1}^N \Psi_j (g_j \circ \psi_j) \quad \text{on } \Gamma,$$

where $\Psi_j \in C_0^\infty(\mathbb{R}^n)$, $\operatorname{supp}\Psi_j \subseteq K_j$, $\Psi \equiv 1$ in a neighborhood of $\operatorname{supp}\varphi_j$.

A straightforward calculation shows for $f \in \mathbf{B}^s_{p,q}(\Gamma)$

$$(R \circ E)f = R(Ef) = R\big(\big((\varphi_j f) \circ \psi^{(j)^{-1}}\big)_{1\leqslant j\leqslant N}\big) = \sum_{j=1}^N \Psi_j \varphi_j f = \sum_{j=1}^N \varphi_j f = f,$$

i.e.,

$$R \circ E = I, \quad \text{the identity operator on } \mathbf{B}^s_{p,q}(\Gamma).$$

One arrives at a standard situation in interpolation theory. Hence, by the method of retraction–coretraction, cf. [39, Sect. 1.2.4, 1.17.1], the results for $\mathbf{B}^s_{p,q}(\mathbb{R}^{n-1})$ carry over to the spaces $\mathbf{B}^s_{p,q}(\varGamma)$. Therefore, (i) and (ii) are a consequence of [7, Cor. 6.2, 6.3]. □

Furthermore, we briefly show that the interpolation results for Besov spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ also hold for spaces on domains $\mathbf{B}^s_{p,q}(\varOmega)$. This is not automatically clear in our context since the extension operator

$$\mathrm{Ex}:\mathbf{B}^s_{p,q}(\varOmega)\longrightarrow\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)$$

constructed in [8] is not linear. The situation is different for spaces $B^s_{p,q}(\varOmega)$. Here Rychkov's (linear) extension operator, cf. [29], automatically yields interpolation results for B-spaces on domains.

Theorem 3.10. *Let $\varOmega$ be a bounded Lipschitz domain.*

(i) *Let $0<p,q,q_0,q_1\leqslant\infty$, $s_0\neq s_1$, and $0<s_i<1$. Then*

$$\big(\mathbf{B}^{s_0}_{p,q_0}(\varOmega),\mathbf{B}^{s_1}_{p,q_1}(\varOmega)\big)_{\theta,q}=\mathbf{B}^s_{p,q}(\varOmega),$$

where $0<\theta<1$ and $s=(1-\theta)s_0+\theta s_1$.

(ii) *Let $0<p_i,q_i\leqslant\infty$, $s_0\neq s_1$ and $0<s_i<1$. Then for each $0<\theta<1$, $s=(1-\theta)s_0+\theta s_1$, $\frac{1}{p}=\frac{1-\theta}{p_0}+\frac{\theta}{p_1}$, and for $\frac{1}{q}=\frac{1-\theta}{q_0}+\frac{\theta}{q_1}$ we have*

$$\big(\mathbf{B}^{s_0}_{p_0,q_0}(\varOmega),\mathbf{B}^{s_1}_{p_1,q_1}(\varOmega)\big)_{\theta,q}=\mathbf{B}^s_{p,q}(\varOmega),$$

provided $p=q$.

Proof. In spite of our remarks before the theorem, we can nevertheless use the extension operator

$$\mathrm{Ex}:\mathbf{B}^s_{p,q}(\varOmega)\longrightarrow\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)$$

constructed in [8] to show that interpolation results for spaces $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ carry over to spaces $\mathbf{B}^s_{p,q}(\varOmega)$. Let $X_i(\varOmega):=\mathbf{B}^{s_i}_{p_i,q_i}(\varOmega)$. By the explanations given in [8, p. 859] we have the estimate

$$K\big(f,t,X_0(\varOmega),X_1(\varOmega)\big)\sim K\big(\mathrm{Ex}\,f,t,X_0\big(\mathbb{R}^n\big),X_1\big(\mathbb{R}^n\big)\big) \tag{3.8}$$

although the operator Ex is not linear. Let $\mathbf{B}^\theta(\varOmega):=(\mathbf{B}^{s_0}_{p_0,q_0}(\varOmega),\mathbf{B}^{s_1}_{p_1,q_1}(\varOmega))_{\theta,q}$ with the given restrictions on the parameters given in (i) and (ii), respectively. We have to prove that

$$\mathbf{B}^\theta(\varOmega)=\mathbf{B}^s_{p,q}(\varOmega),$$

but this follows immediately from [7, Cor. 6.2, 6.3] using (3.8), since

$$\big\|f\big|\mathbf{B}^\theta(\varOmega)\big\|\sim\big\|\mathrm{Ex}\,f\big|\mathbf{B}^\theta\big(\mathbb{R}^n\big)\big\|\sim\big\|\mathrm{Ex}\,f\big|\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)\big\|\sim\big\|f\big|\mathbf{B}^s_{p,q}(\varOmega)\big\|.\qquad\Box$$

3.3. Properties of Besov spaces on Lipschitz domains

The non-smooth atomic decomposition enables us to generalize [32, Prop. 2.5] and obtain new results concerning diffeomorphisms and pointwise multipliers in $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ in the following way. For related matters we also refer to [21, Th. 3.3.3].

Proposition 3.11. *Let* $0 < p, q \leqslant \infty$, $0 < s < 1$ *and* $\sigma > s$.

- (i) (*Diffeomorphisms*) *Let* ψ *be a Lipschitz diffeomorphism. Then* $f \longrightarrow f \circ \psi$ *is a linear and bounded operator from* $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ *onto itself.*
- (ii) (*Pointwise multipliers*) *Let* $h \in \mathcal{C}^\sigma(\mathbb{R}^n)$. *Then* $f \longrightarrow hf$ *is a linear and bounded operator from* $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ *into itself.*

Proof. Concerning (i), we make use of the atomic decomposition as in (2.19) with the Lip-atoms from Definition 2.1. Then we have

$$f \circ \psi = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} a_{j,m} \circ \psi$$

and $a \circ \psi$ is a Lip-atom based on a new cube, and multiplied with a constant depending on ψ, since

$$\bigl|(a_{j,m} \circ \psi)(x) - (a_{j,m} \circ \psi)(y)\bigr| \leqslant 2^j \bigl|\psi(x) - \psi(y)\bigr| \lesssim 2^j |x - y|.$$

To prove (ii) we argue as follows. First, we may suppose that $0 < s < \sigma < 1$. Furthermore, we choose a real parameter σ' with $s < \sigma' < \sigma$. We take the smooth atomic decomposition (1.5) with K-atoms $a_{j,m}$, where $K = 1$. Multiplied with $h \in \mathcal{C}^\sigma$, it gives a new (non-smooth) atomic decomposition of hf. Its convergence in $L_p(\mathbb{R}^n)$ follows from the convergence of (1.5) in $L_p(\mathbb{R}^n)$ and the boundedness of h.

It remains to verify, that $ha_{j,m}$ are non-smooth (σ', p)-atoms. The support property follows immediately from the support property of $a_{j,m}$. We use the bounded support of $(ha_{j,m})(2^{-j}\cdot)$ and the multiplier assertion for $\mathbf{B}^\sigma_\infty(\mathbb{R}^n)$ as presented in [28, Section 4.6.1, Theorem 2] to get

$$\begin{aligned} \bigl\|(ha_{j,m})(2^{-j}\cdot)\bigl|\mathbf{B}^{\sigma'}_p(\mathbb{R}^n)\bigr\| &\leqslant \bigl\|(ha_{j,m})(2^{-j}\cdot)\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\| \\ &= \bigl\|h(2^{-j}\cdot) \cdot a_{j,m}(2^{-j}\cdot)\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\| \\ &\lesssim \bigl\|h(2^{-j}\cdot)\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\| \cdot \bigl\|a_{j,m}(2^{-j}\cdot)\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\|. \end{aligned}$$

The last product is bounded by a constant due to the inequality

$$\bigl\|h(2^{-j}\cdot)\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\| \lesssim \bigl\|h\bigl|\mathbf{B}^\sigma_\infty(\mathbb{R}^n)\bigr\|, \quad j \in \mathbb{N}_0,$$

which may be verified directly (or found in [1, Section 1.7] or [10, Section 2.3.1]), combined with the fact that $a_{j,m}$ are K-atoms for $K = 1$. □

Furthermore, we establish an equivalent quasi-norm for $\mathbf{B}^s_{p,q}(\Omega)$.

Proposition 3.12. *Let $0 < p, q \leqslant \infty$, $0 < s < 1$, and Ω be a bounded Lipschitz domain. Then*

$$\left\|\varphi_0 f\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| + \sum_{j=1}^N \left\|(\varphi_j f)\left(\psi^{(j)}(\cdot)\right)^{-1}\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n_+\right)\right\| \tag{3.9}$$

is an equivalent quasi-norm in $\mathbf{B}^s_{p,q}(\Omega)$.

Proof. Let Ω_1 be a bounded domain with

$$\overline{\Omega}_1 \subset \left\{x \in \mathbb{R}^n\colon \sum_{j=0}^N \varphi_j(x) = 1\right\}$$

and $\overline{\Omega} \subset \Omega_1$. Let $f \in \mathbf{B}^s_{p,q}(\Omega)$. If we restrict the infimum in (1.5) to $g \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with

$$g|_\Omega = f \quad \text{and} \quad \operatorname{supp} g \subset \Omega_1, \tag{3.10}$$

then we obtain a new equivalent quasi-norm in $\mathbf{B}^s_{p,q}(\Omega)$. This follows from Proposition 3.11(ii) if one multiplies an arbitrary element $g \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with a fixed infinitely differentiable function $\varkappa(x)$ with

$$\varkappa(x) = 1 \quad \text{if } x \in \Omega \quad \text{and} \quad \operatorname{supp} \varkappa \subset \Omega_1.$$

For elements $g \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with (3.10),

$$\sum_{k=0}^N \left\|\varphi_k g\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\|$$

is an equivalent quasi-norm. This is also a consequence of Proposition 3.11(ii). Applying part (i) of that proposition to $g(x) \to g(\psi^{(j)}(x))$, we see that

$$\left\|\varphi_0 g\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| + \sum_{k=1}^N \left\|(\varphi_k g)\left(\psi^{(k)}(\cdot)\right)^{-1}\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\|$$

is an equivalent quasi-norm for all $g \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with (3.10). But the infimum over all admissible g with (3.10) yields (3.9). □

4. Trace results on Lipschitz domains

Now we can look for traces of $f \in \mathbf{B}^s_{p,q}(\Omega)$ on the boundary Γ. We briefly explain our understanding of the trace operator since when dealing with $L_p(\mathbb{R}^n)$ functions the pointwise trace has no obvious meaning.

Let $Y(\Gamma)$ denote one of the spaces $\mathbf{B}^\sigma_{u,v}(\Gamma)$ or $L_u(\Gamma)$. Since $\mathcal{S}(\Omega)$ is dense in $\mathbf{B}^s_{p,q}(\Omega)$ for $0 < p, q < \infty$ (both spaces can be interpreted as restrictions of their counterparts defined on $\mathbb{R}^n$), one asks first whether there is a constant $c > 0$ such that

$$\left\|\operatorname{Tr}\varphi\middle|Y(\Gamma)\right\| \leqslant c\left\|\varphi\middle|\mathbf{B}^s_{p,q}(\Omega)\right\| \quad \text{for all } \varphi \in \mathcal{S}(\Omega), \tag{4.1}$$

where $\mathcal{S}(\Omega)$ stands for the restriction of the Schwartz space $\mathcal{S}(\mathbb{R}^n)$ to a domain Ω. If this is the case, then one defines $\operatorname{Tr} f \in Y(\Gamma)$ for $f \in \mathbf{B}^s_{p,q}(\Omega)$ by completion and obtains

$$\left\|\operatorname{Tr} f \middle| Y(\Gamma)\right\| \leqslant c\left\| f \middle| \mathbf{B}^s_{p,q}(\Omega)\right\|, \quad f \in \mathbf{B}^s_{p,q}(\Omega),$$

for the linear and bounded trace operator

$$\operatorname{Tr} : \mathbf{B}^s_{p,q}(\Omega) \hookrightarrow Y(\Gamma).$$

Remark 4.1. We can extend (4.1) to spaces $\mathbf{B}^s_{p,q}(\Omega)$ with $p = \infty$ and/or $q = \infty$ by using embeddings for B- and F-spaces from [17,31]. The results stated there can be generalized to domains Ω, since the spaces $\mathbf{B}^s_{p,q}(\Omega)$ are defined by restriction of the corresponding spaces on $\mathbb{R}^n$, cf. Remark 1.5.

If $p = \infty$, we have that $\mathbf{B}^s_{\infty,q}(\Omega)$ with $s > 0$ is embedded in the space of continuous functions and Tr makes sense pointwise. If $q = \infty$,

$$\mathbf{B}^s_{p,\infty}(\Omega) \hookrightarrow \mathbf{B}^{s-\varepsilon}_{p,1}(\Omega) \quad \text{for any } \varepsilon > 0.$$

Let $s > \frac{1}{p}$ and $\varepsilon > 0$ be small enough such that one has

$$s > s - \varepsilon > \frac{1}{p}.$$

Since by [44, Rem. 13] traces are independent of the source spaces and of the target spaces one can now define Tr for $\mathbf{B}^s_{p,\infty}(\Omega)$ by restriction of Tr for $\mathbf{B}^{s-\varepsilon}_{p,1}(\Omega)$ to $\mathbf{B}^s_{p,\infty}(\Omega)$. Hence (4.1) is always meaningful.

4.1. Boundedness of the trace operator

Now we are able to state and prove our first main theorem concerning traces of Besov spaces on Lipschitz domains.

Theorem 4.2. *Let $n \geqslant 2$, $0 < p, q \leqslant \infty$, $0 < s < 1$, and let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ with boundary Γ. Then the operator*

$$\operatorname{Tr} : \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega) \longrightarrow \mathbf{B}^s_{p,q}(\Gamma) \tag{4.2}$$

is linear and bounded.

Proof. The linearity of the operator follows directly from its definition as discussed above. To prove the boundedness, we take an optimal representation of a smooth function $f \in \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega)$ as described in (1.5), i.e.,

$$f = \sum_{j=0}^{\infty} \sum_{m\in\mathbb{Z}^n}^{j,\Omega} \lambda_{j,m} a_{j,m} \quad \text{with } \left\| f \middle| \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega)\right\| \sim \left\|\lambda \middle| b^{s+\frac{1}{p}}_{p,q}(\Omega)\right\|. \tag{4.3}$$

We put

$$\operatorname{Tr} f := \left(\sum_{j,m}^{j,\Omega} \lambda_{j,m} a_{j,m}\right)\Bigg|_{\Gamma} = \sum_{j,m}^{j,\Gamma} \lambda_{j,m} a_{j,m}\Bigg|_{\Gamma} = \sum_{j,m}^{j,\Gamma} \lambda_{j,m} a_{j,m}^{\Gamma}. \tag{4.4}$$

The proof follows by Theorem 3.8 and the following four facts:

(i) $a_{j,m}^{\Gamma}$ are Lip^{Γ}-atoms,
(ii) $\|\lambda|b_{p,q}^{s}(\Gamma)\| \lesssim \|\lambda|b_{p,q}^{s+\frac{1}{p}}(\Omega)\|$,
(iii) the decomposition (4.4) converges in $L_p(\Gamma)$,
(iv) the trace operator Tr coincides with the trace operator discussed above.

To prove the first point, we observe that

$$\operatorname{supp} a_{j,m}^{\Gamma} \subseteq \operatorname{supp} a_{j,m} \cap \Gamma \subseteq Q_{j,m}^{\Gamma}.$$

Furthermore, we have $\|a_{j,m}^{\Gamma}|L_\infty(\Gamma)\| \leqslant \|a_{j,m}|L_\infty(dQ_{j,m})\| \leqslant c$ and

$$\sup_{\substack{x,y\in Q_{j,m}^{\Gamma}\\ x\neq y}} \frac{a_{j,m}^{\Gamma}(x) - a_{j,m}^{\Gamma}(y)}{|x-y|} \leqslant \sup_{\substack{x,y\in dQ_{j,m}\\ x\neq y}} \frac{a_{j,m}(x) - a_{j,m}(y)}{|x-y|} \lesssim 2^{j}.$$

The proof of the second point follows directly by

$$\begin{aligned}\big\|\lambda\big|b_{p,q}^{s}(\Gamma)\big\| &= \left(\sum_{j} 2^{j(s-\frac{n-1}{p})q}\left(\sum_{m}^{j,\Gamma} |\lambda_{j,m}|^{p}\right)^{q/p}\right)^{1/p} \\ &\leqslant \left(\sum_{j} 2^{j[(s+\frac{1}{p})-\frac{n}{p}]q}\left(\sum_{m}^{j,\Omega} |\lambda_{j,m}|^{p}\right)^{q/p}\right)^{1/p} = \big\|\lambda\big|b_{p,q}^{s+\frac{1}{p}}(\Omega)\big\|.\end{aligned}$$

The proof of the third point follows in the same way as the proof in Step 3 of Theorem 3.8.

The proof of (iv) is based on the fact that for $f \in \mathcal{S}(\Omega)$ there is an optimal atomic decomposition (4.3) which converges also pointwise. This may be observed by a detailed inspection of [16]. Therefore also the series (4.4) converges pointwise and the trace operator Tr may be understood in the pointwise sense for smooth f. □

4.2. *Extension of atoms*

In order to compute the exact trace space we still need to construct an extension operator

$$\operatorname{Ext} : \mathbf{B}_{p,q}^{s}(\Gamma) \longrightarrow \mathbf{B}_{p,q}^{s+\frac{1}{p}}(\Omega)$$

and show its boundedness. The main problem will be to show that we can extend the Lip^{Γ}-atoms from the source spaces in a nice way to obtain suitable atoms for the target spaces. We start with a simple variant of the Gagliardo–Nirenberg inequality, cf. [27, Chapter 5].

Lemma 4.3. *Let $0 < s_0, s_1 < \infty$, $0 < p_0, p_1, q_0, q_1 \leqslant \infty$ and $0 < \theta < 1$. Put*

$$s = (1-\theta)s_0 + \theta s_1, \qquad \frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}, \qquad \frac{1}{q} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1}. \tag{4.5}$$

Then

$$\| f | \mathbf{B}^s_{p,q}(\Omega) \| \lesssim \| f | \mathbf{B}^{s_0}_{p_0,q_0}(\Omega) \|^{1-\theta} \cdot \| f | \mathbf{B}^{s_1}_{p_1,q_1}(\Omega) \|^{\theta} \tag{4.6}$$

for all $f \in \mathbf{B}^{s_0}_{p_0,q_0}(\Omega) \cap \mathbf{B}^{s_1}_{p_1,q_1}(\Omega)$.

Proof. The straightforward proof uses the characterization of B-spaces through differences and Hölder's inequality. □

Our approach is based on the classical Whitney decomposition of $\mathbb{R}^n \setminus \Gamma$ and the corresponding decomposition of unity. We summarize the most important properties of this method in the next lemma and refer to [37, pp. 167–170] and [19, pp. 21–26] for details and proofs.

Lemma 4.4. 1. *Let $\Gamma \subset \mathbb{R}^n$ be a closed set. Then there exists a collection of cubes $\{Q_i\}_{i\in\mathbb{N}}$, such that*

(i) $\mathbb{R}^n \setminus \Gamma = \bigcup_i Q_i$.
(ii) *The interiors of the cubes are mutually disjoint.*
(iii) *The inequality*

$$\operatorname{diam} Q_i \leqslant \operatorname{dist}(Q_i, \Gamma) \leqslant 4 \operatorname{diam} Q_i$$

holds for every cube Q_i. Here $\operatorname{diam} Q_i$ *is the diameter of Q_i and* $\operatorname{dist}(Q_i, \Gamma)$ *is its distance from Γ.*
(iv) *Each point of $\mathbb{R}^n \setminus \Gamma$ is contained in at most N_0 cubes $6/5 \cdot Q_i$, where N_0 depends only on n.*
(v) *If Γ is the boundary of a Lipschitz domain then there is a number $\gamma > 0$, which depends only on n, such that $\sigma(\gamma Q_i \cap \Gamma) > 0$ for all $i \in \mathbb{N}$.*

2. *The are C^∞-functions $\{\psi_i\}_{i\in\mathbb{N}}$ such that*

(i) $\sum_i \psi_i(x) = 1$ *for every $x \in \mathbb{R}^n \setminus \Gamma$.*
(ii) $\operatorname{supp} \psi_i \subset 6/5 \cdot Q_i$.
(iii) *For every $\alpha \in \mathbb{N}_0^n$ there is a constant A_α such that $|D^\alpha \psi_i(x)| \leqslant A_\alpha (\operatorname{diam} Q_i)^{-|\alpha|}$ holds for all $i \in \mathbb{N}$ and all $x \in \mathbb{R}^n$.*

If a is a Lipschitz function on the Lipschitz boundary Γ of Ω, then the Whitney extension operator Ext is defined by

$$\operatorname{Ext} a(x) = \begin{cases} a(x), & x \in \Gamma, \\ \sum_i \mu_i \psi_i(x), & x \in \Omega, \end{cases} \tag{4.7}$$

where we use the notation of Lemma 4.4 and $\mu_i := \frac{1}{\sigma(\gamma Q_i \cap \Gamma)} \int_{\gamma Q_i \cap \Gamma} a(y)\, d\sigma(y)$ with the number $\gamma > 0$ as described in Lemma 4.4. It satisfies $\operatorname{Tr} \circ \operatorname{Ext} a = a$ for a Lipschitz continuous on Γ.

This follows directly from the celebrated Whitney's extension theorem (cf. [19, p. 23]) as Γ is a closed set if Ω is a bounded Lipschitz domain.

Lemma 4.5. *Let a be a Lipschitz function on the Lipschitz boundary Γ of Ω. Then* $\operatorname{Ext} a \in C^\infty(\Omega)$ *and*

$$\max_{|\alpha|=k} \left| D^\alpha \operatorname{Ext} a(x) \right| \leqslant c_k \delta(x)^{1-k} \cdot \left\| a \,|\, \operatorname{Lip}(\Gamma) \right\|, \quad k \in \mathbb{N},\ x \in \Omega. \tag{4.8}$$

Here, $\delta(x)$ is the distance of x to Γ and c_k depends only on k and Ω.

Proof. First, let us note that

$$D^\alpha \operatorname{Ext} a(x) = \sum_i \mu_i D^\alpha \psi_i(x), \quad x \in \Omega,\ \alpha \in \mathbb{N}_0^n,\ |\alpha| = k.$$

By Lemma 4.4 we have for every $x \in \Omega$

$$\left| D^\alpha \psi_i(x) \right| \leqslant c_k \delta(x)^{-k}, \quad |\alpha| = k,$$

and

$$\sum_i D^\alpha \psi_i(x) = D^\alpha \sum_i \psi_i(x) = 0.$$

Furthermore, the Lipschitz continuity of a implies

$$|\mu_i - \mu_j| \lesssim \delta(x) \cdot \left\| a \,|\, \operatorname{Lip}(\Gamma) \right\| \tag{4.9}$$

for $x \in \operatorname{supp} \psi_i \cap \operatorname{supp} \psi_j$. To justify (4.9), we consider natural numbers i and j with $x \in \operatorname{supp} \psi_i \cap \operatorname{supp} \psi_j$, choose any $x_i \in \gamma Q_i \cap \Gamma$ and $x_j \in \gamma Q_j \cap \Gamma$ and calculate

$$\begin{aligned}
|\mu_i - \mu_j| &\leqslant \left| \frac{1}{\sigma(\gamma Q_i \cap \Gamma)} \int_{\gamma Q_i \cap \Gamma} a(x)\, \mathrm{d}\sigma(x) - a(x_i) \right| + \left| a(x_i) - a(x_j) \right| \\
&\quad + \left| a(x_j) - \frac{1}{\sigma(\gamma Q_j \cap \Gamma)} \int_{\gamma Q_j \cap \Gamma} a(x)\, \mathrm{d}\sigma(x) \right| \\
&\leqslant \left\| a \,|\, \operatorname{Lip}(\Gamma) \right\| \cdot \left\{ \operatorname{diam}(\gamma Q_i \cap \Gamma) + |x_i - x_j| + \operatorname{diam}(\gamma Q_j \cap \Gamma) \right\} \\
&\lesssim \left\| a \,|\, \operatorname{Lip}(\Gamma) \right\| \cdot \left\{ \operatorname{diam}(Q_i) + |x_i - x| + |x - x_j| + \operatorname{diam}(Q_j) \right\} \\
&\lesssim \delta(x) \cdot \left\| a \,|\, \operatorname{Lip}(\Gamma) \right\|.
\end{aligned}$$

Let us now fix $x \in \Omega$ and let us denote by $\{i_1, \ldots, i_N\}$, $N \leqslant N_0$, the indices for which x lies in the support of ψ_i. Then we write

$$\left|\sum_{j=1}^{N}\mu_{i_j}D^\alpha\psi_{i_j}(x)\right| \leqslant \left|\sum_{j=1}^{N}(\mu_{i_j}-\mu_{i_1})D^\alpha\psi_{i_j}(x)\right| + \left|\sum_{j=1}^{N}\mu_{i_1}D^\alpha\psi_{i_j}(x)\right|$$
$$\leqslant \sum_{j=1}^{N}|\mu_{i_j}-\mu_{i_1}|\cdot\left|D^\alpha\psi_{i_j}(x)\right| \lesssim \delta(x)^{1-k}\cdot\left\|a|\operatorname{Lip}(\Gamma)\right\|. \qquad \square$$

Remark 4.6. Let a be a function defined on Γ as in Lemma 4.5 with $\operatorname{diam}(\operatorname{supp} a)\leqslant 1$. Then the extension operator from Lemma 4.5 may be combined with a multiplication with a smooth cut-off function. This ensures, that (4.8) still holds and, in addition, $\operatorname{diam}(\operatorname{supp}\operatorname{Ext} a)\lesssim 1$.

The following lemma describes a certain geometrical property of Lipschitz domains, which shall be useful later on. It resembles very much the notion of Minkowski content, cf. [11].

Lemma 4.7. *Let Ω be a bounded Lipschitz domain and let $k\in\mathbb{N}$. Let $h\in\mathbb{R}^n$ with $0<|h|\leqslant 1$ and put $\Omega^h=\{x\in\Omega\colon [x,x+kh]\subset\Omega\}$. Furthermore, for $j\in\mathbb{N}_0$ we define $\Omega^h_j=\{x\in\Omega^h\colon 2^{-j}\leqslant \min_{y\in[x,x+kh]}\delta(y)\leqslant 2^{-j+1}\}$, where $\delta(y)=\operatorname{dist}(y,\Gamma)$. Then*

$$\left|\Omega^h_j\right|\lesssim 2^{-j} \tag{4.10}$$

with a constant independent of j and h.

Proof. To simplify the notation, we shall assume that Ω is a simple Lipschitz domain of the type $\Omega=\{(x',x_n)=(x_1,\dots,x_{n-1},x_n)\in\mathbb{R}^n\colon\ x_n>\psi(x'),\ |x'|<1\}$, where ψ is a Lipschitz function, and we identify Γ with $\{(x',x_n)\colon\ x_n=\psi(x'),\ |x'|<1\}$.

Step 1: First, let us observe that

$$\operatorname{dist}(x,\Gamma)\approx\bigl(x_n-\psi\bigl(x'\bigr)\bigr)\quad\text{for } x=\bigl(x',x_n\bigr)\in\Omega \tag{4.11}$$

and the constants in this equivalence depend only on the Lipschitz constant of ψ. The simple proof of this fact is based on the inner cone property of Lipschitz domains. We refer to [37, Chapter VI, Section 3.2, Lemma 2] for details.

Step 2: Let $j\in\mathbb{N}_0$ and $0<|h|\leqslant 1$ be fixed and let

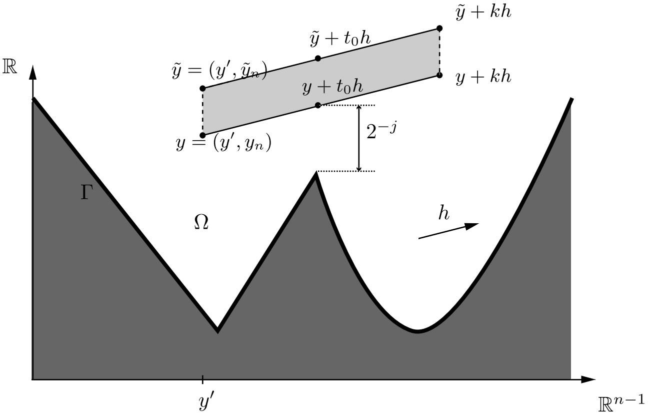
$$y=\bigl(y',y_n\bigr)\in\Omega^h_j$$

and let also

$$\tilde y=\bigl(y',\tilde y_n\bigr)\in\Omega^h_j$$

with $\tilde y_n>y_n$.

As $\tilde{y} \in \Omega_j^h$, there is a $t_0 \in [0,k]$ such that $\operatorname{dist}(\tilde{y}+t_0h,\Gamma) \leqslant 2^{-j+1}$.



Then we use $\psi(y'+t_0h) < t_0h_n + y_n$ (which follows from $y \in \Omega^h$ and $y+t_0h \in \Omega$) and (4.11) to get

$$\begin{aligned}\tilde{y}_n - y_n &= \left[\tilde{y}_n + t_0h_n - \psi\left(y'+t_0h'\right)\right] + \left[\psi\left(y'+t_0h'\right) - t_0h_n - y_n\right]\\ &\lesssim \operatorname{dist}(\tilde{y}+t_0h,\Gamma) \lesssim 2^{-j}.\end{aligned} \tag{4.12}$$

Step 3: Using (4.12), we observe that the set $\Omega(x') = \{x_n \in \mathbb{R}\colon (x',x_n) \in \Omega_j^h\}$ has for every $|x'| < 1$ length smaller than $c2^{-j}$. From this, the inequality (4.10) quickly follows. □

We shall use this geometrical observation together with the extension operator (4.7) to prove the following.

Lemma 4.8. *Let Ω be a bounded Lipschitz domain and let Γ be its boundary. Let a be a Lipschitz function on Γ. Let $0 < p \leqslant \infty$ and $0 < s < k$ for some $k \in \mathbb{N}$ with $k < 1/p + 1$. Then the extension operator defined by* (4.7) *satisfies*

$$\left\|\operatorname{Ext} a\middle|\mathbf{B}^s_{p,p}(\Omega)\right\| \lesssim \left\|a\middle|\operatorname{Lip}(\Gamma)\right\| \tag{4.13}$$

with the constant independent of $a \in \operatorname{Lip}(\Gamma)$.

Proof. Using the characterization by differences, we obtain

$$\begin{aligned}\left\|\operatorname{Ext} a\middle|\mathbf{B}^s_{p,p}(\Omega)\right\| &\lesssim \left\|\operatorname{Ext} a\middle|\mathbf{B}^{s'}_{p,\infty}(\Omega)\right\|\\ &\lesssim \left\|\operatorname{Ext} a|L_p(\Omega)\right\| + \sup_{0<|h|\leqslant 1} |h|^{-s'}\left\|\Delta_h^k \operatorname{Ext} a(\cdot,\Omega)|L_p(\Omega)\right\|,\end{aligned}$$

for $s' > 0$ with $s < s' < k$. Furthermore, we observe that one may modify the definition of $\Delta_h^r f(x,\Omega)$ given in (1.10) to be zero also if the whole segment $[x, x+kh]$ is not a subset of Ω. This follows by a detailed inspection of [40, Section 2.5.12] as well as [9] and [8], which are all based on the integration in cones.

Using the definition of μ_i, the first term may be estimated easily as

$$\big\|\mathrm{Ext}\,a|L_p(\Omega)\big\| \lesssim \big\|\mathrm{Ext}\,a|L_\infty(\Omega)\big\| \leqslant \big\|a|L_\infty(\Gamma)\big\|.$$

To estimate the second term, we shall need the following relationship between differences and derivatives. If $f \in C^k(\mathbb{R}^n)$ and $x, h \in \mathbb{R}^n$, we put $g(t) = f(x+th)$ for $t \in \mathbb{R}$ and obtain

$$\Delta_h^k f(x) = \Delta_1^k g(0) = \int_0^k g^{(k)}(t) B_k(t)\,\mathrm{d}t, \tag{4.14}$$

where B_k is the standard B spline of order k, i.e. the k-fold convolution of $\chi_{[0,1]}$ given by $B_k = \chi_{[0,1]} * \cdots * \chi_{[0,1]}$. Although (4.14) is a classical result of approximation theory (cf. [6, Section 4.7]), let us give a short proof using Fubini's theorem and induction over k:

$$\begin{aligned}
\Delta_1^{k+1} g(0) = \Delta_1^k g(1) - \Delta_1^k g(0) &= \int_0^k \big(g^{(k)}(t+1) - g^{(k)}(t)\big) B_k(t)\,\mathrm{d}t \\
&= \int_0^k B_k(t) \int_t^{t+1} g^{(k+1)}(u)\,\mathrm{d}u\,\mathrm{d}t = \int_0^{k+1} g^{(k+1)}(u) \int_{u-1}^{u} B_k(t)\,\mathrm{d}t\,\mathrm{d}u \\
&= \int_0^{k+1} g^{(k+1)}(u) B_{k+1}(u)\,\mathrm{d}u.
\end{aligned}$$

Hence if $[x, x+kh] \subset \Omega$ for some $x \in \Omega$, we obtain

$$\begin{aligned}
\big|\Delta_h^k \mathrm{Ext}\,a(x,\Omega)\big| &\lesssim |h|^k \int_0^k \max_{|\alpha|=k} \big|D^\alpha \mathrm{Ext}\,a(x+th)\big| \cdot B_k(t)\,\mathrm{d}t \\
&\lesssim |h|^k \cdot \big\|a|\operatorname{Lip}(\Gamma)\big\| \cdot \int_0^k \delta(x+th)^{1-k} \cdot B_k(t)\,\mathrm{d}t.
\end{aligned}$$

Let us fix $h \in \mathbb{R}^n$ with $0 < |h| \leqslant 1$ and let us denote $\Omega^h = \{x \in \Omega\colon [x, x+kh] \subset \Omega\}$ as in Lemma 4.7. We obtain

$$\begin{aligned}
&|h|^{-s'} \big\|\Delta_h^k \mathrm{Ext}\,a(\cdot,\Omega)|L_p(\Omega)\big\| \\
&\quad\lesssim |h|^{k-s'} \big\|a|\operatorname{Lip}(\Gamma)\big\| \left(\int_{\Omega^h} \left(\int_0^k \delta(x+th)^{1-k} \cdot B_k(t)\,\mathrm{d}t\right)^p \mathrm{d}x\right)^{1/p} \\
&\quad\lesssim \big\|a|\operatorname{Lip}(\Gamma)\big\| \left(\int_{\Omega^h} \max_{y \in [x, x+kh]} \delta(y)^{(1-k)p}\,\mathrm{d}x\right)^{1/p}
\end{aligned}$$

$$\lesssim \|a|\operatorname{Lip}(\Gamma)\| \left(\sum_{j=0}^{\infty} 2^{-j(1-k)p} |\Omega_j^h| \right)^{1/p}.$$

This, together with Lemma 4.7 and with $k < 1/p + 1$ finishes the proof. □

Lemma 4.9. *Let $0 < s' < 1$ be fixed. There is a non-linear extension operator* (*denoted by* **Ext**)*, which extends $\operatorname{Lip}^{\Gamma}$-atoms $a_{j,m}$ to $(s' + 1/p, p)$-atoms on $\mathbb{R}^n$.*

Proof. As the definition of $\operatorname{Lip}^{\Gamma}$-atoms as well as the definition of $(s' + 1/p, p)$-atoms works with $a_j(2^{-j}\cdot)$, by homogeneity arguments it is enough to prove

$$\left\| \mathbf{Ext}\, a_{0,m} \middle| \mathbf{B}_{p,p}^{s'+1/p}\left(\mathbb{R}^n\right) \right\| \lesssim \|a_{0,m}|\operatorname{Lip}(\Gamma)\| \tag{4.15}$$

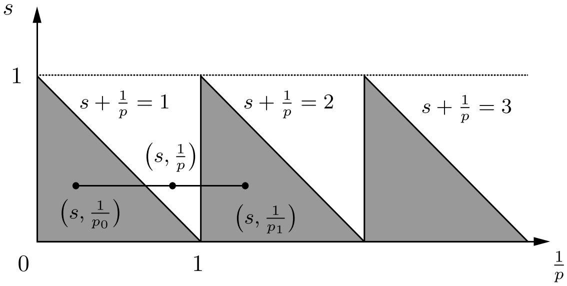
for $\operatorname{Lip}^{\Gamma}$-atoms $a_{j,m}$ with $j = 0$. First we show that

$$\left\| \operatorname{Ext} a_{0,m} \middle| \mathbf{B}_{p,p}^{s'+1/p}(\Omega) \right\| \lesssim \|a_{0,m}|\operatorname{Lip}(\Gamma)\| \tag{4.16}$$

for the extension operator constructed in (4.7). Let $0 < s' < 1$ and $0 < p \leqslant \infty$. We observe, that Lemma 4.8 implies (4.16) for all $0 < s' < 1$ for which there is a $k \in \mathbb{N}_0$ with

$$s' + 1/p < k < 1 + 1/p.$$

In the diagram below these points correspond to all $(s', \frac{1}{p})$ in the gray-shaded triangles.



Then Lemma 4.3 yields (4.16) for all $0 < s' < 1$ and $0 < p \leqslant \infty$ with $s_0 = s_1 = s'$ and $p_0 < p < p_1$ chosen in an appropriate way, see the attached diagram.

Finally, by Remark 1.5, we know that there is a function (denoted by $\mathbf{Ext}\, a_{0,m}$), such that

$$\left\| \mathbf{Ext}\, a_{0,m} \middle| \mathbf{B}_{p,p}^{s'+1/p}\left(\mathbb{R}^n\right) \right\| \lesssim \left\| \operatorname{Ext} a_{0,m} \middle| \mathbf{B}_{p,p}^{s'+1/p}(\Omega) \right\|.$$

This together with (4.16) finishes the proof of (4.15). □

We are now able to complete the proof of the missing part of the trace theorem.

Theorem 4.10. *Let $n \geqslant 2$ and Ω be a bounded Lipschitz domain with boundary Γ. Then for $0 < s < 1$ and $0 < p, q \leqslant \infty$ there is a bounded non-linear extension operator*

$$\mathrm{Ext} : \mathbf{B}^s_{p,q}(\Gamma) \longrightarrow \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega). \tag{4.17}$$

Proof. Let $f \in \mathbf{B}^s_{p,q}(\Gamma)$ with optimal decomposition in the sense of Theorem 3.8

$$f(x) = \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} a^{\Gamma}_{j,m}(x), \tag{4.18}$$

where $a^{\Gamma}_{j,m}$ are Lip^{Γ}-atoms, (4.18) converges in $L_p(\Gamma)$, and $\|f|\mathbf{B}^s_{p,q}(\Gamma)\| \sim \|\lambda|b^s_{p,q}(\Gamma)\|$.

We use the extension operator constructed in Lemma 4.9 and define by

$$\mathrm{Ext}\, f := \sum_{j=0}^{\infty} \sum_{m \in \mathbb{Z}^n} \lambda_{j,m} \big(\mathbf{Ext}\, a^{\Gamma}_{j,m}\big)\Big|_{\Omega} \tag{4.19}$$

an atomic decomposition of f in the space $\mathbf{B}^{s+1/p}_{p,q}(\Omega)$ with non-smooth $(s' + 1/p, p)$-atoms $\mathbf{Ext}\, a^{\Gamma}_{j,m}$, where $s < s' < 1$. The convergence of (4.19) in $L_p(\Omega)$ follows in the same way as in the proof of Step 3 of Theorem 3.8.

Together with $\|\lambda|b^s_{p,q}(\Gamma)\| \sim \|\lambda|b^{s+1/p}_{p,q}(\Omega)\|$, this shows that

$$\big\|\mathrm{Ext}\, f\big|\mathbf{B}^{s+1/p}_{p,q}(\Omega)\big\| \lesssim \big\|\lambda\big|b^{s+1/p}_{p,q}(\Omega)\big\| \sim \big\|\lambda\big|b^s_{p,q}(\Gamma)\big\| < \infty$$

is bounded. □

Theorems 4.2 and 4.10 together now allow us to state the general result for traces on Lipschitz domains without any restrictions on the parameters s, p and q.

Theorem 4.11. *Let $n \geqslant 2$ and Ω be a bounded Lipschitz domain with boundary Γ. Then for $0 < s < 1$ and $0 < p, q \leqslant \infty$,*

$$\mathrm{Tr}\, \mathbf{B}^{s+\frac{1}{p}}_{p,q}(\Omega) = \mathbf{B}^s_{p,q}(\Gamma). \tag{4.20}$$

The above theorem extends the trace results obtained in [33, Th. 3.4] from C^k domains with $k > s + \frac{1}{p}$ to Lipschitz domains.

Furthermore, the trace results for spaces of Triebel–Lizorkin type carry over as well to the case of Lipschitz domains. The proof follows [33, Th. 2.6] where the independence of the trace on q was established for F-spaces. Let us mention that the sequence spaces $f^s_{p,q}(\Omega)$ are defined similarly as $b^s_{p,q}(\Omega)$, cf. Definition 1.1, with ℓ_p and ℓ_q summation interchanged. The corresponding function spaces (denoted by $\mathfrak{F}^s_{p,q}(\Omega)$) are then defined as in Definition 1.4.

The main ingredient in the study of traces for Triebel–Lizorkin spaces $\mathfrak{F}^s_{p,q}(\Omega)$ is then the fact that the corresponding sequence spaces $f^s_{p,q}(\Gamma)$ are independent of q,

$$f^s_{p,q}(\Gamma) = b^s_{p,p}(\Gamma). \tag{4.21}$$

A proof may be found in [43, Prop. 9.22, p. 394] for Γ being a compact porous set in $\mathbb{R}^n$ with [12] as an important forerunner. In [45, Prop. 3.6] it is shown that the boundaries $\partial\Omega = \Gamma$ of (ε, δ)-domains Ω are porous. Therefore, this result is also true for boundaries of Lipschitz domains.

For completeness we state the trace results for F-spaces below.

Corollary 4.12. *Let* $0 < p < \infty$, $0 < q \leqslant \infty$, $0 < s < 1$, *and let* $\Omega \subset \mathbb{R}^n$ *be a bounded Lipschitz domain with boundary* Γ. *Then*

$$\operatorname{Tr} \mathfrak{F}^{s+\frac{1}{p}}_{p,q}(\Omega) = \mathbf{B}^s_{p,p}(\Gamma). \tag{4.22}$$

4.3. The limiting case

We briefly discuss what happens in the limiting case $s = 0$. In [34, Th. 2.7] traces for Besov and Triebel–Lizorkin spaces on d-sets Γ, $0 < d < n$, were studied. In particular, it was shown that for $0 < p < \infty$ and $0 < q \leqslant \infty$,

$$\operatorname{Tr} \mathbf{B}^{\frac{n-d}{p}}_{p,q}\left(\mathbb{R}^n\right) = L_p(\Gamma), \quad 0 < q \leqslant \min(1, p), \tag{4.23}$$

and

$$\operatorname{Tr} \mathfrak{F}^{\frac{n-d}{p}}_{p,q}\left(\mathbb{R}^n\right) = L_p(\Gamma), \quad 0 < p \leqslant 1. \tag{4.24}$$

Since the boundary Γ of a Lipschitz domain Ω is a d-set with $d = n - 1$ the results follow almost immediately from these previous results, using the fact that the B- and F-spaces on domains Ω are defined as restrictions of the corresponding spaces on $\mathbb{R}^n$, cf. Remark 1.5.

Corollary 4.13. *Let* Ω *be a bounded Lipschitz domain with boundary* Γ. *Furthermore, let* $0 < p < \infty$ *and* $0 < q \leqslant \infty$.

(i) *Then*

$$\operatorname{Tr} \mathbf{B}^{\frac{1}{p}}_{p,q}(\Omega) = L_p(\Gamma), \quad 0 < q \leqslant \min(1, p). \tag{4.25}$$

(ii) *Furthermore,*

$$\operatorname{Tr} \mathfrak{F}^{\frac{1}{p}}_{p,q}(\Omega) = L_p(\Gamma), \quad 0 < p \leqslant 1. \tag{4.26}$$

5. Pointwise multipliers in function spaces

As an application we now use our results on non-smooth atomic decompositions to deal with pointwise multipliers in the respective function spaces.

A function m in $L^{loc}_{\min(1,p)}(\mathbb{R}^n)$ is called a *pointwise multiplier* for $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ if

$$f \mapsto mf$$

generates a bounded map in $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$. The collection of all multipliers for $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ is denoted by $M(\mathbf{B}^s_{p,q}(\mathbb{R}^n))$. In the following, let ψ stand for a non-negative C^∞ function with

$$\operatorname{supp}\psi \subset \left\{y \in \mathbb{R}^n\colon |y| \leqslant \sqrt{n}\right\} \tag{5.1}$$

and

$$\sum_{l\in\mathbb{Z}^n} \psi(x-l) = 1, \quad x \in \mathbb{R}^n. \tag{5.2}$$

Definition 5.1. Let $s > 0$ and $0 < p, q \leqslant \infty$. We define the space $\mathbf{B}^s_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$ to be the set of all $f \in L^{loc}_{\min(1,p)}(\mathbb{R}^n)$ such that

$$\left\| f \middle| \mathbf{B}^s_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right)\right\| := \sup_{j\in\mathbb{N}_0,\, l\in\mathbb{Z}^n} \left\| \psi(\cdot - l) f\left(2^{-j}\cdot\right) \middle| \mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \tag{5.3}$$

is finite.

Remark 5.2. The study of pointwise multipliers is one of the key problems of the theory of function spaces. As far as classical Besov spaces and (fractional) Sobolev spaces with $p > 1$ are concerned we refer to [22–24]. Pointwise multipliers in general spaces $B^s_{p,q}(\mathbb{R}^n)$ and $F^s_{p,q}(\mathbb{R}^n)$ have been studied in great detail in [28, Ch. 4].

Self-similar spaces were first introduced in [42] and then considered in [43, Sect. 2.3]. Corresponding results for anisotropic function spaces may be found in [25]. We also mention their forerunners, the uniform spaces $\mathbf{B}^s_{p,q,\mathrm{unif}}(\mathbb{R}^n)$, studied in detail in [28, Sect. 4.9]. As stated in [20], for these spaces it is known that

$$M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right) = \mathbf{B}^s_{p,q,\mathrm{unif}}\left(\mathbb{R}^n\right), \quad 1 \leqslant p \leqslant q \leqslant \infty,\ s > \frac{n}{p},$$

cf. [36] concerning the proof. Self-similar spaces are also closely connected with pointwise multipliers. We shall use the abbreviation

$$\mathbf{B}^s_{p,\mathrm{selfs}}\left(\mathbb{R}^n\right) := \mathbf{B}^s_{p,p,\mathrm{selfs}}\left(\mathbb{R}^n\right).$$

One can easily show

$$\mathbf{B}^s_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right) \hookrightarrow L_\infty\left(\mathbb{R}^n\right). \tag{5.4}$$

To see this applying homogeneity gives

$$\begin{aligned}\left\|\psi(\cdot - l) f\left(2^{-j}\cdot\right) \middle| \mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| &\sim 2^{j\frac{n}{p}} \left\|\psi\left(2^j\cdot - l\right) f \middle| L_p\left(\mathbb{R}^n\right)\right\| \\ &\quad + 2^{-j(s-\frac{n}{p})}\left(\int_0^1 t^{-sq}\omega_r\left(\psi\left(2^j\cdot - l\right) f, t\right)_p{}^q \frac{\mathrm{d}t}{t}\right)^{1/q}\end{aligned}$$

uniformly for all $j \in \mathbb{N}_0$ and $l \in \mathbb{Z}^n$. Consequently,

$$2^{jn} \int_{\mathbb{R}^n} \left|\psi\left(2^j y - l\right)\right|^p \left|f(y)\right|^p \, \mathrm{d}y \leqslant c \left\| f \middle| \mathbf{B}^s_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right) \right\|^p. \tag{5.5}$$

Thus, the right-hand side of (5.5) is just a uniform bound for $|f(\cdot)|^p$ at its Lebesgue points, cf. [38, Cor., p. 13], which proves the desired embedding (5.4).

Definition 5.3. Let $s > 0$ and $0 < p, q \leqslant \infty$. We define

$$\mathbf{B}^{s+}_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right) := \bigcup_{\sigma > s} \mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right).$$

We have the following relation between pointwise multipliers and self-similar spaces.

Theorem 5.4. *Let $s > 0$ and $0 < p, q \leqslant \infty$. Then*

(i) $\mathbf{B}^{s+}_{p,q,\mathrm{selfs}}(\mathbb{R}^n) \subset M(\mathbf{B}^s_{p,q}(\mathbb{R}^n)) \hookrightarrow \mathbf{B}^s_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$.
(ii) *Additionally, if $0 < p \leqslant 1$,*

$$M\left(\mathbf{B}^s_p\left(\mathbb{R}^n\right)\right) = \mathbf{B}^s_{p,\mathrm{selfs}}\left(\mathbb{R}^n\right).$$

Proof. We first prove the right-hand side embedding in (i). Let $m \in M(\mathbf{B}^s_{p,q}(\mathbb{R}^n))$. An application of the homogeneity property from Theorem 1.8 yields

$$\begin{aligned}\left\|\psi(\cdot - l)m\left(2^{-j}\cdot\right)\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| &\sim 2^{-j(s-\frac{n}{p})}\left\|\psi\left(2^j\cdot - l\right)m\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\lesssim 2^{-j(s-\frac{n}{p})}\left\|m\middle|M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right)\right\| \cdot \left\|\psi\left(2^j\cdot - l\right)\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &= 2^{-j(s-\frac{n}{p})}\left\|m\middle|M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right)\right\| \cdot \left\|\psi\left(2^j\cdot\right)\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\sim \left\|m\middle|M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right)\right\|\left\|\psi\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \lesssim \left\|m\middle|M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right)\right\|\end{aligned}$$

for all $l \in \mathbb{Z}^n$, $j \in \mathbb{N}_0$, and hence,

$$\begin{aligned}\left\|m\middle|\mathbf{B}^s_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right)\right\| &= \sup_{j\in\mathbb{N}_0,\, l\in\mathbb{Z}^n}\left\|\psi(\cdot - l)m\left(2^{-j}\right)\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\lesssim \left\|m\middle|M\left(\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right)\right\|.\end{aligned}$$

We make use of the non-smooth atomic decompositions for $\mathbf{B}^s_{p,q}(\mathbb{R}^n)$ from Theorem 2.6 in order to prove the first inclusion in (i). Let $m \in \mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}$ with $\sigma > s$. Let $f \in \mathbf{B}^s_{p,q}(\mathbb{R}^n)$ with optimal smooth atomic decomposition

$$f = \sum_{j=0}^{\infty}\sum_{l\in\mathbb{Z}^n}\lambda_{j,l}a_{j,l} \quad \text{with } \left\|f\middle|\mathbf{B}^s_{p,q}\left(\mathbb{R}^n\right)\right\| \sim \left\|\lambda\middle|b^s_{p,q}\right\|, \tag{5.6}$$

where $a_{j,m}$ are K-atoms with $K > \sigma$. Then

$$mf = \sum_{j=0}^{\infty} \sum_{l \in \mathbb{Z}^n} \lambda_{j,l}(ma_{j,l}), \tag{5.7}$$

and we wish to prove that, up to normalizing constants, the $ma_{j,l}$ are (σ, p)-atoms. The support condition is obvious:

$$\operatorname{supp} ma_{j,l} \subset \operatorname{supp} a_{j,l} \subset dQ_{j,l}, \quad j \in \mathbb{N}_0, l \in \mathbb{Z}^n.$$

If $l = 0$ we put $a_j = a_{j,l}$. Note that

$$\operatorname{supp} a_j\left(2^{-j}\right) \subset \left\{y\colon\ |y_i| \leqslant \frac{d}{2}\right\}$$

and we can assume that

$$\psi(y) > 0 \quad \text{if } y \in \left\{x\colon\ |x_i| \leqslant d\right\}.$$

Then – using multiplier assertions from [33, Prop. 2.15(ii)] – we have for any $g \in \mathbf{B}^{\sigma}_{p,q}(\mathbb{R}^n)$,

$$\begin{aligned}\left\|a_j\left(2^{-j}\right)\psi^{-1}g\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\| &\lesssim \left\|a_j\left(2^{-j}\right)\psi^{-1}\middle|C^K\left(\mathbb{R}^n\right)\right\|\left\|g\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\lesssim \left\|g\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\|\end{aligned}$$

and hence

$$\left\|a_j\left(2^{-j}\right)\psi^{-1}\middle|M\left(\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right)\right\| \lesssim 1, \quad j \in \mathbb{N}_0. \tag{5.8}$$

By (5.8) and the homogeneity property we then get, for any $\sigma > \sigma' > s$ and $j \in \mathbb{N}_0$,

$$\begin{aligned}\left\|(ma_j)\left(2^{-j}\cdot\right)\middle|\mathbf{B}^{\sigma'}_{p}\left(\mathbb{R}^n\right)\right\| &\lesssim \left\|m\left(2^{-j}\cdot\right)a_j\left(2^{-j}\cdot\right)\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\lesssim \left\|a_j\left(2^{-j}\cdot\right)\psi^{-1}\middle|M\left(\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right)\right\|\left\|m\left(2^{-j}\cdot\right)\psi\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &\lesssim \left\|m\left(2^{-j}\cdot\right)\psi\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\|.\end{aligned} \tag{5.9}$$

In the case of $a_{j,l}$ with $l \in \mathbb{Z}^n$ one arrives at (5.9) with $a_{j,l}$ and $\psi(\cdot - l)$ in place of a_j and ψ, respectively. Hence

$$\begin{aligned}\left\|ma_{j,l}\left(2^{-j}\cdot\right)\middle|\mathbf{B}^{\sigma'}_{p}\left(\mathbb{R}^n\right)\right\| &\lesssim \sup_{j,l}\left\|m\left(2^{-j}\cdot\right)\psi(\cdot - l)\middle|\mathbf{B}^{\sigma}_{p,q}\left(\mathbb{R}^n\right)\right\| \\ &= \left\|m\middle|\mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right)\right\|, \quad j \in \mathbb{N}_0, l \in \mathbb{Z}^n,\end{aligned} \tag{5.10}$$

and therefore, $ma_{j,l}$ is a (σ', p)-atom where $\sigma' > s$. By Theorem 2.6, in view of (5.7), $mf \in \mathbf{B}^{s}_{p,q}(\mathbb{R}^n)$ and

$$\left\|mf\middle|\mathbf{B}^{s}_{p,q}\left(\mathbb{R}^n\right)\right\| \leqslant \left\|\lambda\middle|b^{s}_{p,q}\right\|\left\|m\middle|\mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right)\right\| \sim \left\|f\middle|\mathbf{B}^{s}_{p,q}\right\|\left\|m\middle|\mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\left(\mathbb{R}^n\right)\right\|,$$

which completes the proof of (i).

We now prove (ii). Restricting ourselves to $p = q$, let now $m \in \mathbf{B}^s_{p,\mathrm{selfs}}(\mathbb{R}^n)$. We can modify (5.9) by choosing $\sigma' = \sigma = s$,

$$\begin{aligned}\big\|(ma_j)\big(2^{-j}\cdot\big)\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\| &= \big\|m\big(2^{-j}\cdot\big)a_j\big(2^{-j}\cdot\big)\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\| \\ &\lesssim \big\|a_j\big(2^{-j}\cdot\big)\psi^{-1}\big|M\big(\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big)\big\|\big\|m\big(2^{-j}\cdot\big)\psi\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\| \\ &\lesssim \big\|m\big(2^{-j}\cdot\big)\psi\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\|, \end{aligned} \tag{5.11}$$

yielding for general atoms $a_{j,l}$,

$$\begin{aligned}\big\|ma_{j,l}\big(2^{-j}\cdot\big)\big|\mathbf{B}^s_{p,}\big(\mathbb{R}^n\big)\big\| &\lesssim \sup_{j,l}\big\|m\big(2^{-j}\cdot\big)\psi(\cdot - l)\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\| \\ &= \big\|m\big|\mathbf{B}^s_{p,\mathrm{selfs}}\big(\mathbb{R}^n\big)\big\|, \quad j \in \mathbb{N}_0, l \in \mathbb{Z}^n. \end{aligned} \tag{5.12}$$

Since $p \leqslant 1$, we have that $\mathbf{B}^s_p(\mathbb{R}^n)$ is a p-Banach space. From (5.6), using (5.7) and (5.12), we obtain

$$\begin{aligned}\big\|mf\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\|^p &\leqslant \sum_{j=0}^{\infty}\sum_{l\in\mathbb{Z}^n}|\lambda_{j,l}|^p 2^{j(s-\frac{n}{p})p}2^{-j(s-\frac{n}{p})p}\big\|ma_{j,l}\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\|^p \\ &\sim \big\|\lambda\big|b^s_{p,p}\big\|^p\big\|(ma_{j,l})\big(2^{-j}\cdot\big)\big|\mathbf{B}^s_p\big(\mathbb{R}^n\big)\big\|^p \\ &\lesssim \big\|\lambda\big|b^s_{p,p}\big\|^p\big\|m\big|\mathbf{B}^s_{p,\mathrm{selfs}}\big(\mathbb{R}^n\big)\big\|^p. \end{aligned} \tag{5.13}$$

Hence $m \in M(\mathbf{B}^s_p(\mathbb{R}^n))$ and, moreover, $\mathbf{B}^s_{p,\mathrm{selfs}}(\mathbb{R}^n) \hookrightarrow M(\mathbf{B}^s_p(\mathbb{R}^n))$. The other embedding follows from part (i). □

Remark 5.5. It remains open whether it is possible or not to generalize Theorem 5.4(ii) to the case when $p \neq q$. The problem in the proof given above is the estimate (5.13), which only holds if $p = q$.

Characteristic functions as multipliers The final part of this work is devoted to the question in which function spaces the characteristic function χ_Ω of a domain $\Omega \subset \mathbb{R}^n$ is a pointwise multiplier. We contribute to this question mainly as an application of Theorem 5.4. The results shed some light on a relationship between some fundamental notion of fractal geometry and pointwise multipliers in function spaces. For complementary remarks and studies in this direction we refer to [42].

There are further considerations of a similar kind in the literature, asking for geometric conditions on the domain Ω such that the corresponding characteristic function χ_Ω provides multiplier properties, cf. [14,15,12], and [28, Sect. 4.6.3].

Definition 5.6. Let Γ be a non-empty compact set in $\mathbb{R}^n$. Let h be a positive non-decreasing function on the interval $(0, 1]$. Then Γ is called an h-set, if there is a finite Radon measure $\mu \in \mathbb{R}^n$ with

$$\operatorname{supp}\mu = \Gamma \quad \text{and} \quad \mu\big(B(\gamma, r)\big) \sim h(r), \quad \gamma \in \Gamma,\ 0 < r \leqslant 1. \tag{5.14}$$

Remark 5.7. A measure μ with (5.14) satisfies the so-called *doubling condition*, meaning there is a constant $c > 0$ such that

$$\mu\big(B(\gamma, 2r)\big) \leqslant c\mu\big(B(\gamma, r)\big), \quad \gamma \in \Gamma,\ 0 < r < 1. \tag{5.15}$$

We refer to [42, p. 476] for further explanations.

Theorem 5.8. *Let Ω be a bounded domain in $\mathbb{R}^n$. Moreover, let $\sigma > 0$, $0 < p < \infty$, $0 < q \leqslant \infty$, and let $\Gamma = \partial\Omega$ be an h-set with*

$$\sup_{j\in\mathbb{N}_0} \sum_{k=0}^{\infty} 2^{k\sigma q} \left(\frac{h(2^{-j})}{h(2^{-j-k})} 2^{-kn} \right)^{q/p} < \infty \tag{5.16}$$

(*with the usual modifications if $q = \infty$*). *Let $\mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$ be the spaces defined in* (5.3). *Then*

$$\chi_\Omega \in \mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}\big(\mathbb{R}^n\big).$$

Proof. It simplifies the argument, and causes no loss of generality, to assume $\operatorname{diam}\Omega < 1$. We define

$$\Omega^k = \big\{x \in \Omega\colon\ 2^{-k-2} \leqslant \operatorname{dist}(x, \Gamma) \leqslant 2^{-k}\big\}, \quad k \in \mathbb{N}_0.$$

Moreover, let

$$\big\{\varphi_l^k\colon\ k \in \mathbb{N}_0,\ l = 1, \ldots, M_k\big\} \subset C_0^\infty(\Omega)$$

be a resolution of unity,

$$\sum_{k\in\mathbb{N}_0} \sum_{l=1}^{M_k} \varphi_l^k(x) = 1 \quad \text{if } x \in \Omega, \tag{5.17}$$

with

$$\operatorname{supp}\varphi_l^k \subset \big\{x\colon\ \big|x - x_l^k\big| \leqslant 2^{-k}\big\} \subset \Omega^k$$

and

$$\big|D^\alpha \varphi_l^k(x)\big| \lesssim 2^{|\alpha|k}, \quad |\alpha| \leqslant K,$$

where $K \in \mathbb{N}$ with $K > \sigma$. It is well known that resolutions of unity with the required properties exist. We now estimate the number M_k in (5.17). Combining the fact that the measure μ satisfies the doubling condition (5.15) together with (5.14) we arrive at

$$M_k h\big(2^{-k}\big) \lesssim 1, \quad k \in \mathbb{N}_0. \tag{5.18}$$

Since the φ_l^k in (5.17) are K-atoms according to Definition 1.3, we obtain

$$\big\|\chi_\Omega \big|\mathbf{B}^\sigma_{p,q}\big(\mathbb{R}^n\big)\big\|^q \leqslant \sum_{k=0}^\infty 2^{k(\sigma-n/p)q} M_k^{q/p} \lesssim \sum_{k=0}^\infty 2^{k\sigma q}\left(\frac{2^{-kn}}{h(2^{-k})}\right)^{q/p} < \infty. \tag{5.19}$$

This shows that $\chi_\Omega \in \mathbf{B}^\sigma_{p,q}(\mathbb{R}^n)$. We now prove that $\chi_\Omega \in \mathbf{B}^\sigma_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$. We consider the non-negative function $\psi \in C^\infty(\mathbb{R}^n)$ satisfying (5.1) and (5.2). By the definition of self-similar spaces, it suffices to consider

$$\chi_\Omega\big(2^{-j}\cdot\big)\psi,$$

assuming in addition that $0 \in 2^j\Gamma = \{2^j\gamma = (2^j\gamma_1,\dots,2^j\gamma_n)\colon \gamma\in\Gamma\}$, $j\in\mathbb{N}$. Let μ^j be the image measure of μ with respect to the dilations $y \mapsto 2^j y$. Then we obtain

$$\mu^j\big(B(0,\sqrt{n}\,)\cap 2^j\Gamma\big) \sim h\big(2^{-j}\big), \quad j\in\mathbb{N}_0.$$

We apply the same argument as above to $B(0,\sqrt{n}\,)\cap 2^j\Omega$ and $B(0,\sqrt{n}\,)\cap 2^j\Gamma$ in place of Ω and Γ, respectively. Let M_k^j be the counterpart of the above number M_k. Then

$$M_k^j h\big(2^{-j-k}\big) \lesssim h\big(2^{-j}\big), \quad j\in\mathbb{N}_0,\ k\in\mathbb{N}_0,$$

is the generalization of (5.18) we are looking for, which completes the proof. □

In view of Theorem 5.4 we have the following result.

Corollary 5.9. *Let Ω be a bounded domain in $\mathbb{R}^n$. Moreover, let $\sigma>0$, $0<p<\infty$, $0<q\leqslant\infty$, and let $\Gamma=\partial\Omega$ be an h-set satisfying (5.16). Then*

$$\chi_\Omega \in M\big(\mathbf{B}^s_{p,q}\big(\mathbb{R}^n\big)\big) \quad \textit{for } 1<p<\infty,\ 0<s<\sigma,$$

and

$$\chi_\Omega \in M\big(\mathbf{B}^\sigma_p\big(\mathbb{R}^n\big)\big) \quad \textit{for } 0<p\leqslant 1.$$

Remark 5.10. As for the assertion (5.16) we mention that

$$\sup_{j\in\mathbb{N}_0,\,k\in\mathbb{N}_0} 2^{k\sigma}\left(\frac{h(2^{-j})}{h(2^{-j-k})}2^{-kn}\right)^{1/p} < \infty$$

is the adequate counterpart for $\mathbf{B}^\sigma_{p,\infty}(\mathbb{R}^n)$. In the special case of d-sets, which corresponds to $h(t)\sim t^d$, the condition (5.16) therefore corresponds to

$$\sigma < \frac{n-d}{p} \quad \text{or} \quad \sigma = \frac{n-d}{p} \quad \text{and} \quad q=\infty.$$

For bounded Lipschitz domains Ω, i.e., $d = n - 1$, Theorem 5.8 therefore yields $\chi_\Omega \in \mathbf{B}^{\sigma}_{p,q,\mathrm{selfs}}(\mathbb{R}^n)$ if

$$\sigma < \frac{1}{p} \quad \text{or} \quad \sigma = \frac{1}{p} \quad \text{and} \quad q = \infty. \tag{5.20}$$

These results are sharp since there exists a Lipschitz domain Ω in $\mathbb{R}^n$ such that

$$\chi_\Omega \in \mathbf{B}^{\frac{1}{p}}_{p,\infty,\mathrm{selfs}}\big(\mathbb{R}^n\big) \quad \text{and} \quad \chi_\Omega \notin \mathbf{B}^{\frac{1}{p}}_{p,q}\big(\mathbb{R}^n\big) \quad \text{if } 0 < q < \infty.$$

In order to see this let $\Omega = [-\frac{1}{2}, \frac{1}{2}]^n$. Observing that

$$\omega_r(\chi_\Omega, t)_p \lesssim t^{\frac{1}{p}}$$

one calculates

$$\left(\int_0^1 t^{-\sigma q} \omega_r(\chi_\Omega, t)_p^q \frac{\mathrm{d}t}{t}\right)^{1/q} \lesssim \left(\int_0^1 t^{(\frac{1}{p}-\sigma)q} \frac{\mathrm{d}t}{t}\right)^{1/q}$$

which is finite if, and only if, σ satisfies (5.20). Therefore, in view of Theorem 5.4, concerning Lipschitz domains there is an

> *alternative s.t. either the trace of $\mathbf{B}^{\sigma}_{p,q}(\mathbb{R}^n)$ on Γ exists or χ_Ω is a pointwise multiplier for $\mathbf{B}^{\sigma}_{p,q}(\mathbb{R}^n)$,*

as was conjectured for F-spaces in [41, p. 36]: For smoothness $\sigma > \frac{1}{p}$ we have traces according to Theorem 4.11 whereas for $\sigma < \frac{1}{p}$ we know that χ_Ω is a pointwise multiplier for $\mathbf{B}^{\sigma}_{p,q}(\mathbb{R}^n)$. The limiting case $\sigma = \frac{1}{p}$ needs to be discussed separately: according to Corollary 4.13 we have traces for B-spaces with $q \leqslant \min(1, p)$, but χ_Ω is (possibly) only a multiplier for $\mathbf{B}^{1/p}_{p,\infty}(\mathbb{R}^n)$. There remains a 'gap' for spaces

$$\mathbf{B}^{1/p}_{p,q}\big(\mathbb{R}^n\big) \quad \text{when } \min(1, p) < q < \infty.$$

Acknowledgments

We thank Winfried Sickel and Hans Triebel for their valuable comments. Jan Vybíral acknowledges the financial support provided by the FWF project Y 432-N15 START-Preis "Sparse Approximation and Optimization in High Dimensions".

References

[1] G. Bourdaud, Sur les opérateurs pseudo-différentiels à coefficients peu réguliers, Habilitation thesis, Université de Paris-Sud, Paris, 1983.

[2] V.I. Burenkov, Sobolev Spaces on Domains, Teubner Texte zur Mathematik, Teubner, Stuttgart, 1998.

[3] A.M. Caetano, S. Lopes, H. Triebel, A homogeneity property for Besov spaces, J. Funct. Spaces Appl. 5 (2) (2007) 123–132.

[4] A.M. Caetano, S. Lopes, Homogeneity, non-smooth atoms and Besov spaces of generalised smoothness on quasi-metric spaces, Dissertationes Math. (Rozprawy Mat.) 460 (2009), 44 pp.

[5] B. Dacorogna, Introduction to the Calculus of Variations, Imperial College Press, London, 2004. Translated from the 1992 French original.

[6] R.A. DeVore, G.G. Lorentz, Constructive Approximation, Grundlehren Math. Wiss., vol. 303, Springer, Berlin, 1993.

[7] R.A. DeVore, V.A. Popov, Interpolation of Besov spaces, Trans. Amer. Math. Soc. 305 (1) (1988) 397–414.

[8] R.A. DeVore, R.C. Sharpley, Besov spaces on domains in $\mathbf{R}^d$, Trans. Amer. Math. Soc. 335 (2) (1993) 843–864.

[9] S. Dispa, Intrinsic characterizations of Besov spaces on Lipschitz domains, Math. Nachr. 260 (2003) 21–33.

[10] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, Differential Operators, Cambridge Univ. Press, Cambridge, 1996.

[11] K. Falconer, Fractal Geometry, Math. Found. Appl., John Wiley & Sons, Ltd., Chichester, 1990.

[12] M. Frazier, B. Jawerth, A discrete transform and decompositions of distribution spaces, J. Funct. Anal. 93 (1) (1990) 34–170.

[13] E. Gagliardo, Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili, Rend. Sem. Mat. Univ. Padova 27 (1957) 284–305 (in Italian).

[14] A.B. Gulisashvili, On multipliers in Besov spaces, Zap. Nautch. Sem. LOMI 135 (1984) 36–50 (in Russian).

[15] A.B. Gulisashvili, Multipliers in Besov spaces and traces of functions on subsets of the Euclidean n-space, DAN SSSR 281 (1985) 777–781 (in Russian).

[16] L.I. Hedberg, Y. Netrusov, An axiomatic approach to function spaces, spectral synthesis, and Luzin approximation, Mem. Amer. Math. Soc. 188 (882) (2007), 97 pp.

[17] D.D. Haroske, C. Schneider, Besov spaces with positive smoothness on $\mathbb{R}^n$, embeddings and growth envelopes, J. Approx. Theory 161 (2) (2009) 723–747.

[18] D. Jerison, C.E. Kenig, The inhomogeneous Dirichlet problem in Lipschitz domains, J. Funct. Anal. 130 (1) (1995) 161–219.

[19] A. Jonsson, H. Wallin, Function spaces on subsets of $\mathbf{R}^n$, Math. Rep. 2 (1984).

[20] H. Koch, W. Sickel, Pointwise multipliers of Besov spaces of smoothness zero and spaces of continuous functions, Rev. Mat. Iberoam. 18 (3) (2002) 587–626.

[21] S. Mayboroda, The Poisson problem on Lipschitz domains, PhD thesis, University of Missouri-Columbia, USA, 2005.

[22] V.G. Maz'ya, Sobolev Spaces, Springer Ser. Soviet Math., Springer-Verlag, Berlin, 1985.

[23] V.G. Maz'ya, T.O. Shaposhnikova, Theory of Multipliers in Spaces of Differentiable Functions, Monogr. Stud. in Math., vol. 23, Pitman (Advanced Publishing Program), Boston, MA, 1985.

[24] V.G. Maz'ya, T.O. Shaposhnikova, Theory of Sobolev Multipliers, Grundlehren Math. Wiss., vol. 337, Springer, Berlin, 2009.

[25] S.D. Moura, I. Piotrowska, M. Piotrowski, Non-smooth atomic decompositions of anisotropic function spaces and some applications, Studia Math. 180 (2) (2007) 169–190.

[26] P. Oswald, Multilevel Finite Element Approximation, Teubner Skripten zur Numerik, Teubner, Stuttgart, 1994.

[27] J. Peetre, New Thoughts on Besov Spaces, Duke Univ. Math. Ser., Duke University, Durham, NC, 1976.

[28] T. Runst, W. Sickel, Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations, de Gruyter Ser. Nonlinear Anal. Appl., vol. 3, Walter de Gruyter & Co., Berlin, 1996.

[29] V.S. Rychkov, Linear extension operators for restrictions of function spaces to irregular open sets, Studia Math. 140 (2) (2000) 141–162.

[30] B. Scharf, Atomic representations in function spaces and applications to pointwise multipliers and diffeomorphisms, a new approach, Math. Nachr., in press, http://dx.doi.org/10.1002/mana.201100336.

[31] C. Schneider, Spaces of Sobolev type with positive smoothness on $\mathbb{R}^n$, embeddings and growth envelopes, J. Funct. Spaces Appl. 7 (3) (2009) 251–288.

[32] C. Schneider, Trace operators in Besov and Triebel–Lizorkin spaces, Z. Anal. Anwend. 29 (3) (2010) 275–302.

[33] C. Schneider, Traces of Besov and Triebel–Lizorkin spaces on domains, Math. Nachr. 284 (5–6) (2011) 572–586.

[34] C. Schneider, Trace operators on fractals, entropy and approximation numbers, Georgian Math. J. 18 (3) (2011) 549–575.

[35] C. Schneider, J. Vybíral, Homogeneity property of Besov and Triebel–Lizorkin spaces, J. Funct. Spaces Appl. (2012), Article ID 281085.

[36] W. Sickel, I. Smirnow, Localization Properties of Besov Spaces and Its Associated Multiplier Spaces, Jena. Schr. Math. Inform., vol. 21/99, Universität Jena, 1999.

[37] E.M. Stein, Singular Integrals and Differentiability Properties of Functions, Princeton Math. Ser., vol. 30, Princeton University Press, Princeton, NJ, 1970.

[38] E.M. Stein, Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals, Princeton Math. Ser., vol. 43, Princeton University Press, Princeton, NJ, 1993. With the assistance of Timothy S. Murphy, Monographs in Harmonic Analysis, III.

[39] H. Triebel, Interpolation Theory, Function Spaces, Differential Operators, North-Holland Math. Library, vol. 18, North-Holland Publishing Co., Amsterdam, 1978.

[40] H. Triebel, Theory of Function Spaces, Monogr. Math., vol. 78, Birkhäuser Verlag, Basel, 1983.

[41] H. Triebel, Function spaces in Lipschitz domains and on Lipschitz manifolds. Characteristic functions as pointwise multipliers, Rev. Mat. Complut. 15 (2) (2002) 475–524.

[42] H. Triebel, Non-smooth atoms and pointwise multipliers in function spaces, Ann. Mat. Pura Appl. (4) 182 (4) (2003) 457–486.

[43] H. Triebel, Theory of Function Spaces III, Monogr. Math., vol. 100, Birkhäuser Verlag, Basel, 2006.

[44] H. Triebel, The dichotomy between traces on d-sets Γ in $\mathbb{R}^n$ and the density of $D(\mathbb{R}^n \setminus \Gamma)$ in function spaces, Acta Math. Sin. (Engl. Ser.) 24 (4) (2008) 539–554.

[45] H. Triebel, Function Spaces and Wavelets on Domains, EMS Tracts Math., vol. 7, EMS Publishing House, Zürich, 2008.

[46] H. Triebel, H. Winkelvoß, Intrinsic atomic characterizations of function spaces on domains, Math. Z. 221 (4) (1996) 647–673.

J Fourier Anal Appl
DOI 10.1007/s00041-012-9224-7

Spaces of Variable Smoothness and Integrability: Characterizations by Local Means and Ball Means of Differences

Henning Kempka · Jan Vybíral

Received: 28 March 2011 / Revised: 8 March 2012

Abstract We study the spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ of Besov and Triebel-Lizorkin type as introduced recently in Almeida and Hästö (J. Funct. Anal. 258(5):1628–2655, 2010) and Diening et al. (J. Funct. Anal. 256(6):1731–1768, 2009). Both scales cover many classical spaces with fixed exponents as well as function spaces of variable smoothness and function spaces of variable integrability.

The spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ have been introduced in Almeida and Hästö (J. Funct. Anal. 258(5):1628–2655, 2010) and Diening et al. (J. Funct. Anal. 256(6):1731–1768, 2009) by Fourier analytical tools, as the decomposition of unity. Surprisingly, our main result states that these spaces also allow a characterization in the time-domain with the help of classical ball means of differences.

To that end, we first prove a local means characterization for $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ with the help of the so-called Peetre maximal functions. Our results do also hold for 2-microlocal function spaces $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ which are a slight generalization of generalized smoothness spaces and spaces of variable smoothness.

Keywords Besov spaces · Triebel-Lizorkin spaces · Variable smoothness · Variable integrability · Ball means of differences · Peetre maximal operator · 2-microlocal spaces

Mathematics Subject Classification 46E35 · 46E30 · 42B25

Communicated by Arieh Iserles.

H. Kempka (✉)
Mathematical Institute, Friedrich-Schiller-University Jena, 07737 Jena, Germany
e-mail: henning.kempka@uni-jena.de

J. Vybíral
Department of Mathematics, TU Berlin, Sec. MA 8-1 Str. des 17. Juni 136, 10623 Berlin, Germany
e-mail: vybiral@math.tu-berlin.de

Published online: 12 April 2012

1 Introduction

Function spaces of variable integrability appeared in a work by Orlicz [41] already in 1931, but the recent interest in these spaces is based on the paper of Kováčik and Rákosnik [32] together with applications in terms of modelling electrorheological fluids [45]. A fundamental breakthrough concerning spaces of variable integrability was the observation that, under certain regularity assumptions on $p(\cdot)$, the Hardy-Littlewood maximal operator is also bounded on $L_{p(\cdot)}(\mathbb{R}^n)$, see [14]. This result has been generalized to wider classes of exponents $p(\cdot)$ in [11, 40] and [15].

Besides electrorheological fluids, the spaces $L_{p(\cdot)}(\mathbb{R}^n)$ possess interesting applications in the theory of PDE's, variational calculus, financial mathematics and image processing. A recent overview of this vastly growing field is given in [17].

Sobolev and Besov spaces with variable smoothness but fixed integrability have been introduced in the late 60s and early 70s in the works of Unterberger [57], Višik and Eskin [58], Unterberger and Bokobza [56] and in the work of Beauzamy [7]. Leopold studied in [33] Besov spaces where the smoothness is determined by a symbol $a(x,\xi)$ of a certain class of hypoelliptic pseudodifferential operators. In the special case $a(x,\xi)=(1+|\xi|^2)^{\sigma(x)/2}$ these spaces coincide with spaces of variable smoothness $B^{\sigma(x)}_{p,p}(\mathbb{R}^n)$.

A more general approach to spaces of variable smoothness are the so-called 2-microlocal function spaces $B^{\boldsymbol{w}}_{p,q}(\mathbb{R}^n)$ and $F^{\boldsymbol{w}}_{p,q}(\mathbb{R}^n)$. Here the smoothness in these spaces gets measured by a weight sequence $\boldsymbol{w}=(w_j)_{j=0}^{\infty}$. Besov spaces with such weight sequences appeared first in the works of Peetre [42] and Bony [9]. Establishing a wavelet characterization for 2-microlocal Hölder-Zygmund spaces in [24] it turned out that 2-microlocal spaces are well adapted in connection to regularity properties of functions [25, 35, 37]. Spaces of variable smoothness are a special case of 2-microlocal function spaces and in [34] and [8] characterizations by differences have been given for certain classes of them.

The theories of function spaces with fixed smoothness and variable integrability and function spaces with variable smoothness and fixed integrability finally crossed each other in [16], where the authors introduced the function spaces of Triebel-Lizorkin type with variable smoothness and simultaneously with variable integrability. It turned out that many of the spaces mentioned above are really included in this new structure, see [16] and references therein. The key point to merge both lines of investigation was the study of traces. From Theorem 3.13 in [16]

$$tr_{\mathbb{R}^{n-1}} F^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big) = F^{s(\cdot)-1/p(\cdot)}_{p(\cdot),p(\cdot)}\big(\mathbb{R}^{n-1}\big)$$

one immediately understands the necessity to take all exponents variable assuming $p(\cdot)$ or $s(\cdot)$ variable. So the trace embeddings may be described in a natural way in the context of these spaces. Furthermore, this was complemented in [59] by showing, that the classical Sobolev embedding theorem

$$F^{s_0(\cdot)}_{p_0(\cdot),q(\cdot)}\big(\mathbb{R}^n\big) \hookrightarrow F^{s_1(\cdot)}_{p_1(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)$$

holds also in this scale of function spaces if the usual condition is replaced by its point-wise analogue

$$s_0(x) - n/p_0(x) = s_1(x) - n/p_1(x), \quad x \in \mathbb{R}^n.$$

Finally, Almeida and Hästö managed in [1] to adapt the definition of Besov spaces to the setting of variable smoothness and integrability and proved the Sobolev and other usual embeddings in this scale.

The properties of Besov and Triebel-Lizorkin spaces of variable smoothness and integrability known so far give a reasonable hope that these new scales of function spaces enjoy sufficiently many properties to allow a local description of many effects, which up to now could only be described in a global way. Subsequently, for the spaces $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ there is a characterization by local means given in [30]. This characterization still works with Fourier analytical tools but the analyzing functions $k_0, k \in \mathcal{S}(\mathbb{R}^n)$ are compactly supported in the time-domain and we only need local values of f around $x \in \mathbb{R}^n$ to calculate the building blocks $k(2^{-j}, f)(x)$. This is in sharp contrast to the definition of the spaces by the decomposition of unity, cf. Definitions 1 and 3. For the spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ we will prove a local means assertion of this type in Sect. 3 which will be helpful later on.

The main aim of this paper is to present another essential property of the function spaces from [16] and [1]. We prove the surprising result that these spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ with variable smoothness and integrability do also allow a characterization purely in the time-domain by classical ball means of differences.

The paper is organized as follows. First of all we provide all necessary notation in Sect. 2. Since the proofs for spaces of variable smoothness and 2-microlocal function spaces work very similar (see Remark 2) we present our results for both scales. The proof for the local means characterization will be given in Sect. 3 in terms of 2-microlocal function spaces and we present the version for spaces of variable smoothness in Sect. 3.2. In Sect. 4 we prove the characterization by ball means of differences for $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and the version for 2-microlocal function spaces will be given in Sect. 4.5.

2 Notation

In this section we collect all the necessary definitions. We start with the variable Lebesgue spaces $L_{p(\cdot)}(\mathbb{R}^n)$. A measurable function $p : \mathbb{R}^n \to (0, \infty]$ is called a *variable exponent function* if it is bounded away from zero, i.e. if $p^- = \text{ess-inf}_{x\in\mathbb{R}^n} p(x) > 0$. We denote the set of all variable exponent functions by $\mathcal{P}(\mathbb{R}^n)$. We put also $p^+ = \text{ess-sup}_{x\in\mathbb{R}^n} p(x)$.

The variable exponent Lebesgue space $L_{p(\cdot)}(\mathbb{R}^n)$ consists of all measurable functions f for which there exist $\lambda > 0$ such that the modular

$$\varrho_{L_{p(\cdot)}(\mathbb{R}^n)}(f/\lambda) = \int_{\mathbb{R}^n} \varphi_{p(x)}\left(\frac{|f(x)|}{\lambda}\right) dx$$

is finite, where

$$\varphi_p(t)=\begin{cases} t^p & \text{if } p\in(0,\infty),\\ 0 & \text{if } p=\infty \text{ and } t\le 1,\\ \infty & \text{if } p=\infty \text{ and } t>1.\end{cases}$$

If we define $\mathbb{R}^n_\infty=\{x\in\mathbb{R}^n : p(x)=\infty\}$ and $\mathbb{R}^n_0=\mathbb{R}^n\setminus\mathbb{R}^n_\infty$, then the Luxemburg norm of a function $f\in L_{p(\cdot)}(\mathbb{R}^n)$ is given by

$$\begin{aligned}&\big\| f|L_{p(\cdot)}(\mathbb{R}^n)\big\|\\ &\quad=\inf\big\{\lambda>0:\varrho_{L_{p(\cdot)}(\mathbb{R}^n)}(f/\lambda)\le 1\big\}\\ &\quad=\inf\left\{\lambda>0:\int_{\mathbb{R}^n_0}\left(\frac{f(x)}{\lambda}\right)^{p(x)}dx<1 \text{ and } \big|f(x)\big|<\lambda \text{ for a.e. } x\in\mathbb{R}^n_\infty\right\}.\end{aligned}$$

If $p(\cdot)\ge 1$, then it is a norm otherwise it is always a quasi-norm.

To define the mixed spaces $\ell_{q(\cdot)}(L_{p(\cdot)})$ we have to define another modular. For $p,q\in\mathcal{P}(\mathbb{R}^n)$ and a sequence $(f_\nu)_{\nu\in\mathbb{N}_0}$ of $L_{p(\cdot)}(\mathbb{R}^n)$ functions we define

$$\varrho_{\ell_{q(\cdot)}(L_{p(\cdot)})}(f_\nu)=\sum_{\nu=0}^{\infty}\inf\left\{\lambda_\nu>0:\varrho_{L_{p(\cdot)}(\mathbb{R}^n)}\left(\frac{f_\nu}{\lambda_\nu^{1/q(\cdot)}}\right)\le 1\right\}. \tag{1}$$

If $q^+<\infty$, then we can replace (1) by the simpler expression

$$\varrho_{\ell_{q(\cdot)}(L_{p(\cdot)})}(f_\nu)=\sum_\nu\Big\| |f_\nu|^{q(\cdot)}|L_{\frac{p(\cdot)}{q(\cdot)}}\Big\|.$$

The (quasi-)norm in the $\ell_{q(\cdot)}(L_{p(\cdot)})$ spaces is defined as usual by

$$\big\| f_\nu|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|=\inf\big\{\mu>0:\varrho_{\ell_{q(\cdot)}(L_{p(\cdot)})}(f_\nu/\mu)\le 1\big\}.$$

It is known, cf. [1, 31], that $\ell_{q(\cdot)}(L_{p(\cdot)})$ is a norm if $q(\cdot)\ge 1$ is constant almost everywhere (a.e.) on $\mathbb{R}^n$ and $p(\cdot)\ge 1$, or if $1/p(x)+1/q(x)\le 1$ a.e. on $\mathbb{R}^n$, or if $1\le q(x)\le p(x)\le\infty$ a.e. on $\mathbb{R}^n$. Surprisingly enough, it turned out in [31] that the condition $\min(p(x),q(x))\ge 1$ a.e. on $\mathbb{R}^n$ is not sufficient for $\ell_{q(\cdot)}(L_{p(\cdot)})$ to be a norm. Nevertheless, it was proven in [1] that it is a quasi-norm for every $p,q\in\mathcal{P}(\mathbb{R}^n)$.

For the sake of completeness, we state also the definition of the space $L_{p(\cdot)}(\ell_{q(\cdot)})$, which is much more intuitive then the definition of $\ell_{q(\cdot)}(L_{p(\cdot)})$. One just takes the $\ell_{q(x)}$ norm of $(f_\nu(x))_{\nu\in\mathbb{N}_0}$ for every $x\in\mathbb{R}^n$ and then the $L_{p(\cdot)}$-norm with respect to $x\in\mathbb{R}^n$, i.e.

$$\big\| f_\nu|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|=\big\|\big\| f_\nu(x)|\ell_{q(x)}\big\||L_{p(\cdot)}\big\|.$$

It is easy to show [16] that $L_{p(\cdot)}(\ell_{q(\cdot)})$ is always a quasi-normed space and it is a normed space, if $\min(p(x),q(x))\ge 1$ holds point-wise.

The summation in the definition of the norms of $\ell_{q(\cdot)}(L_{p(\cdot)})$ and $L_{p(\cdot)}(\ell_{q(\cdot)})$ can also be taken for $\nu\in\mathbb{Z}$. It always comes out of the context over which interval the summation is taken. Occasionally, we may indicate it by $\|(f_\nu)_{\nu=-\infty}^{\infty}|\ell_{q(\cdot)}(L_{p(\cdot)})\|$.

By $\hat{f} = \mathcal{F}f$ and $f^\vee = \mathcal{F}^{-1}f$ we denote the usual Fourier transform and its inverse on $\mathcal{S}(\mathbb{R}^n)$, the Schwartz space of smooth and rapidly decreasing functions, and on $\mathcal{S}'(\mathbb{R}^n)$, the dual of the Schwartz space.

2.1 Spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$

The definition of Besov and Triebel-Lizorkin spaces of variable smoothness and integrability is based on the technique of *decomposition of unity* exactly in the same manner as in the case of constant exponents.

Definition 1 Let $\varphi_0 \in \mathcal{S}(\mathbb{R}^n)$ with $\varphi_0(x) = 1$ for $|x| \le 1$ and $\operatorname{supp}\varphi_0 \subseteq \{x \in \mathbb{R}^n : |x| \le 2\}$. For $j \ge 1$ we define

$$\varphi_j(x) = \varphi_0\big(2^{-j}x\big) - \varphi_0\big(2^{-j+1}x\big).$$

One may verify easily that

$$\sum_{j=0}^{\infty} \varphi_j(x) = 1 \quad \text{for all } x \in \mathbb{R}^n.$$

The following regularity classes for the exponents are necessary to make the definition of the spaces independent on the chosen decomposition of unity.

Definition 2 Let $g \in C(\mathbb{R}^n)$.

(i) We say that g is *locally* log-*Hölder continuous*, abbreviated $g \in C^{\log}_{loc}(\mathbb{R}^n)$, if there exists $c_{\log}(g) > 0$ such that

$$\big|g(x) - g(y)\big| \le \frac{c_{\log}(g)}{\log(e + 1/|x-y|)} \tag{2}$$

holds for all $x, y \in \mathbb{R}^n$.

(ii) We say that g is *globally* log-*Hölder continuous*, abbreviated $g \in C^{\log}(\mathbb{R}^n)$, if g is locally log-Hölder continuous and there exists $g_\infty \in \mathbb{R}$ such that

$$\big|g(x) - g_\infty\big| \le \frac{c_{\log}}{\log(e + |x|)}$$

holds for all $x \in \mathbb{R}^n$.

Remark 1 With (2) we obtain

$$\big|g(x)\big| \le c_{\log}(g) + \big|g(0)\big|, \quad \text{for all } x \in \mathbb{R}^n.$$

This implies that all functions $g \in C^{\log}_{loc}(\mathbb{R}^n)$ always belong to $L_\infty(\mathbb{R}^n)$.

If an exponent $p \in \mathcal{P}(\mathbb{R}^n)$ satisfies $1/p \in C^{\log}(\mathbb{R}^n)$, then we say it belongs to the class $\mathcal{P}^{\log}(\mathbb{R}^n)$. We recall the definition of the spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$, as given in [16] and [1].

Definition 3 (i) Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ with $0 < p^- \le p^+ < \infty$, $0 < q^- \le q^+ < \infty$ and let $s \in C_{loc}^{\log}(\mathbb{R}^n)$. Then

$$F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) = \{f \in \mathcal{S}'(\mathbb{R}^n) : \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|_\varphi < \infty\},$$
$$\text{where } \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|_\varphi = \|2^{js(\cdot)}(\varphi_j \hat{f})^\vee | L_{p(\cdot)}(\ell_{q(\cdot)})\|.$$

(ii) Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ and let $s \in C_{loc}^{\log}(\mathbb{R}^n)$. Then

$$B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) = \{f \in \mathcal{S}'(\mathbb{R}^n) : \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|_\varphi < \infty\},$$
$$\text{where } \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|_\varphi = \|2^{js(\cdot)}(\varphi_j \hat{f})^\vee | \ell_{q(\cdot)}(L_{p(\cdot)})\|.$$

The subscript φ at the norm symbolizes that the definition formally does depend on the resolution of unity. From [30] and [1] we have that the definition of the spaces $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ is independent of the chosen resolution of unity if $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ and $s \in C_{loc}^{\log}(\mathbb{R}^n)$. That means that different start functions φ_0 and $\tilde{\varphi}_0$ from Definition 1 induce equivalent norms in the above definition. So we will suppress the subscript φ in the notation of the norms.

Let us comment on the conditions on $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ for the Triebel-Lizorkin spaces. The condition $0 < p^- \le p^+ < \infty$ is quite natural since there exists also the restriction $p < \infty$ in the case of constant exponents, see [51] and [63]. The second one, $0 < q^- \le q^+ < \infty$, is a bit unnatural and comes from the use of the convolution Lemma 21 [16, Theorem 3.2]. There is some hope that this convolution lemma can be generalized and the case $q^+ = \infty$ can be incorporated in the definition of the F-spaces.

The Triebel-Lizorkin spaces with variable smoothness have first been introduced in [16] under much more restrictive conditions on $s(\cdot)$. These conditions have been relaxed in [30] in the context of 2-microlocal function spaces (see the next subsection).

Besov spaces with variable $p(\cdot)$, $q(\cdot)$ and $s(\cdot)$ have been introduced in [1].

Both scales contain as special cases a lot of well known function spaces. If s, p and q are constants, then we derive the well known Besov and Triebel-Lizorkin spaces with usual Hölder and Sobolev spaces included, see [51] and [52]. If the smoothness $s \in \mathbb{R}$ is a constant and $p \in \mathcal{P}^{\log}(\mathbb{R}^n)$ with $p^- > 1$, then $F^s_{p(\cdot),2}(\mathbb{R}^n) = \mathcal{L}^s_{p(\cdot)}(\mathbb{R}^n)$ are the variable Bessel potential spaces from [2] and [23] with its special cases $F^0_{p(\cdot),2}(\mathbb{R}^n) = L_{p(\cdot)}(\mathbb{R}^n)$ and $F^k_{p(\cdot),2}(\mathbb{R}^n) = W^k_{p(\cdot)}(\mathbb{R}^n)$ for $k \in \mathbb{N}_0$, see [16].

Taking $s \in \mathbb{R}$ and $q \in (0, \infty]$ as constants we derive the spaces $F^s_{p(\cdot),q}(\mathbb{R}^n)$ and $B^s_{p(\cdot),q}(\mathbb{R}^n)$ studied by Xu in [61] and [62].

Furthermore it holds $F^{s(\cdot)}_{p(\cdot),p(\cdot)}(\mathbb{R}^n) = B^{s(\cdot)}_{p(\cdot),p(\cdot)}(\mathbb{R}^n)$ and $B^{s(\cdot)}_{\infty,\infty}(\mathbb{R}^n)$ equals the variable Hölder-Zygmund space $\mathcal{C}^{s(\cdot)}(\mathbb{R}^n)$ introduced in [3, 4] and [44] with $0 < s^- \le s^+ \le 1$, see [1].

2.2 2-Microlocal Spaces

The definition of Besov and Triebel-Lizorkin spaces of variable smoothness and integrability is a special case of the so-called 2-microlocal spaces of variable integrability. As some of the results presented here get proved in this more general scale, we present also the definition of 2-microlocal spaces. It is based on the dyadic decomposition of unity as presented above combined with the concept of admissible weight sequences.

Definition 4 Let $\alpha \geq 0$ and let $\alpha_1, \alpha_2 \in \mathbb{R}$ with $\alpha_1 \leq \alpha_2$. A sequence of non-negative measurable functions $\boldsymbol{w} = (w_j)_{j=0}^{\infty}$ belongs to the class $\mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$ if and only if

(i) there exists a constant $C > 0$ such that

$$0 < w_j(x) \leq C w_j(y)\big(1 + 2^j|x-y|\big)^{\alpha} \quad \text{for all } j \in \mathbb{N}_0 \text{ and all } x, y \in \mathbb{R}^n$$

(ii) and for all $j \in \mathbb{N}_0$ and all $x \in \mathbb{R}^n$ we have

$$2^{\alpha_1} w_j(x) \leq w_{j+1}(x) \leq 2^{\alpha_2} w_j(x).$$

Such a system $(w_j)_{j=0}^{\infty} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$ is called an *admissible weight sequence*.

Finally, here is the definition of the spaces under consideration.

Definition 5 Let $\boldsymbol{w} = (w_j)_{j\in\mathbb{N}_0} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$. Further, let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ (with $p^+, q^+ < \infty$ in the F-case), then we define

$$B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big) = \big\{f \in \mathcal{S}'\big(\mathbb{R}^n\big) : \big\|f|B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|_{\varphi} < \infty\big\},$$
$$\text{where } \big\|f|B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|_{\varphi} = \big\|w_j(\varphi_j \hat{f})^{\vee}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|$$

and

$$F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big) = \big\{f \in \mathcal{S}'\big(\mathbb{R}^n\big) : \big\|f|F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|_{\varphi} < \infty\big\},$$
$$\text{where } \big\|f|F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|_{\varphi} = \big\|w_j(\varphi_j \hat{f})^{\vee}|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|.$$

The independence of the decomposition of unity for the 2-microlocal spaces from Definition 5 follows from the local means characterization (see [30] for the Triebel-Lizorkin and Sect. 3 for the Besov spaces).

The 2-microlocal spaces with the special weight sequence

$$w_j(x) = 2^{js}\big(1 + 2^j|x - x_0|\big)^{s'} \quad \text{with } s, s' \in \mathbb{R} \text{ and } x_0 \in \mathbb{R}^n \tag{3}$$

have first been introduced by Peetre in [42] and by Bony in [9]. Later on, Jaffard and Meyer gave a characterization in [24] and [25] with wavelets of the spaces $C^{s,s'}_{x_0} = B^{\boldsymbol{w}}_{\infty,\infty}(\mathbb{R}^n)$ and $H^{s,s'}_{x_0} = B^{\boldsymbol{w}}_{2,2}(\mathbb{R}^n)$ with the weight sequence (3). It turned out that spaces of this type are very useful to study regularity properties of functions.

Subsequently, Lévy-Véhel and Seuret developed in [35] the 2-microlocal formalism and studied the behavior of cusps, chirps and fractal functions with respect to the spaces $C^{s,s'}_{x_0}$.

A first step to a more general weight sequence $\boldsymbol{w}$ has been taken by Moritoh and Yamada in [38] and wider ranges of function spaces have been studied by Xu in [60] and by Andersson in [5].

The above definition for 2-microlocal weight sequences was presented by Besov in [8] and also in [30] by Kempka.

A different line of study for spaces of variable smoothness—using different methods—are the spaces of generalized smoothness introduced by Goldman and Kalyabin in [20, 21, 26] and [27]. A systematic treatment of these spaces based on differences has been given by Goldman in [22], see also the survey [29] and references therein.

Later on, spaces of generalized smoothness appeared in interpolation theory and have been investigated in [10, 36] and [39]. For further information on these spaces see the survey paper [19] where also a characterization by atoms and local means for these spaces is given.

From the definition of admissible sequences, $d_1\sigma_j \le \sigma_{j+1} \le d_2\sigma_j$, it follows directly that the spaces of generalized smoothness $B^{(\sigma_j)}_{p,q}(\mathbb{R}^n)$ and $F^{(\sigma_j)}_{p,q}(\mathbb{R}^n)$ of Farkas and Leopold [19] and $B^{(s,\Psi)}_{p,q}(\mathbb{R}^n)$ and $F^{(s,\Psi)}_{p,q}(\mathbb{R}^n)$ from Moura [39] are a special subclass of 2-microlocal function spaces with $2^{\alpha_1} = d_1$, $2^{\alpha_2} = d_2$ and $\alpha = 0$.

In a different approach Schneider in [48] studied spaces of varying smoothness. Here the smoothness at a point gets determined by a global smoothness $s_0 \in \mathbb{R}$ and a local smoothness function $s(\cdot)$. These spaces can not be incorporated into the scale of 2-microlocal function spaces, but there exist some embeddings.

Remark 2 Surprisingly, these 2-microlocal weight sequences are directly connected to variable smoothness functions $s : \mathbb{R}^n \to \mathbb{R}$ if we set

$$w_j(x) = 2^{js(x)}. \tag{4}$$

If $s \in C^{\log}_{loc}(\mathbb{R}^n)$ (which is the standard condition on $s(\cdot)$), then $\boldsymbol{w} = (w_j(x))_{j\in\mathbb{N}_0} = (2^{js(x)})_{j\in\mathbb{N}_0}$ belongs to $\mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$ with $\alpha_1 = s^-$ and $\alpha_2 = s^+$. For the third index α we use Lemma 19 with $m = 0$ and obtain $\alpha = c_{\log}(s)$, where $c_{\log}(s)$ is the constant for $s(\cdot)$ from (2). That means that spaces of variable smoothness from Definition 3 are a special case of 2-microlocal function spaces from Definition 5. Both types of function spaces are very closely connected and the properties used in the proofs are either

$$2^{k|s(x)-s(y)|} \le c \quad \text{or} \quad \frac{w_k(x)}{w_k(y)} \le c \tag{5}$$

for $|x - y| \le c2^{-k}$. This property follows directly either from the definition of $s \in C^{\log}_{loc}(\mathbb{R}^n)$ or from Definition 4.

Nevertheless there exist examples of admissible weight sequences which can not be expressed in terms of variable smoothness functions. For example the important and well studied case of the weight sequence $\boldsymbol{w}$ from (3) can not be expressed via (4) if $s' \neq 0$. Another example are the spaces of generalized smoothness which can not be identified as spaces of variable smoothness.

Since spaces of variable smoothness are included in the scale of 2-microlocal function spaces all special cases of the previous subsection can be identified in the definition of 2-microlocal spaces.

Although the 2-microlocal spaces include the scales of spaces of variable smoothness, we will give some of our proofs in the notation of variable smoothness, since this notation is more common. We will then reformulate the results in terms of 2-microlocal spaces, the proof works then very similar; we just have to use (5).

3 Local Means Characterization

The main result of this section is the local means characterization of the spaces $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$. For the spaces $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ there already exists a local means characterization [30, Corollary 4.7]. We shall first give the full proof for the 2-microlocal spaces and later on (in Sect. 3.2) we restate the result also for spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

The crucial tool will be the Peetre maximal operator, as defined by Peetre in [42]. The operator assigns to each system $(\Psi_k)_{k\in\mathbb{N}_0} \subset \mathcal{S}(\mathbb{R}^n)$, to each distribution $f \in \mathcal{S}'(\mathbb{R}^n)$ and to each number $a > 0$ the following quantities

$$\left(\Psi_k^* f\right)_a(x) := \sup_{y\in\mathbb{R}^n} \frac{|(\Psi_k * f)(y)|}{1 + |2^k(y-x)|^a}, \quad x \in \mathbb{R}^n \text{ and } k \in \mathbb{N}_0. \tag{6}$$

We start with two given functions $\psi_0, \psi_1 \in \mathcal{S}(\mathbb{R}^n)$. We define

$$\psi_j(x) = \psi_1\left(2^{-j+1}x\right), \quad \text{for } x \in \mathbb{R}^n \text{ and } j \in \mathbb{N}.$$

Furthermore, for all $j \in \mathbb{N}_0$ we write $\Psi_j = \hat{\psi}_j$. The main theorem of this section reads as follows.

Theorem 6 *Let $\boldsymbol{w} = (w_k)_{k\in\mathbb{N}_0} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$, $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ and let $a > 0$, $R \in \mathbb{N}_0$ with $R > \alpha_2$. Further, let ψ_0, ψ_1 belong to $\mathcal{S}(\mathbb{R}^n)$ with*

$$D^{\beta}\psi_1(0) = 0, \quad \textit{for } 0 \le |\beta| < R, \tag{7}$$

and

$$\left|\psi_0(x)\right| > 0 \quad \textit{on } \left\{x \in \mathbb{R}^n : |x| < \varepsilon\right\}, \tag{8}$$

$$\left|\psi_1(x)\right| > 0 \quad \textit{on } \left\{x \in \mathbb{R}^n : \varepsilon/2 < |x| < 2\varepsilon\right\} \tag{9}$$

for some $\varepsilon > 0$. For $a > \frac{n + c_{\log}(1/q)}{p^-} + \alpha$ and all $f \in \mathcal{S}'(\mathbb{R}^n)$ we have

$$\left\| f | B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\left(\mathbb{R}^n\right)\right\| \approx \left\| (\Psi_k * f) w_k | \ell_{q(\cdot)}(L_{p(\cdot)})\right\| \approx \left\| \left(\Psi_k^* f\right)_a w_k | \ell_{q(\cdot)}(L_{p(\cdot)})\right\|.$$

Remark 3

(i) The proof relies on [46] and will be shifted to the next section. Moreover, Theorem 6 shows that the definition of the 2-microlocal spaces of variable integrability is independent of the resolution of unity used in the Definition 5.

(ii) The conditions (7) are usually called *moment conditions* while (8) and (9) are the so called *Tauberian conditions*.

(iii) If $R = 0$, then there are no moment conditions (7) on ψ_1.

(iv) The notation $c_{\log}(1/q)$ stands for the constant from (2) with $1/q(\cdot)$.

Next we reformulate the abstract Theorem 6 in the sense of classical local means (see Sects. 2.4.6 and 2.5.3 in [52]). Since the proof is the same as the one from Theorem 2.4 in [30] we just state the result.

Corollary 1 *There exist functions* $k_0, k \in \mathcal{S}(\mathbb{R}^n)$ *with* $\operatorname{supp} k_0, \operatorname{supp} k \subset \{x \in \mathbb{R}^n : |x| < 1\}$ *and* $D^\beta \hat{k}(0) = 0$ *for all* $0 \le |\beta| < \alpha_2$ *such that for all* $f \in \mathcal{S}'(\mathbb{R}^n)$

$$\left\| k_0(1, f) w_0 | L_{p(\cdot)}\big(\mathbb{R}^n\big) \right\| + \left\| k\big(2^{-j}, f\big) w_j | \ell_{q(\cdot)}(L_{p(\cdot)}) \right\|$$

is an equivalent norm on $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

The building blocks get calculated by

$$k(t, f)(x) = \int_{\mathbb{R}^n} k(y) f(x + ty) dy = t^{-n} \int_{\mathbb{R}^n} k\left(\frac{y - x}{t}\right) f(y) dy$$

and similarly for $k_0(1, f)(x)$.

A similar characterization for $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and details how these functions $k_0, k \in \mathcal{S}(\mathbb{R}^n)$ can be constructed can be found in [30].

3.1 Proof of Local Means

The proof of Theorem 6 is divided into three parts. The next section is devoted to some technical lemmas needed later. Section 3.1.2 is devoted to the proof of Theorem 12, which gives an inequality between different Peetre maximal operators. Finally, Sect. 3.1.3 proves the boundedness of the Peetre maximal operator in Theorem 13. These two theorems combined give immediately the proof of Theorem 6.

3.1.1 Helpful Lemmas

Before proving the local means characterization we recall some technical lemmas, which appeared in the paper of Rychkov [46]. For some of them we need adapted versions to our situation.

The first lemma describes the use of the so called moment conditions.

Lemma 7 ([46], Lemma 1) *Let* $g, h \in \mathcal{S}(\mathbb{R}^n)$ *and let* $M \in \mathbb{N}_0$. *Suppose that*

$$\big(D^\beta \hat{g}\big)(0) = 0 \quad \textit{for } 0 \le |\beta| < M.$$

Then for each $N \in \mathbb{N}_0$ there is a constant C_N such that

$$\sup_{z\in\mathbb{R}^n} \big|(g_t * h)(z)\big|\big(1+|z|^N\big) \le C_N t^M, \quad \textit{for } 0<t<1,$$

where $g_t(x)=t^{-n}g(x/t)$.

The next lemma is a discrete convolution inequality. We formulate it in a rather abstract notation and point out later on the conclusions we need.

Lemma 8 *Let $X \subset \{(f_k)_{k\in\mathbb{Z}} : f_k : \mathbb{R}^n \to [-\infty,\infty]$ measurable$\}$ be a quasi-Banach space of sequences of measurable functions. Further we assume that its quasi-norm is shift-invariant, i.e. it satisfies*

$$\big\|(f_{k+l})_{k\in\mathbb{Z}}|X\big\| = \big\|(f_k)_{k\in\mathbb{Z}}|X\big\| \quad \textit{for every } l\in\mathbb{Z} \textit{ and } (f_k)_{k\in\mathbb{Z}}\in X.$$

For a sequence of non-negative functions $(g_k)_{k\in\mathbb{Z}} \in X$ and $\delta>0$ we denote

$$G_\nu(x) = \sum_{k=-\infty}^{\infty} 2^{-|\nu-k|\delta} g_k(x), \quad x\in\mathbb{R}^n,\ \nu\in\mathbb{Z}.$$

Then there exists a constant $c>0$ depending only on δ and X such that for every sequence $(g_k)_{k\in\mathbb{Z}}$

$$\big\|(G_\nu)_\nu|X\big\| \le c\big\|(g_k)_k|X\big\|.$$

Proof Since X is a quasi-Banach space, there exists a $r>0$ such that $\|\cdot|X\|$ is equivalent to some r-norm, cf. [6, 43]. We have then the following

$$\begin{aligned}\big\|(G_\nu)_\nu|X\big\|^r &= \left\|\left(\sum_{k=-\infty}^{\infty} 2^{-|\nu-k|\delta} g_k\right)_\nu \middle| X\right\|^r = \left\|\left(\sum_{l\in\mathbb{Z}} 2^{-|l|\delta} g_{\nu+l}\right)_\nu \middle| X\right\|^r \\ &\lesssim \sum_{l\in\mathbb{Z}} 2^{-|l|r\delta}\big\|(g_{\nu+l})_\nu|X\big\|^r \le c\big\|(g_\nu)_\nu|X\big\|^r.\end{aligned}$$

Now taking the power $1/r$ yields the desired estimate. □

The spaces $L_{p(\cdot)}(\ell_{q(\cdot)})$ and $\ell_{q(\cdot)}(L_{p(\cdot)})$ are quasi-Banach spaces which fulfill the conditions of Lemma 8. Therefore, we obtain the following

Lemma 9 *Let $p,q\in\mathcal{P}(\mathbb{R}^n)$ and $\delta>0$. Let $(g_k)_{k\in\mathbb{Z}}$ be a sequence of non-negative measurable functions on $\mathbb{R}^n$ and denote*

$$G_\nu(x) = \sum_{k\in\mathbb{Z}} 2^{-|\nu-k|\delta} g_k(x), \quad x\in\mathbb{R}^n,\ \nu\in\mathbb{Z}.$$

Then there exist a constants $C_1, C_2 > 0$, depending on $p(\cdot), q(\cdot)$ and δ, such that

$$\big\|G_\nu|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| \le C_1\big\|g_k|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| \quad \textit{and}$$
$$\big\|G_\nu|L_{p(\cdot)}(\ell_{q(\cdot)})\big\| \le C_2\big\|g_k|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|.$$

Remark 4 Of course, Lemma 9 holds true also if the indices k and ν run only over natural numbers.

Since the maximal operator is in general not bounded on $\ell_{q(\cdot)}(L_{p(\cdot)})$ (see [1, Example 4.1]) we need a replacement for that. It turned out that a convolution with radial decreasing functions fits very well into the scheme. A careful evaluation of the proof in [1, Lemma 4.7] together with Lemma 19 gives us the following convolution inequality.

Lemma 10 *Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ with $p(\cdot) \ge 1$ and let $\eta_{\nu,m}(x) = 2^{n\nu}(1+2^\nu|x|)^{-m}$. For all $m > n + c_{\log}(1/q)$ there exists a constant $c > 0$ such that for all sequences $(f_j)_{j\in\mathbb{N}_0} \in \ell_{q(\cdot)}(L_{p(\cdot)})$ it holds*

$$\big\|(\eta_{\nu,m} * f_\nu)_{\nu\in\mathbb{N}_0}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| \le c\big\|(f_j)_{j\in\mathbb{N}_0}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|.$$

The last technical lemma is overtaken literally from [46].

Lemma 11 ([46], Lemma 3) *Let $0 < r \le 1$ and let $(\gamma_\nu)_{\nu\in\mathbb{N}_0}$, $(\beta_\nu)_{\nu\in\mathbb{N}_0}$ be two sequences taking values in $(0,\infty)$. Assume that for some $N^0 \in \mathbb{N}_0$,*

$$\limsup_{\nu\to\infty} \frac{\gamma_\nu}{2^{\nu N^0}} < \infty. \tag{10}$$

Furthermore, we assume that for any $N \in \mathbb{N}$

$$\gamma_\nu \le C_N \sum_{k=0}^{\infty} 2^{-kN} \beta_{k+\nu}\gamma_{k+\nu}^{1-r}, \quad \nu \in \mathbb{N}_0,\ C_N < \infty$$

holds, then for any $N \in \mathbb{N}$

$$\gamma_\nu^r \le C_N \sum_{k=0}^{\infty} 2^{-kNr}\beta_{k+\nu}, \quad \nu \in \mathbb{N}_0$$

holds with the same constants C_N.

3.1.2 Comparison of Different Peetre Maximal Operators

In this subsection we present an inequality between different Peetre maximal operators. Let us recall the notation given before Theorem 6. For two given functions $\psi_0, \psi_1 \in \mathcal{S}(\mathbb{R}^n)$ we define

$$\psi_j(x) = \psi_1\big(2^{-j+1}x\big), \quad \text{for } x \in \mathbb{R}^n \text{ and } j \in \mathbb{N}.$$

Furthermore, for all $j \in \mathbb{N}_0$ we write $\Psi_j = \hat{\psi}_j$ and in an analogous manner we define Φ_j from two starting functions $\phi_0, \phi_1 \in \mathcal{S}(\mathbb{R}^n)$. Using this notation we are ready to formulate the theorem.

Theorem 12 *Let $\boldsymbol{w} = (w_j)_{j \in \mathbb{N}_0} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$, $p, q \in \mathcal{P}(\mathbb{R}^n)$ and $a > 0$. Moreover, let $R \in \mathbb{N}_0$ with $R > \alpha_2$,*

$$D^\beta \psi_1(0) = 0, \quad 0 \le |\beta| < R \tag{11}$$

and for some $\varepsilon > 0$

$$\bigl|\phi_0(x)\bigr| > 0 \quad \text{on } \bigl\{x \in \mathbb{R}^n : |x| < \varepsilon\bigr\}, \tag{12}$$

$$\bigl|\phi_1(x)\bigr| > 0 \quad \text{on } \bigl\{x \in \mathbb{R}^n : \varepsilon/2 < |x| < 2\varepsilon\bigr\}, \tag{13}$$

then

$$\bigl\| \bigl(\Psi_k^* f\bigr)_a w_k | \ell_{q(\cdot)}(L_{p(\cdot)}) \bigr\| \le c \bigl\| \bigl(\Phi_k^* f\bigr)_a w_k | \ell_{q(\cdot)}(L_{p(\cdot)}) \bigr\|$$

holds for every $f \in \mathcal{S}'(\mathbb{R}^n)$.

Remark 5 Observe that there are no restrictions on $a > 0$ and $p, q \in \mathcal{P}(\mathbb{R}^n)$ in the theorem above.

Proof We have the fixed resolution of unity from Definition 1 and define the sequence of functions $(\lambda_j)_{j \in \mathbb{N}_0}$ by

$$\lambda_j(x) = \frac{\varphi_j(\frac{2x}{\varepsilon})}{\phi_j(x)}.$$

It follows from the Tauberian conditions (12) and (13) that they satisfy

$$\sum_{j=0}^{\infty} \lambda_j(x)\phi_j(x) = 1, \quad x \in \mathbb{R}^n, \tag{14}$$

$$\lambda_j(x) = \lambda_1\bigl(2^{-j+1}x\bigr), \quad x \in \mathbb{R}^n,\ j \in \mathbb{N}, \tag{15}$$

and

$$\operatorname{supp} \lambda_0 \subset \bigl\{x \in \mathbb{R}^n : |x| \le \varepsilon\bigr\} \quad \text{and} \quad \operatorname{supp} \lambda_1 \subset \bigl\{x \in \mathbb{R}^n : \varepsilon/2 \le |x| \le 2\varepsilon\bigr\}. \tag{16}$$

Furthermore, we denote $\Lambda_k = \hat{\lambda}_k$ for $k \in \mathbb{N}_0$ and obtain together with (14) the following identities (convergence in $\mathcal{S}'(\mathbb{R}^n)$)

$$f = \sum_{k=0}^{\infty} \Lambda_k * \Phi_k * f, \qquad \Psi_\nu * f = \sum_{k=0}^{\infty} \Psi_\nu * \Lambda_k * \Phi_k * f. \tag{17}$$

We have

$$\begin{aligned}\big|(\Psi_\nu * \Lambda_k * \Phi_k * f)(y)\big| &\le \int_{\mathbb{R}^n} \big|(\Psi_\nu * \Lambda_k)(z)\big|\big|(\Phi_k * f)(y-z)\big| dz \\ &\le \big(\Phi_k^* f\big)_a(y) \int_{\mathbb{R}^n} \big|(\Psi_\nu * \Lambda_k)(z)\big|\big(1+\big|2^k z\big|^a\big) dz \\ &=: \big(\Phi_k^* f\big)_a(y) I_{\nu,k}, \end{aligned} \tag{18}$$

where

$$I_{\nu,k} := \int_{\mathbb{R}^n} \big|(\Psi_\nu * \Lambda_k)(z)\big|\big(1+\big|2^k z\big|^a\big) dz.$$

According to Lemma 7 we get

$$I_{\nu,k} \le c \begin{cases} 2^{(k-\nu)R}, & k \le \nu, \\ 2^{(\nu-k)(a+1+|\alpha_1|)}, & \nu \le k. \end{cases} \tag{19}$$

Namely, we have for $1 \le k < \nu$ with the change of variables $2^k z \mapsto z$

$$\begin{aligned} I_{\nu,k} &= 2^{-n} \int_{\mathbb{R}^n} \big|\big(\Psi_{\nu-k} * \Lambda_1(\cdot/2)\big)(z)\big|\big(1+|z|^a\big) dz \\ &\le c \sup_{z\in\mathbb{R}^n} \big|\big(\Psi_{\nu-k} * \Lambda_1(\cdot/2)\big)(z)\big|\big(1+|z|\big)^{a+n+1} \le c 2^{(k-\nu)R}. \end{aligned}$$

Similarly, we get for $1 \le \nu < k$ with the substitution $2^\nu z \mapsto z$

$$\begin{aligned} I_{\nu,k} &= 2^{-n} \int_{\mathbb{R}^n} \big|\big(\Psi_1(\cdot/2) * \Lambda_{k-\nu}\big)(z)\big|\big(1+\big|2^{k-\nu} z\big|^a\big) dz \\ &\le c 2^{(\nu-k)(M-a)}. \end{aligned}$$

M can be taken arbitrarily large because Λ_1 has infinitely many vanishing moments. Taking $M = 2a + |\alpha_1| + 1$ we derive (19) for the cases $k, \nu \ge 1$ with $k \ne \nu$. The missing cases can be treated separately in an analogous manner. The needed moment conditions are always satisfied by (11) and (16). The case $k = \nu = 0$ is covered by the constant c in (19).

Furthermore, we have

$$\begin{aligned} \big(\Phi_k^* f\big)_a(y) &\le \big(\Phi_k^* f\big)_a(x)\big(1+\big|2^k(x-y)\big|^a\big) \\ &\le \big(\Phi_k^* f\big)_a(x)\big(1+\big|2^\nu(x-y)\big|^a\big) \max\big(1, 2^{(k-\nu)a}\big). \end{aligned}$$

We put this into (18) and get

$$\sup_{y\in\mathbb{R}^n} \frac{|(\Psi_\nu * \Lambda_k * \Phi_k * f)(y)|}{1+|2^\nu(x-y)|^a} \le c\big(\Phi_k^* f\big)_a(x) \begin{cases} 2^{(k-\nu)R}, & k \le \nu, \\ 2^{(\nu-k)(1+|\alpha_1|)}, & k \ge \nu. \end{cases}$$

Multiplying both sides with $w_\nu(x)$ and using

$$w_\nu(x) \le w_k(x) \begin{cases} 2^{(k-\nu)(-\alpha_2)}, & k \le \nu, \\ 2^{(\nu-k)\alpha_1}, & k \ge \nu, \end{cases}$$

leads us to

$$\sup_{y\in\mathbb{R}^n} \frac{|(\Psi_\nu * \Lambda_k * \Phi_k * f)(y)|}{1+|2^\nu(x-y)|^a} w_\nu(x) \le c\big(\Phi_k^* f\big)_a(x) w_k(x) \begin{cases} 2^{(k-\nu)(R-\alpha_2)}, & k \le \nu, \\ 2^{(\nu-k)}, & k \ge \nu. \end{cases}$$

This inequality together with (17) gives for $\delta := \min(1, R-\alpha_2) > 0$

$$\big(\Psi_\nu^* f\big)_a(x) w_\nu(x) \le c \sum_{k=0}^{\infty} 2^{-|k-\nu|\delta} \big(\Phi_k^* f\big)_a(x) w_k(x), \quad x \in \mathbb{R}^n.$$

Taking the $\ell_{q(\cdot)}(L_{p(\cdot)})$ norm and using Lemma 9 yields immediately the desired result. □

3.1.3 Boundedness of the Peetre Maximal Operator

We will present a theorem which describes the boundedness of the Peetre maximal operator. We use the same notation introduced at the beginning of the last subsection. Especially, we have the functions $\psi_k \in \mathcal{S}(\mathbb{R}^n)$ and $\Psi_k = \hat{\psi}_k \in \mathcal{S}(\mathbb{R}^n)$ for all $k \in \mathbb{N}_0$.

Theorem 13 *Let $(w_k)_{k\in\mathbb{N}_0} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$, $a > 0$ and $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$. For some $\varepsilon > 0$ we assume $\psi_0, \psi_1 \in \mathcal{S}(\mathbb{R}^n)$ with*

$$|\psi_0| > 0 \quad \text{on } \big\{x \in \mathbb{R}^n : |x| < \varepsilon\big\},$$
$$|\psi_1| > 0 \quad \text{on } \big\{x \in \mathbb{R}^n : \varepsilon/2 < |x| < 2\varepsilon\big\}.$$

For $a > \frac{n+c_{\log}(1/q)}{p^-} + \alpha$

$$\big\|\big(\Psi_k^* f\big)_a w_k | \ell_{q(\cdot)}(L_{p(\cdot)})\big\| \le c \big\|(\Psi_k * f) w_k | \ell_{q(\cdot)}(L_{p(\cdot)})\big\|$$

holds for all $f \in \mathcal{S}'(\mathbb{R}^n)$.

Remark 6 Observe that in the theorem above no moment conditions on ψ_1 are stated but this time there are restrictions on a and $p(\cdot), q(\cdot)$.

Proof As in the last proof we find the functions $(\lambda_j)_{j\in\mathbb{N}_0}$ with the properties (15), (16) and

$$\sum_{k=0}^{\infty} \lambda_k\big(2^{-\nu}x\big)\psi_k\big(2^{-\nu}x\big) = 1 \quad \text{for all } \nu \in \mathbb{N}_0.$$

Instead of (17) we get the identity

$$\Psi_\nu * f = \sum_{k=0}^{\infty} \Lambda_{k,\nu} * \Psi_{k,\nu} * \Psi_\nu * f, \tag{20}$$

where

$$\Lambda_{k,\nu}(\xi) = \big[\lambda_k\big(2^{-\nu}\cdot\big)\big]^\wedge(\xi) = 2^{\nu n}\Lambda_k\big(2^\nu\xi\big) \quad \text{for all } \nu, k \in \mathbb{N}_0.$$

The $\Psi_{k,\nu}$ are defined similarly. For $k \ge 1$ and $\nu \in \mathbb{N}_0$ we have $\Psi_{k,\nu} = \Psi_{k+\nu}$ and with the notation

$$\sigma_{k,\nu}(x) = \begin{cases} \psi_0(2^{-\nu}x) & \text{if } k = 0, \\ \psi_\nu(x) & \text{otherwise} \end{cases}$$

we get $\psi_k(2^{-\nu}x)\psi_\nu(x) = \sigma_{k,\nu}(x)\psi_{k+\nu}(x)$. Hence, we can rewrite (20) as

$$\Psi_\nu * f = \sum_{k=0}^{\infty} \Lambda_{k,\nu} * \hat{\sigma}_{k,\nu} * \Psi_{k+\nu} * f. \tag{21}$$

For $k \ge 1$ we get from Lemma 7

$$\big|(\Lambda_{k,\nu} * \hat{\sigma}_{k,\nu})(z)\big| = 2^{(\nu-1)n}\big|\big(\Lambda_k * \Psi_1(\cdot/2)\big)\big(2^\nu z\big)\big| \le C_M 2^{\nu n}\frac{2^{-kM}}{(1+|2^\nu z|^a)} \tag{22}$$

for all $k, \nu \in \mathbb{N}_0$ and arbitrary large $M \in \mathbb{N}$. For $k = 0$ we get the estimate (22) by using Lemma 7 with $M = 0$. This together with (21) gives us

$$\big|(\Psi_\nu * f)(y)\big| \le C_M 2^{\nu n}\sum_{k=0}^{\infty}\int_{\mathbb{R}^n}\frac{2^{-kM}}{(1+|2^\nu(y-z)|^a)}\big|(\Psi_{k+\nu} * f)(z)\big|dz. \tag{23}$$

For fixed $r \in (0,1]$ we divide both sides of (23) by $(1+|2^\nu(x-y)|^a)$ and we take the supremum with respect to $y \in \mathbb{R}^n$. Using the inequalities

$$\big(1+\big|2^\nu(y-z)\big|^a\big)\big(1+\big|2^\nu(x-y)\big|^a\big) \ge c\big(1+\big|2^\nu(x-z)\big|^a\big),$$

$$\big|(\Psi_{k+\nu} * f)(z)\big| \le \big|(\Psi_{k+\nu} * f)(z)\big|^r\big(\Psi_{k+\nu}^* f\big)_a(x)^{1-r}\big(1+\big|2^{k+\nu}(x-z)\big|^a\big)^{1-r}$$

and

$$\frac{(1+|2^{k+\nu}(x-z)|^a)^{1-r}}{(1+|2^\nu(x-z)|^a)} \le \frac{2^{ka}}{(1+|2^{k+\nu}(x-z)|^a)^r},$$

we get

$$\big(\Psi_\nu^* f\big)_a(x) \le C_M\sum_{k=0}^{\infty}2^{-k(M+n-a)}\big(\Psi_{k+\nu}^* f\big)_a(x)^{1-r}\int_{\mathbb{R}^n}\frac{2^{(k+\nu)n}|(\Psi_{k+\nu} * f)(z)|^r}{(1+|2^{k+\nu}(x-z)|^a)^r}dz. \tag{24}$$

Now, we apply Lemma 11 with

$$\gamma_\nu = \left(\Psi_\nu^* f\right)_a(x), \qquad \beta_\nu = \int_{\mathbb{R}^n} \frac{2^{\nu n}|(\Psi_\nu * f)(z)|^r}{(1+|2^\nu(x-z)|^a)^r} dz, \quad \nu \in \mathbb{N}_0$$

$N = M + n - a$, $C_N = C_M + n - a$ and N^0 in (10) equals the order of the distribution $f \in \mathcal{S}'(\mathbb{R}^n)$.

By Lemma 11 we obtain for every $N \in \mathbb{N}$, $x \in \mathbb{R}^n$ and $\nu \in \mathbb{N}_0$

$$\left(\Psi_\nu^* f\right)_a(x)^r \le C_N \sum_{k=0}^{\infty} 2^{-kNr} \int_{\mathbb{R}^n} \frac{2^{(k+\nu)n}|(\Psi_{k+\nu} * f)(z)|^r}{(1+|2^{k+\nu}(x-z)|^a)^r} dz \tag{25}$$

provided that $(\Psi_\nu^* f)_a(x) < \infty$.

Since $f \in \mathcal{S}'(\mathbb{R}^n)$, we see that $(\Psi_\nu^* f)_a(x) < \infty$ for all $x \in \mathbb{R}^n$ and all $\nu \in \mathbb{N}_0$ at least if $a > N^0$, where N^0 is the order of the distribution. Thus we have (25) with C_N independent of $f \in \mathcal{S}'(\mathbb{R}^n)$ for $a \ge N^0$ and therefore with $C_N = C_{N,f}$ for all $a > 0$ (the right side of (25) decreases as a increases). One can easily check that (25) with $C_N = C_{N,f}$ implies that if for some $a > 0$ the right side of (25) is finite, then $(\Psi_\nu^* f)_a(x) < \infty$. Now, repeating the above argument resurrects the independence of C_N. If the right side of (25) is infinite, there is nothing to prove. More exhaustive arguments of this type have been used in [54] and [47].

We point out that (25) holds also for $r > 1$, where the proof is much simpler. We only have to take (23) with $a + n$ instead of a, divide both sides by $(1 + |2^\nu(x-y)|^a)$ and apply Hölder's inequality with respect to k and then z.

Multiplying (25) by $w_\nu(x)^r$ we derive with the properties of our weight sequence

$$\left(\Psi_\nu^* f\right)_a(x)^r w_\nu(x)^r \le C_N' \sum_{k=0}^{\infty} 2^{-k(N+\alpha_1)r} \int_{\mathbb{R}^n} \frac{2^{(k+\nu)n}|(\Psi_{k+\nu} * f)(z)|^r w_{k+\nu}(z)^r}{(1+|2^{k+\nu}(x-z)|^{a-\alpha})^r} dz, \tag{26}$$

for all $x \in \mathbb{R}^n$, $\nu \in \mathbb{N}_0$ and all $N \in \mathbb{N}$.

Now, we choose $r = p^-$ and we have $r(a-\alpha) > n + c_{\log}(1/q)$. We denote $g_{k+\nu}^r(z) = |(\Psi_{k+\nu} * f)(z)|^r w_{k+\nu}(z)^r$ then we can rewrite (26) by

$$\left(\Psi_\nu^* f\right)_a(x)^r w_\nu(x)^r \le C_N' \sum_{l=\nu}^{\infty} 2^{-(l-\nu)(N+\alpha_1)r} \left(g_l^r * \eta_{l,r(a-\alpha)}\right)(x). \tag{27}$$

For fixed $N > 0$ with $\delta = N + \alpha_1 > 0$ we apply the $\ell_{\frac{q(\cdot)}{r}}(L_{\frac{p(\cdot)}{r}})$ norm and derive from (27)

$$\left\| \left(\Psi_k^* f\right)_a^r w_k^r | \ell_{q(\cdot)/r}(L_{p(\cdot)/r}) \right\| \le C_N \left\| \sum_{l=\nu}^{\infty} 2^{-(l-\nu)\delta} \left(g_l^r * \eta_{l,r(a-\alpha)}\right) \middle| \ell_{q(\cdot)/r}(L_{p(\cdot)/r}) \right\|.$$

Now application of Lemma 9 and Lemma 10 ($r(a-\alpha) > n + c_{\log}(1/q)$) on the formula above give us

$$\left\| \left(\Psi_k^* f\right)_a^r(\cdot) w_k^r(\cdot) | \ell_{q(\cdot)/r}(L_{p(\cdot)/r}) \right\| \le C_N' \left\| \left|(\Psi_\nu * f)(\cdot)\right|^r w_\nu(\cdot)^r | \ell_{q(\cdot)/r}(L_{p(\cdot)/r}) \right\|$$

which proves the theorem. □

3.2 Local Means Characterization of $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$

In this section we reformulate the local means characterization for $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ from above and for $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ from Corollary 4.7 in [30] in terms of variable smoothness. If we have a variable smoothness function $s \in C^{\log}_{loc}(\mathbb{R}^n)$ given, then $w_j(x) = 2^{js(x)}$ defines an admissible weight sequence $\boldsymbol{w} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$ with $\alpha_1 = s^-$, $\alpha_2 = s^+$ and $\alpha = c_{\log}(s)$, cf. Remark 2. Here, we denote by $c_{\log}(s)$ the constant in (2) for $s(\cdot)$.

Theorem 14 *Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ ($p^+, q^+ < \infty$ in the F-case) and $s \in C^{\log}_{loc}(\mathbb{R}^n)$. Further let $a > 0$, $R \in \mathbb{N}_0$ with $R > s^+$ and let ψ_0, ψ_1 belong to $\mathcal{S}(\mathbb{R}^n)$ with*

$$D^\beta \psi_1(0) = 0, \quad \text{for } 0 \le |\beta| < R,$$

and

$$\left|\psi_0(x)\right| > 0 \quad \text{on } \left\{x \in \mathbb{R}^n : |x| < \varepsilon\right\},$$
$$\left|\psi_1(x)\right| > 0 \quad \text{on } \left\{x \in \mathbb{R}^n : \varepsilon/2 < |x| < 2\varepsilon\right\}$$

for some $\varepsilon > 0$.

1. *For $a > \frac{n + c_{\log}(1/q)}{p^-} + c_{\log}(s)$ and all $f \in \mathcal{S}'(\mathbb{R}^n)$ we have*

$$\left\| f | B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \right\| \approx \left\| 2^{ks(\cdot)}(\Psi_k * f) | \ell_{q(\cdot)}(L_{p(\cdot)}) \right\| \approx \left\| 2^{ks(\cdot)} \left(\Psi_k^* f\right)_a | \ell_{q(\cdot)}(L_{p(\cdot)}) \right\|.$$

2. *For $a > \frac{n}{\min(p^-, q^-)} + c_{\log}(s)$ and all $f \in \mathcal{S}'(\mathbb{R}^n)$ we have*

$$\left\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \right\| \approx \left\| 2^{ks(\cdot)}(\Psi_k * f) | L_{p(\cdot)}(\ell_{q(\cdot)}) \right\| \approx \left\| 2^{ks(\cdot)} \left(\Psi_k^* f\right)_a | L_{p(\cdot)}(\ell_{q(\cdot)}) \right\|.$$

Remark 7 During the referee process of this work, there appeared in [18] a characterization by local means and a characterization by atoms for $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$. The author moved the smoothness sequence $2^{ks(\cdot)}$ into the Peetre maximal operator (6) and modified it to

$$\left(\Psi_k^* 2^{ks(\cdot)} f\right)_a(x) = \sup_{y \in \mathbb{R}^n} \frac{2^{ks(y)} |(\Psi_k * f)(y)|}{1 + |2^k(y - x)|^a}.$$

For this modified Peetre maximal operator he obtained in [18, Theorem 2] an equivalence of the norms similar to our Theorem 14 for $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$. The advantage of his method is that the condition on $a > 0$ weakens to $a > \frac{n}{p^-}$.

4 Ball Means of Differences

This section is devoted to the characterization of Besov and Triebel-Lizorkin spaces $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ by ball means of differences. In the case of constant indices p,q and s, this is a classical part of the theory of function spaces. We refer especially to [51, Sect. 2.5] and references given there. It turns out that, under the restriction

$$s>\sigma_p=n\left(\frac{1}{\min(p,1)}-1\right) \tag{28}$$

in the B-case and

$$s>\sigma_{p,q}=n\left(\frac{1}{\min(p,q,1)}-1\right) \tag{29}$$

in the F-case, Besov and Triebel-Lizorkin spaces with constant indices may be characterized by expressions involving only the differences of the function values without any use of Fourier analysis. This was complemented in [49] and [50] by showing that these conditions are also indispensable. Of course, we are limited by (28) and (29) also in the case of variable exponents.

The characterization by (local means of) differences for 2-microlocal spaces with constant $p,q>1$ was given by Besov [8] and a similar characterization for Besov spaces with $p=q=\infty$ and the special weight sequence from (3) was given by Seuret and Levy Véhél in [34]. We refer to [19] and [28] for the treatment of spaces of generalized smoothness.

Our approach follows essentially [51] with some modifications described in [53]. The main obstacle on this way is the unboundedness of the maximal operator in the frame of $L_{p(\cdot)}(\ell_{q(\cdot)})$ and $\ell_{q(\cdot)}(L_{p(\cdot)})$ spaces, cf. [16, Sect. 5] and [1, Example 4.1]. This is circumvented by the use of convolution with radial functions in the sense of [16] and [1] together with a certain bootstrapping argument, which shall be described in detail below.

The plan of this part of the work is as follows. First we give in Sect. 4.1 the necessary notation. We state the main assertions of this part in Sect. 4.2. Then we prove in Sect. 4.3 a certain preliminary version of these assertions. In Sect. 4.4 we prove a characterization by ball means of differences for spaces with $q\in(0,\infty]$ constant (where the maximal operator is bounded) and use this together with our preliminary characterization from Sect. 4.3 to conclude the proof. Finally, in Sect. 4.5 we will present the ball means of differences characterization also for the 2-microlocal function spaces $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ and in Sect. 4.6 we present separately some useful Lemmas, not to disturb the main proofs of this part.

4.1 Notation

Let f be a function on $\mathbb{R}^n$ and let $h\in\mathbb{R}^n$. Then we define

$$\Delta^1_h f(x)=f(x+h)-f(x),\quad x\in\mathbb{R}^n.$$

The higher order differences are defined inductively by

$$\Delta_h^M f(x) = \Delta_h^1\big(\Delta_h^{M-1} f\big)(x), \quad M = 2, 3, \dots$$

This definition also allows a direct formula

$$\Delta_h^M f(x) := \sum_{j=0}^{M} (-1)^j \binom{M}{j} f\big(x + (M-j)h\big). \tag{30}$$

By *ball means of differences* we mean the quantity

$$d_t^M f(x) = t^{-n} \int_{|h| \le t} \big|\Delta_h^M f(x)\big| dh = \int_B \big|\Delta_{th}^M f(x)\big| dh,$$

where $B = \{y \in \mathbb{R}^n : |y| < 1\}$ is the unit ball of $\mathbb{R}^n$, $t > 0$ is a real number and M is a natural number.

Let us now introduce the (quasi-)norms, which shall be the main subject of our study. We define

$$\begin{aligned}\big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^* &:= \big\| f | L_{p(\cdot)}\big(\mathbb{R}^n\big)\big\| \\ &\quad + \left\| \left(\int_0^\infty t^{-s(x)q(x)} \big(d_t^M f(x)\big)^{q(x)} \frac{dt}{t} \right)^{1/q(x)} \middle| L_{p(\cdot)}\big(\mathbb{R}^n\big) \right\|\end{aligned} \tag{31}$$

and its partially discretized counterpart

$$\begin{aligned}\big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^{**} &:= \big\| f | L_{p(\cdot)}\big(\mathbb{R}^n\big)\big\| \\ &\quad + \left\| \left(\sum_{k=-\infty}^{\infty} 2^{ks(x)q(x)} \big(d_{2^{-k}}^M f(x)\big)^{q(x)} \right)^{1/q(x)} \middle| L_{p(\cdot)}\big(\mathbb{R}^n\big) \right\| \\ &= \big\| f | L_{p(\cdot)}\big(\mathbb{R}^n\big)\big\| + \big\| \big(2^{ks(x)} d_{2^{-k}}^M f(x)\big)_{k=-\infty}^{\infty} \big| L_{p(\cdot)}(\ell_{q(\cdot)})\big\|.\end{aligned}$$

The norm $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ admits a direct counterpart also for Besov spaces, namely

$$\big\| f | B^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^{**} := \big\| f | L_{p(\cdot)}\big(\mathbb{R}^n\big)\big\| + \big\| \big(2^{ks(x)} d_{2^{-k}}^M f(x)\big)_{k=-\infty}^{\infty} \big| \ell_{q(\cdot)}(L_{p(\cdot)})\big\|. \tag{32}$$

Finally, we shall use as a technical tool also the analogues of (31)–(32) with the integration over t restricted to $0 < t < 1$. This leads to the following expressions

$$\begin{aligned}\big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|_1^* &:= \big\| f | L_{p(\cdot)}\big(\mathbb{R}^n\big)\big\| \\ &\quad + \left\| \left(\int_0^1 t^{-s(x)q(x)} \big(d_t^M f(x)\big)^{q(x)} \frac{dt}{t} \right)^{1/q(x)} \middle| L_{p(\cdot)}\big(\mathbb{R}^n\big) \right\|,\end{aligned}$$

$$\begin{aligned}\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}_1 &:= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| \\ &\quad + \left\|\left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)}\big(d^M_{2^{-k}}f(x)\big)^{q(x)}\right)^{1/q(x)}\middle| L_{p(\cdot)}(\mathbb{R}^n)\right\| \\ &= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| + \big\|\big(2^{ks(x)}d^M_{2^{-k}}f(x)\big)^\infty_{k=0}\big|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|, \\ \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}_1 &:= \|f|L_{p(\cdot)}(\mathbb{R}^n)\| + \big\|\big(2^{ks(x)}d^M_{2^{-k}}f(x)\big)^\infty_{k=0}\big|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|.\end{aligned}$$

4.2 Main Theorem

Using the notation introduced above, we may now state the main result of this section.

Theorem 15 (i) *Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ with $p^+, q^+ < \infty$ and $s \in C^{\log}_{loc}(\mathbb{R}^n)$. Let $M \in \mathbb{N}$ with $M > s^+$ and let*

$$s^- > \sigma_{p^-,q^-} \cdot \left[1 + \frac{c_{\log}(s)}{n} \cdot \min\big(p^-, q^-\big)\right]. \tag{33}$$

Then

$$F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) = \big\{f \in L_{p(\cdot)}(\mathbb{R}^n) \cap \mathcal{S}'(\mathbb{R}^n) : \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^* < \infty\big\}$$

and $\|\cdot|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ and $\|\cdot|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^*$ are equivalent on $F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$. The same holds for $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$.

(ii) *Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ and $s \in C^{\log}_{loc}(\mathbb{R}^n)$. Let $M \in \mathbb{N}$ with $M > s^+$ and let*

$$s^- > \sigma_{p^-} \cdot \left[1 + \frac{c_{\log}(1/q)}{n} + \frac{c_{\log}(s)}{n} \cdot p^-\right]. \tag{34}$$

Then

$$B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) = \big\{f \in L_{p(\cdot)}(\mathbb{R}^n) \cap \mathcal{S}'(\mathbb{R}^n) : \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**} < \infty\big\}$$

and $\|\cdot|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ and $\|\cdot|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ are equivalent on $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

Remark 8 Let us comment on the rather technical conditions (33) and (34).

- If $\min(p^-, q^-) \geq 1$, then (33) becomes just $s^- > 0$. Furthermore, if p, q and s are constant functions, then (33) coincides with (29).
- If $p^- \geq 1$, then (34) reduces also to $s^- > 0$ and in the case of constant exponents we again recover (28).

As indicated already above, the proof is divided into several parts.

4.3 Preliminary Version of Theorem 15

This subsection contains a preliminary version of Theorem 15 (Lemma 16). Its proof represents the heart of the proof of Theorem 15. For better lucidity, it is again divided into more parts.

Lemma 16 *Under the conditions of Theorem 15, the following estimates hold for all* $f \in L_{p(\cdot)}(\mathbb{R}^n) \cap \mathcal{S}'(\mathbb{R}^n)$:

$$\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^* \approx \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}, \tag{35}$$

$$\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^*_1 \approx \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}_1, \tag{36}$$

$$\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}_1 \lesssim \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\| \lesssim \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}, \tag{37}$$

$$\|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}_1 \lesssim \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\| \lesssim \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}. \tag{38}$$

Proof Part I. First we prove (35) and (36). We discretize the inner part of $\|\cdot\|^*$ and obtain

$$\begin{aligned}
&\left[\int_0^\infty t^{-s(x)q(x)}\left(\int_B |\Delta^M_{th} f(x)|dh\right)^{q(x)}\frac{dt}{t}\right]^{1/q(x)}\\
&\quad=\left[\int_0^\infty t^{-s(x)q(x)}\left(t^{-n}\int_{tB} |\Delta^M_{\varkappa} f(x)|d\varkappa\right)^{q(x)}\frac{dt}{t}\right]^{1/q(x)}\\
&\quad=\left[\sum_{k=-\infty}^{\infty}\int_{2^{-k-1}}^{2^{-k}} t^{-s(x)q(x)}\left(t^{-n}\int_{tB} |\Delta^M_{\varkappa} f(x)|d\varkappa\right)^{q(x)}\frac{dt}{t}\right]^{1/q(x)}.
\end{aligned}\tag{39}$$

If $2^{-k-1}\le t\le 2^{-k}$, then $2^{ks(x)q(x)}\le t^{-s(x)q(x)}\le 2^{(k+1)s(x)q(x)}$ and

$$2^{kn}\int_{2^{-(k+1)}B} |\Delta^M_{\varkappa} f(x)|d\varkappa \lesssim t^{-n}\int_{tB} |\Delta^M_{\varkappa} f(x)|d\varkappa \lesssim 2^{(k+1)n}\int_{2^{-k}B} |\Delta^M_{\varkappa} f(x)|d\varkappa.$$

Plugging these estimates into (39), we may further estimate

$$\begin{aligned}
&\left[\int_0^\infty t^{-s(x)q(x)}\left(\int_B |\Delta^M_{th} f(x)|dh\right)^{q(x)}\frac{dt}{t}\right]^{1/q(x)}\\
&\quad\lesssim\left[\sum_{k=-\infty}^{\infty} 2^{(k+1)s(x)q(x)}\left(2^{kn}\int_{2^{-k}B} |\Delta^M_{\varkappa} f(x)|d\varkappa\right)^{q(x)}\right]^{1/q(x)}\\
&\quad\lesssim\left[\sum_{k=-\infty}^{\infty} 2^{ks(x)q(x)}\left(\int_{B} |\Delta^M_{2^{-k}\varkappa} f(x)|d\varkappa\right)^{q(x)}\right]^{1/q(x)}.
\end{aligned}$$

The estimate from below follows in the same manner. Finally, the proof of (36) is almost the same.

Part II. This part is devoted to the proof of the left hand side of (37). It is divided into several steps to make the presentation clearer.

Step 1. First, we point out that the estimate

$$\|f|L_{p(\cdot)}(\mathbb{R}^n)\| \lesssim \|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$$

follows from the characterization of $B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$ in terms of Nikol'skij representations (cf. Theorem 8.1 of [1]). We refer also to Remark 2.5.3/1 in [51]. The extension to F-spaces is then given by the simple embedding

$$\|f|L_{p(\cdot)}(\mathbb{R}^n)\| \lesssim \|f|B^{s(\cdot)-\varepsilon}_{p(\cdot),p(\cdot)}(\mathbb{R}^n)\| \lesssim \|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$$

with $\varepsilon > 0$ chosen small enough.

Step 2. Let $(\varphi_j)_{j\in\mathbb{N}_0}$ be the functions used in Definition 3. We use the decomposition

$$f = \sum_{l=-\infty}^{\infty} f_{(k+l)}, \quad k \in \mathbb{Z},$$

where $f_{(k+l)} = (\varphi_{k+l}\hat{f})^\vee$, or $= 0$ if $k+l<0$ and get

$$\begin{aligned}(\clubsuit) &:= \sum_{k=0}^{\infty} 2^{ks(x)q(x)}\left(\int_B |\Delta^M_{2^{-k}h} f(x)|dh\right)^{q(x)}\\ &= \sum_{k=0}^{\infty} 2^{ks(x)q(x)}\left(\int_B \left|\Delta^M_{2^{-k}h}\left(\sum_{l=-\infty}^{\infty} f_{(k+l)}\right)(x)\right|dh\right)^{q(x)}.\end{aligned}$$

If $q(x) \le 1$ then we proceed further

$$\begin{aligned}(\clubsuit) &\le \sum_{k=0}^{\infty} 2^{ks(x)q(x)}\left(\int_B \sum_{l=-\infty}^{\infty} |\Delta^M_{2^{-k}h} f_{(k+l)}(x)|dh\right)^{q(x)}\\ &\le \sum_{k=0}^{\infty}\sum_{l=-\infty}^{\infty} 2^{ks(x)q(x)}\left(\int_B |\Delta^M_{2^{-k}h} f_{(k+l)}(x)|dh\right)^{q(x)}.\end{aligned}$$

If $q(x) > 1$, we use Minkowski's inequality

$$\begin{aligned}(\clubsuit)^{1/q(x)} &\le \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)}\left(\int_B \sum_{l=-\infty}^{\infty} |\Delta^M_{2^{-k}h} f_{(k+l)}(x)|dh\right)^{q(x)}\right)^{1/q(x)}\\ &\le \sum_{l=-\infty}^{\infty}\left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)}\left(\int_B |\Delta^M_{2^{-k}h} f_{(k+l)}(x)|dh\right)^{q(x)}\right)^{1/q(x)}.\end{aligned}$$

We split in both cases

$$\sum_{l=-\infty}^{\infty} \cdots = I + II = \sum_{l=-\infty}^{0} \cdots + \sum_{l=1}^{\infty} \cdots \tag{40}$$

Step 3. We estimate the first summand with $l \leq 0$.

We use Lemma 22 in the form

$$\left|\Delta_h^M f_{(k+l)}(x)\right| \leq C \max\big(1, |bh|^a\big) \cdot \min\big(1, |bh|^M\big) P_{b,a} f_{(k+l)}(x),$$

where $a > 0$ is arbitrary, $b = 2^{k+l}$ and

$$P_{b,a} f(x) = \sup_{z \in \mathbb{R}^n} \frac{|f(x-z)|}{1 + |bz|^a}.$$

Furthermore, we use this estimate with $2^{-k}h$ instead of h. We obtain

$$\begin{aligned} \int_B \left|\Delta_{2^{-k}h}^M f_{(k+l)}(x)\right| dh &\lesssim \int_B \max\big(1, \left|b2^{-k}h\right|^a\big) \cdot \min\big(1, \left|b2^{-k}h\right|^M\big) P_{b,a} f_{(k+l)}(x) dh \\ &\lesssim 2^{lM} P_{2^{k+l},a} f_{(k+l)}(x). \end{aligned} \tag{41}$$

The last inequality follows from $\max(1, |b2^{-k}h|^a) \leq 1$ (recall that $l \leq 0$ and $|h| \leq 1$) and $\min(1, |b2^{-k}h|^M) \leq 2^{lM}$.

If $q(x) \leq 1$, we estimate the first sum in (40)

$$\begin{aligned} I &\leq \sum_{l=-\infty}^{0} \sum_{k=0}^{\infty} 2^{ks(x)q(x)} \left(\int_B \left|\Delta_{2^{-k}h}^M f_{(k+l)}(x)\right| dh\right)^{q(x)} \\ &\lesssim \sum_{l=-\infty}^{0} \sum_{k=0}^{\infty} 2^{ks(x)q(x)} \big(2^{lM} P_{2^{k+l},a} f_{(k+l)}(x)\big)^{q(x)} \\ &= \sum_{l=-\infty}^{0} 2^{l(M-s(x))q(x)} \sum_{k=0}^{\infty} 2^{(k+l)s(x)q(x)} P_{2^{k+l},a}^{q(x)} f_{(k+l)}(x) \\ &\approx \sum_{k=0}^{\infty} 2^{ks(x)q(x)} P_{2^k,a}^{q(x)} f_{(k)}(x), \end{aligned}$$

where the last estimate makes use of $M > s^+$, $q^- > 0$ and the fact that $f_{(k+l)} = 0$ for $k + l < 0$.

If $q(x) > 1$, we proceed in a similar way to obtain

$$I^{1/q(x)} \leq \sum_{l=-\infty}^{0} \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)} \left(\int_B \left|\Delta_{2^{-k}h}^M f_{(k+l)}(x)\right| dh\right)^{q(x)}\right)^{1/q(x)}$$

$$
\begin{aligned}
&\lesssim \sum_{l=-\infty}^{0} \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)} \big(2^{lM} P_{2^{k+l},a} f_{(k+l)}(x)\big)^{q(x)} \right)^{1/q(x)} \\
&= \sum_{l=-\infty}^{0} 2^{l(M-s(x))} \left(\sum_{k=0}^{\infty} 2^{(k+l)s(x)q(x)} P^{q(x)}_{2^{k+l},a} f_{(k+l)}(x) \right)^{1/q(x)} \\
&\lesssim \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)} P^{q(x)}_{2^{k},a} f_{(k)}(x) \right)^{1/q(x)} .
\end{aligned}
$$

We have used in the last estimate again $M > s^+$ and the definition of $f_{(k+l)}$.

Hence,

$$
I^{1/q(x)} \lesssim \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)} P^{q(x)}_{2^{k},a} f_{(k)}(x) \right)^{1/q(x)}
$$

holds for all $x \in \mathbb{R}^n$.

Finally, we obtain

$$
\begin{aligned}
\big\| I^{1/q(\cdot)} | L_{p(\cdot)}(\mathbb{R}^n) \big\| &\lesssim \left\| \left(\sum_{k=0}^{\infty} 2^{ks(x)q(x)} P^{q(x)}_{2^{k},a} f_{(k)}(x) \right)^{1/q(x)} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \\
&= \big\| \big(2^{ks(x)} P_{2^k,a} f_{(k)}(x)\big)_{k=0}^{\infty} | L_{p(\cdot)}(\ell_{q(\cdot)}) \big\| \\
&\lesssim \big\| \big(2^{ks(\cdot)} f_{(k)}\big)_{k=0}^{\infty} | L_{p(\cdot)}(\ell_{q(\cdot)}) \big\| = \big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \big\|, \qquad (42)
\end{aligned}
$$

where we used the boundedness of Peetre maximal operator as described in Theorem 14 for $a > 0$ large enough.

Step 4. We estimate the second summand in (40) with $l > 0$. If $\min(p^-, q^-) > 1$, then we put $\lambda = 1$. Otherwise we choose real parameters $0 < \lambda < \min(p^-, q^-)$ and $a > 0$ such that

$$
a > \frac{n}{\min(p^-, q^-)} + c_{\log}(s)
$$

and $a(1-\lambda) < s^-$. Due to (33), this is always possible.

We start again with estimates of the ball means of differences. We use Lemma 22 and (30) to obtain

$$
\begin{aligned}
&\int_B \big|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\big| dh \\
&\quad = \int_B \big|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\big|^{\lambda} \cdot \big|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\big|^{1-\lambda} dh \\
&\quad \lesssim \int_B \big(\max\big(1, |2^{k+l}2^{-k}h|^a\big) \min\big(1, |2^{k+l}2^{-k}h|^M\big) P_{2^{k+l},a} f_{(k+l)}(x)\big)^{1-\lambda} \\
&\qquad \cdot \big|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\big|^{\lambda} dh
\end{aligned}
$$

$$\le \left(2^{la} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{1-\lambda} \int_B \left|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\right|^\lambda dh$$

$$\le \left(2^{la} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{1-\lambda} \sum_{j=0}^{M} c_{j,M} \int_B \left|f_{(k+l)}\left(x+j2^{-k}h\right)\right|^\lambda dh, \tag{43}$$

where the constants $c_{j,M}$ are given by (30).

We shall deal in detail only with the term with $j=1$. The term with $j=0$ is much simpler to handle (as there the integration over $h \in B$ immediately disappears) and this case reduces essentially to Hölder's inequality and boundedness of the Peetre maximal operator. The terms with $2 \le j \le M$ may be handled in the same way as the one with $j=1$.

We use Lemma 20 with $r=\lambda$ in the form

$$\left|f_{(k+l)}(y)\right|^\lambda \lesssim \left(\eta_{k+l,2m} * |f_{(k+l)}|^\lambda\right)(y),$$

with $m > \max(n, c_{\log}(s))$, Lemmas 23 and 19 to get

$$\begin{aligned}
&2^{ks(x)\lambda} \int_B \left|f_{(k+l)}\left(x+2^{-k}h\right)\right|^\lambda dh \\
&\quad\lesssim 2^{ks(x)\lambda} \int_B \left(\eta_{k+l,2m} * |f_{(k+l)}|^\lambda\right)\left(x+2^{-k}h\right) dh \\
&\quad= 2^{ks(x)\lambda}\left(\left[2^{kn}\chi_{2^{-k}B}\right] * \eta_{k+l,2m} * |f_{(k+l)}|^\lambda\right)(x) \\
&\quad\lesssim 2^{ks(x)\lambda}\left(\eta_{k,2m} * |f_{(k+l)}|^\lambda\right)(x) \\
&\quad\lesssim \left(\eta_{k,m} * \left|2^{ks(\cdot)} f_{(k+l)}\right|^\lambda\right)(x) \\
&\quad\le 2^{-ls^-\lambda}\left(\eta_{k,m} * \left|2^{(k+l)s(\cdot)} f_{(k+l)}\right|^\lambda\right)(x).
\end{aligned} \tag{44}$$

We insert (44) into (43) and arrive at

$$\begin{aligned}
&2^{ks(x)} \int_B \left|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\right| dh \\
&\quad\lesssim 2^{la(1-\lambda)-ls^-}\left(2^{(k+l)s(x)} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{1-\lambda}\left(\eta_{k,m} * \left|2^{(k+l)s(\cdot)} f_{(k+l)}\right|^\lambda\right)(x).
\end{aligned} \tag{45}$$

If $q(x) > 1$, we proceed further with the use of Hölder's inequality

$$\begin{aligned}
II^{1/q(x)} &\lesssim \sum_{l=1}^{\infty} 2^{la(1-\lambda)-ls^-}\Bigg(\sum_{k=0}^{\infty}\left(2^{(k+l)s(x)} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{(1-\lambda)q(x)} \\
&\qquad\cdot\left(\eta_{k,m} * \left|2^{(k+l)s(\cdot)} f_{(k+l)}\right|^\lambda\right)^{q(x)}(x)\Bigg)^{1/q(x)} \\
&\le \sum_{l=1}^{\infty} 2^{la(1-\lambda)-ls^-}\left(\sum_{k=0}^{\infty}\left(2^{(k+l)s(x)} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{q(x)}\right)^{(1-\lambda)/q(x)}
\end{aligned}$$

$$\cdot\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda/q(x)}$$

$$=\left(\sum_{k=0}^{\infty}\left(2^{ks(x)}P_{2^k,a}f_{(k)}(x)\right)^{q(x)}\right)^{(1-\lambda)/q(x)}$$

$$\cdot\sum_{l=1}^{\infty}2^{la(1-\lambda)-ls^-}\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda/q(x)}.$$

If $q(x)\le 1$, we obtain in a similar way

$$II\lesssim\sum_{l=1}^{\infty}2^{(la(1-\lambda)-ls^-)q(x)}\sum_{k=0}^{\infty}\left(2^{(k+l)s(x)}P_{2^{k+l},a}f_{(k+l)}(x)\right)^{(1-\lambda)q(x)}$$

$$\cdot\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)}(x)$$

$$\le\sum_{l=1}^{\infty}2^{(la(1-\lambda)-ls^-)q(x)}\left(\sum_{k=0}^{\infty}\left(2^{(k+l)s(x)}P_{2^{k+l},a}f_{(k+l)}(x)\right)^{q(x)}\right)^{1-\lambda}$$

$$\cdot\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda}$$

$$=\left(\sum_{k=0}^{\infty}\left(2^{ks(x)}P_{2^k,a}f_{(k)}(x)\right)^{q(x)}\right)^{1-\lambda}$$

$$\cdot\sum_{l=1}^{\infty}2^{(la(1-\lambda)-ls^-)q(x)}\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda}$$

and further (with use of Lemma 24)

$$II^{1/q(x)}\lesssim\left(\sum_{k=0}^{\infty}\left(2^{ks(x)}P_{2^k,a}f_{(k)}(x)\right)^{q(x)}\right)^{(1-\lambda)/q(x)}$$

$$\cdot\left(\sum_{l=1}^{\infty}2^{(la(1-\lambda)-ls^-)q(x)}\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda}\right)^{1/q(x)}$$

$$\lesssim\left(\sum_{k=0}^{\infty}\left(2^{ks(x)}P_{2^k,a}f_{(k)}(x)\right)^{q(x)}\right)^{(1-\lambda)/q(x)}$$

$$\cdot\sum_{l=1}^{\infty}2^{1/2\cdot(la(1-\lambda)-ls^-)}\left(\sum_{k=0}^{\infty}\left(\eta_{k,m}*\left|2^{(k+l)s(\cdot)}f_{(k+l)}\right|^{\lambda}\right)^{q(x)/\lambda}(x)\right)^{\lambda/q(x)}.$$

If we denote

$$F(x) := \left(\sum_{k=0}^{\infty} \bigl(2^{ks(x)} P_{2^k,a} f_{(k)}(x) \bigr)^{q(x)} \right)^{1/q(x)}, \quad x \in \mathbb{R}^n$$

and

$$B_{k+l}(x) := \left| 2^{(k+l)s(x)} f_{(k+l)}(x) \right|, \quad x \in \mathbb{R}^n$$

we get for $\delta := -1/2 \cdot (a(1-\lambda) - s^-) > 0$

$$II^{1/q(x)} \lesssim F(x)^{1-\lambda} \cdot \sum_{l=1}^{\infty} 2^{-l\delta} \left(\sum_{k=0}^{\infty} \bigl(\eta_{k,m} * B_{k+l}^{\lambda} \bigr)^{q(x)/\lambda}(x) \right)^{\lambda/q(x)}. \tag{46}$$

We use $\|F_1^{1-\lambda} F_2^{\lambda}\|_{p(\cdot)} \le 2\|F_1\|_{p(\cdot)}^{1-\lambda} \|F_2\|_{p(\cdot)}^{\lambda}$, cf. [17, Lemma 3.2.20], and suppose that the $L_{p(\cdot)}$-(quasi-)norm is equivalent to an r-norm with $0 < r \le 1$. Together with Lemma 21 we arrive at

$$\begin{aligned}
\|II^{1/q(x)}\|_{p(\cdot)}^{r} &\lesssim \|F(x)\|_{p(\cdot)}^{(1-\lambda)r} \cdot \left\| \sum_{l=1}^{\infty} 2^{-l\delta} \left(\sum_{k=0}^{\infty} \bigl(\eta_{k,m} * B_{k+l}^{\lambda} \bigr)^{q(x)/\lambda}(x) \right)^{1/q(x)} \right\|_{p(\cdot)}^{\lambda r} \\
&\lesssim \|F(x)\|_{p(\cdot)}^{(1-\lambda)r} \cdot \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| \left(\sum_{k=0}^{\infty} \bigl(\eta_{k,m} * B_{k+l}^{\lambda} \bigr)^{q(x)/\lambda}(x) \right)^{1/q(x)} \right\|_{p(\cdot)}^{\lambda r} \\
&\lesssim \left\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \right\|^{(1-\lambda)r} \\
&\quad \cdot \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| \bigl(\eta_{k,m} * B_{k+l}^{\lambda}(x) \bigr)_{k=0}^{\infty} \right\|_{L_{p(\cdot)/\lambda}(\ell_{q(\cdot)/\lambda})}^{r} \\
&\lesssim \left\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \right\|^{(1-\lambda)r} \cdot \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| \bigl(B_{k+l}^{\lambda}(x) \bigr)_{k=0}^{\infty} \right\|_{L_{p(\cdot)/\lambda}(\ell_{q(\cdot)/\lambda})}^{r} \\
&\lesssim \left\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \right\|^{(1-\lambda)r} \cdot \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| \bigl(B_{k}^{\lambda}(x) \bigr)_{k=0}^{\infty} \right\|_{L_{p(\cdot)/\lambda}(\ell_{q(\cdot)/\lambda})}^{r} \\
&\lesssim \left\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \right\|^{(1-\lambda)r} \cdot \left\| B_k(x) \right\|_{L_{p(\cdot)}(\ell_{q(\cdot)})}^{\lambda r} \\
&\lesssim \left\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \right\|^{r},
\end{aligned} \tag{47}$$

which finishes the proof.

Part III. We prove the right hand side of (37). We follow again essentially [51, Sect. 2.5.9] with some modifications as presented in [53]. Roughly speaking, compared to the case of constant exponents, only minor modifications are necessary.

Let $\psi \in C_0^\infty(\mathbb{R}^n)$ with $\psi(x)=1$, $|x|\le 1$ and $\psi(x)=0$, $|x|>3/2$. We define

$$\varphi_0(x)=(-1)^{M+1}\sum_{\mu=0}^{M-1}(-1)^\mu\binom{M}{\mu}\psi\bigl((M-\mu)x\bigr).$$

It follows that $\varphi_0\in C_0^\infty(\mathbb{R}^n)$ with $\varphi(x)=0$, $|x|>3/2$ and $\varphi(x)=1$, $|x|<1/M$. We also put $\varphi_j(x)=\varphi_0(2^{-j}x)-\varphi_0(2^{-j+1}x)$ for $j\ge 1$. This is the decomposition of unity we used in the definition of $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$, cf. Definition 3. Recall that due to [16] and [30], this (quasi-)norm of $\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ does not depend on the choice of the decomposition of unity.

We observe that

$$\varphi_0(x)=(-1)^{M+1}\bigl(\Delta_x^M\psi(0)-(-1)^M\bigr),$$

and

$$\bigl(\mathcal{F}^{-1}\varphi_j\mathcal{F}f\bigr)(x)=\begin{cases}(\mathcal{F}^{-1}\Delta_\xi^M\psi(0)\mathcal{F}f)(x)+(-1)^{M+1}f(x), & j=0,\\ (\mathcal{F}^{-1}(\Delta^M_{2^{-j}\xi}\psi(0)-\Delta^M_{2^{-j+1}\xi}\psi(0))\mathcal{F}f)(x), & j\ge 1.\end{cases}\tag{48}$$

Furthermore, a straightforward calculation shows that

$$\begin{aligned}\bigl|\bigl(\mathcal{F}^{-1}\bigl(\Delta^M_{2^{-j}\xi}\psi(0)\bigr)\mathcal{F}f\bigr)(x)\bigr| &= \left|\sum_{u=0}^M(-1)^u\mathcal{F}^{-1}\bigl[\psi\bigl((M-u)2^{-j}\cdot\bigr)\mathcal{F}f\bigr](x)\right|\\ &\approx\left|\sum_{u=0}^M(-1)^u\mathcal{F}^{-1}\bigl[\psi\bigl((M-u)2^{-j}\cdot\bigr)\bigr]*f(x)\right|\\ &\approx\left|\sum_{u=0}^M(-1)^u\int_{\mathbb{R}^n}\mathcal{F}^{-1}\psi(h)f\bigl(x-(M-u)2^{-j}h\bigr)dh\right|\\ &=\left|\int_{\mathbb{R}^n}\hat{\psi}(h)\Delta^M_{2^{-j}h}f(x)dh\right|\\ &\le\int_{\mathbb{R}^n}\bigl|\hat{\psi}(h)\bigr|\cdot\bigl|\Delta^M_{2^{-j}h}f(x)\bigr|dh \end{aligned}\tag{49}$$

holds for every $j\in\mathbb{N}_0$. We denote $g=\hat{\psi}\in\mathcal{S}(\mathbb{R}^n)$ and obtain

$$\begin{aligned}\bigl\|f|F^{s(\cdot)}_{p(\cdot),q(\cdot)}\bigl(\mathbb{R}^n\bigr)\bigr\| &\approx\bigl\|2^{js(x)}\bigl(\mathcal{F}^{-1}\varphi_j\mathcal{F}f\bigr)(x)|L_{p(\cdot)}(\ell_{q(\cdot)})\bigr\|\\ &\lesssim\bigl\|f|L_{p(\cdot)}\bigl(\mathbb{R}^n\bigr)\bigr\|\\ &\quad+\left\|2^{js(x)}\int_{\mathbb{R}^n}\bigl|g(h)\bigr|\cdot\bigl|\Delta^M_{2^{-j}h}f(x)\bigr|dh|L_{p(\cdot)}(\ell_{q(\cdot)})\right\|.\end{aligned}\tag{50}$$

The rest of this part consists essentially of using the property of $g\in\mathcal{S}(\mathbb{R}^n)$ to come from (50) to $\|\cdot\|^{**}$.

We denote

$$I_0 := B, \qquad I_u := 2^u B \setminus 2^{u-1}B, \quad u \in \mathbb{N}$$

and use $|g(h)| \le c2^{-ur}$, $h \in I_u$ with r taken large enough (recall that $g \in \mathcal{S}(\mathbb{R}^n)$) and estimate

$$\begin{aligned}\int_{\mathbb{R}^n} |g(h)| \cdot |\Delta^M_{2^{-j}h} f(x)| dh &= \sum_{u=0}^{\infty} \int_{I_u} |g(h)| \cdot |\Delta^M_{2^{-j}h} f(x)| dh \\ &\lesssim \sum_{u=0}^{\infty} 2^{-ur} 2^{jn} \int_{2^{u-j}B} |\Delta^M_h f(x)| dh \\ &= \sum_{u=0}^{\infty} 2^{u(n-r)} 2^{-(u-j)n} \int_{2^{u-j}B} |\Delta^M_h f(x)| dh. \qquad (51)\end{aligned}$$

We put

$$G_j(x) := 2^{js(x)} \left| \left(\mathcal{F}^{-1} \left(\Delta^M_{2^{-j}\xi} \psi(0) \right) \mathcal{F} f \right)(x) \right|, \quad j \in \mathbb{N}_0$$

and

$$g_k(x) := 2^{ks(x)} 2^{kn} \int_{2^{-k}B} |\Delta^M_h f(x)| dh, \quad k \in \mathbb{Z}.$$

Using (48), (49) and (51), we obtain the estimate

$$\begin{aligned} G_j(x) &\lesssim 2^{js(x)} \sum_{u=0}^{\infty} 2^{u(n-r)} 2^{-(u-j)n} \int_{2^{u-j}B} |\Delta^M_h f(x)| dh \\ &= \sum_{k=-\infty}^{j} 2^{(j-k)s(x)} 2^{(j-k)(n-r)} 2^{ks(x)} 2^{kn} \int_{2^{-k}B} |\Delta^M_h f(x)| dh \\ &= \sum_{k=-\infty}^{j} 2^{(j-k)(s(x)+n-r)} g_k(x) \le \sum_{k=-\infty}^{\infty} 2^{|j-k| \cdot (s(x)+n-r)} g_k(x). \qquad (52)\end{aligned}$$

Choosing $r > s^+ + n$ and applying Lemma 9 then finishes the proof.

Part IV. The proof of the left hand side of (38) follows in the same manner as in Part II. We shall describe the necessary modifications. First, let us mention, that the condition $q^+ < \infty$ was used only in the application of Lemma 21. In the rest of the arguments also the case $q(x) = \infty$ may be incorporated with only slight change of notation.

Let us put

$$f^{(k)}(x) := 2^{ks(x)} \int_B |\Delta^M_{2^{-k}h} f(x)| dh, \quad x \in \mathbb{R}^n.$$

We obtain (in analogue to (40))

$$f^{(k)} \le f^{(k),I} + f^{(k),II} := \sum_{l=-\infty}^{0} 2^{ks(x)} \int_B \left|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\right| dh \\ + \sum_{l=1}^{\infty} 2^{ks(x)} \int_B \left|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\right| dh.$$

We estimate the first sum using (41) and get

$$f^{(k),I} \lesssim \sum_{l=-\infty}^{0} 2^{l(M-s^+)} g^1_{k+l} = \sum_{u=0}^{k} 2^{(u-k)(M-s^+)} g^1_u \le \sum_{u=0}^{\infty} 2^{-|u-k|(M-s^+)} g^1_u,$$

where $g^1_u := 2^{us(x)} P_{2^u,a} f_{(u)}(x)$. The application of Lemma 9 and Theorem 14 with $a > 0$ large enough gives

$$\left\|\left(f^{(k),I}\right)_{k=0}^{\infty} | \ell_{q(\cdot)}(L_{p(\cdot)})\right\| \lesssim \left\|(g_u)_{u=0}^{\infty} | \ell_{q(\cdot)}(L_{p(\cdot)})\right\| \lesssim \left\| f | B^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\right\|.$$

To estimate $f^{(k),II}$, we proceed as in the Step 4 of Part II. If $p^- > 1$, we choose again $\lambda = 1$, otherwise we take $0 < \lambda < p^-$ and

$$a > \frac{n + c_{\log}(1/q)}{p^-} + c_{\log}(s)$$

such that $a(1-\lambda) < s^-$. This is possible due to (34).

We use (44) with $m > \max(n + c_{\log}(1/q), c_{\log}(s))$ to get

$$f^{(k),II} \lesssim \sum_{l=1}^{\infty} 2^{la(1-\lambda)-ls^-} \left(2^{(k+l)s(x)} P_{2^{k+l},a} f_{(k+l)}(x)\right)^{1-\lambda} \\ \cdot \left(\eta_{k,m} * \left|2^{(k+l)s(\cdot)} f_{(k+l)}(\cdot)\right|^{\lambda}\right)(x) \\ = \sum_{l=1}^{\infty} 2^{la(1-\lambda)-ls^-} \left(g^1_{k+l}(x)\right)^{1-\lambda} \cdot \left(\eta_{k,m} * \left(g^2_{k+l}\right)^{\lambda}\right)(x), \qquad (53)$$

where $g^2_{k+l}(x) := |2^{(k+l)s(x)} f_{(k+l)}(x)|$. We take the $\ell_{q(\cdot)}(L_{p(\cdot)})$ (quasi-)norm of the last expression—and assume that it is equivalent to some r-norm. This gives for $\delta := s^- - a(1-\lambda) > 0$ the following estimate

$$\left\| f^{(k),II} | \ell_{q(\cdot)}(L_{p(\cdot)})\right\|^r \\ \lesssim \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| \left(g^1_{k+l}(x)\right)^{1-\lambda} \cdot \left(\eta_{k,m} * \left(g^2_{k+l}\right)^{\lambda}\right)(x) | \ell_{q(\cdot)}(L_{p(\cdot)})\right\|^r \\ \lesssim \sum_{l=1}^{\infty} 2^{-l\delta r} \left\| g^1_{k+l} | \ell_{q(\cdot)}(L_{p(\cdot)})\right\|^{(1-\lambda)r} \cdot \left\| \left[\eta_{k,m} * \left(g^2_{k+l}\right)^{\lambda}\right]^{1/\lambda} | \ell_{q(\cdot)}(L_{p(\cdot)})\right\|^{\lambda r}$$

$$
\begin{aligned}
&\lesssim \sum_{l=1}^{\infty} 2^{-l\delta r} \big\| g_k^1 | \ell_{q(\cdot)}(L_{p(\cdot)}) \big\|^{(1-\lambda)r} \cdot \big\| \eta_{k,m} * \big(g_{k+l}^2\big)^{\lambda} | \ell_{q(\cdot)/\lambda}(L_{p(\cdot)/\lambda}) \big\|^{r} \\
&\lesssim \sum_{l=1}^{\infty} 2^{-l\delta r} \big\| g_k^1 | \ell_{q(\cdot)}(L_{p(\cdot)}) \big\|^{(1-\lambda)r} \cdot \big\| \big(g_{k+l}^2\big)^{\lambda} | \ell_{q(\cdot)/\lambda}(L_{p(\cdot)/\lambda}) \big\|^{r} \\
&\lesssim \sum_{l=1}^{\infty} 2^{-l\delta r} \big\| g_k^1 | \ell_{q(\cdot)}(L_{p(\cdot)}) \big\|^{(1-\lambda)r} \cdot \big\| g_{k+l}^2 | \ell_{q(\cdot)}(L_{p(\cdot)}) \big\|^{\lambda r} \\
&\lesssim \big\| f | B_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \big\|^{r}. \qquad (54)
\end{aligned}
$$

We have used Lemma 10 and Lemma 25.

Part V. The right hand side inequality of (38) follows also along the same line as in Part III. We just combine (52) with the choice $r > s^+ + n$ and apply Lemma 9. □

4.4 Proof of Theorem 15

This section is devoted to the proof of Theorem 15. We start with the case of constant q. In that case, the usual Hardy-Littlewood maximal operator

$$Mf(x) = \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |f(y)| dy$$

is bounded on $\ell_q(L_{p(\cdot)})$ and $L_{p(\cdot)}(\ell_q)$. Indeed, the following lemma is a consequence of [12] and [17, Theorem 4.3.8].

Lemma 17

(i) *Let* $p \in \mathcal{P}^{\log}(\mathbb{R}^n)$ *with* $1 < p^- \le p^+ < \infty$ *and* $1 < q < \infty$. *Then*

$$\big\| (Mf_j)_{j=-\infty}^{\infty} | L_{p(\cdot)}(\ell_q) \big\| \lesssim \big\| (f_j)_{j=-\infty}^{\infty} | L_{p(\cdot)}(\ell_q) \big\|$$

for all $(f_j)_{j=-\infty}^{\infty} \in L_{p(\cdot)}(\ell_q)$.

(ii) *Let* $p \in \mathcal{P}^{\log}(\mathbb{R}^n)$ *with* $p^- > 1$ *and* $0 < q \le \infty$. *Then*

$$\big\| (Mf_j)_{j=-\infty}^{\infty} | \ell_q(L_{p(\cdot)}) \big\| \lesssim \big\| (f_j)_{j=-\infty}^{\infty} | \ell_q(L_{p(\cdot)}) \big\|$$

for all $(f_j)_{j=-\infty}^{\infty} \in \ell_q(L_{p(\cdot)})$.

Proof of Theorem 15 With the help of Lemma 17, we prove Theorem 15 for q constant. In view of Lemma 16, it is enough to prove

$$\big\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \big\|^{**} \lesssim \big\| f | F_{p(\cdot),q(\cdot)}^{s(\cdot)}(\mathbb{R}^n) \big\| \qquad (55)$$

and a corresponding analogue for the B-spaces.

Part I. In this part we point out the necessary modifications in the proof of Lemma 16 to obtain a characterization by ball means of differences for $B_{p(\cdot),q}^{s(\cdot)}(\mathbb{R}^n)$

and $F^{s(\cdot)}_{p(\cdot),q}(\mathbb{R}^n)$. The proof follows the scheme of Part II of the proof of Lemma 16. We start with $\sum_{k=-\infty}^{\infty}$ instead of $\sum_{k=0}^{\infty}$. With this modification the Steps 1–3 go through without any other changes and we obtain (42) again (just recall that $f_{(k)}=0$ if $k<0$).

Due to the boundedness of the maximal operator there is no need for the use of r-trick and convolution with $\eta_{\nu,m}$. The analogue of (43), (44) and (45) now reads as follows:

$$\begin{aligned}
&2^{ks(x)}\int_B \big|\Delta^M_{2^{-k}h} f_{(k+l)}(x)\big|dh\\
&\quad\lesssim \big(2^{ks(x)}2^{la}P_{2^{k+l},a}f_{(k+l)}(x)\big)^{1-\lambda}\\
&\qquad\cdot\sum_{j=0}^{M} c_{j,M}\int_B 2^{ks(x+j2^{-kh})\lambda}\big|f_{(k+l)}\big(x+j2^{-k}h\big)\big|^{\lambda}dh\\
&\quad\le 2^{la(1-\lambda)-ls^-}\big(2^{(k+l)s(x)}P_{2^{k+l},a}f_{(k+l)}(x)\big)^{1-\lambda}\\
&\qquad\cdot\sum_{j=0}^{M} c_{j,M}\int_B 2^{(k+l)s(x+j2^{-kh})\lambda}\big|f_{(k+l)}\big(x+j2^{-k}h\big)\big|^{\lambda}dh\\
&\quad\lesssim 2^{la(1-\lambda)-ls^-}\big(2^{(k+l)s(x)}P_{2^{k+l},a}f_{(k+l)}(x)\big)^{1-\lambda}\\
&\qquad\cdot\sum_{j=0}^{M} c_{j,M}M\big(\big|2^{(k+l)s(\cdot)}f_{(k+l)}(\cdot)\big|^{\lambda}\big)(x),
\end{aligned}$$

where we used Hölder's regularity of $s(\cdot)$, see (5). As a consequence, we obtain

$$II^{1/q(x)}\lesssim F(x)^{1-\lambda}\cdot\sum_{l=1}^{\infty}2^{-l\delta}\left(\sum_{k=-l}^{\infty}\big(MB^{\lambda}_{k+l}\big)^{q(x)/\lambda}(x)\right)^{\lambda/q(x)}$$

instead of (46). The rest then follows in the same manner with the help of Lemma 17 and the proof of (55) is finished.

The proof of

$$\big\|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^{**}\lesssim\big\|f|B^{s(\cdot)}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|$$

follows along the same lines. Especially, we get

$$f^{(k),II}\lesssim\sum_{l=1}^{\infty}2^{la(1-\lambda)-ls^-}\big(g^1_{k+l}(x)\big)^{1-\lambda}\cdot\big(M\big(g^2_{k+l}\big)^{\lambda}\big)(x)$$

instead of (53). The rest follows again by Lemma 17.

Part II. Finally, we present how the characterization for q constant can help us to improve on the case of variable exponent $q(\cdot)$.

In view of Lemma 16, it is enough to show that

$$\left\| \left(\sum_{k=-\infty}^{0} 2^{ks(x)q(x)} \big(d^M_{2^{-k}} f(x)\big)^{q(x)} \right)^{1/q(x)} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \lesssim \big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \big\|.$$

But this is a consequence of

$$\begin{aligned} &\left\| \left(\sum_{k=-\infty}^{0} 2^{ks(x)q(x)} \big(d^M_{2^{-k}} f(x)\big)^{q(x)} \right)^{1/q(x)} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \\ &\quad\lesssim \left\| \left(\sum_{k=-\infty}^{0} 2^{k(s(x)-\varepsilon)q^-} \big(d^M_{2^{-k}} f(x)\big)^{q^-} \right)^{1/q^-} \middle| L_{p(\cdot)}(\mathbb{R}^n) \right\| \\ &\quad\lesssim \big\| f | F^{s(\cdot)-\varepsilon}_{p(\cdot),q^-}(\mathbb{R}^n) \big\| \lesssim \big\| f | F^{s(\cdot)}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \big\|, \end{aligned}$$

where $\varepsilon > 0$ is small enough and we used the differences characterization for fixed q and a trivial embedding theorem.

The same arguments apply for the Besov spaces and the proof is finished. □

Remark 9 The somewhat complicated proof of Theorem 15 would work more direct and simpler if we could use versions of Lemmas 10 and 21 in (47) and (54) where the $\ell_{q(\cdot)}$ summation runs over $\nu \in \mathbb{Z}$.

For Triebel-Lizorkin spaces there seems to exist such an extension [13], but for Besov spaces the proof of Lemma 10 in [1] seems to be to customized to the situation $\nu \in \mathbb{N}_0$.

4.5 Ball Means of Differences for 2-Microlocal Spaces

As already remarked in Sect. 2.2 all the proofs for spaces of variable smoothness do also serve for 2-microlocal spaces. One just has to use the definition of admissible weight sequences and the property (5), see Remark 2.

First of all we give the notation for the (quasi-)norms. For simplicity we just use the discrete versions, although it is also possible to give continuous versions of 2-microlocal weights, see [55, Definition 4.1]. In analogy to the spaces of variable smoothness we introduce the following norms

$$\big\| f | B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \big\|^{**} = \big\| f | L_{p(\cdot)}(\mathbb{R}^n) \big\| + \big\| \big(w_k(x) d^M_{2^{-k}} f(x)\big)^{\infty}_{k=-\infty} \big| \ell_{q(\cdot)}(L_{p(\cdot)}) \big\|$$

and

$$\big\| f | F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n) \big\|^{**} = \big\| f | L_{p(\cdot)}(\mathbb{R}^n) \big\| + \big\| \big(w_k(x) d^M_{2^{-k}} f(x)\big)^{\infty}_{k=-\infty} \big| L_{p(\cdot)}(\ell_{q(\cdot)}) \big\|.$$

Finally, the preceding calculations show that the following theorem is true.

Theorem 18 (i) *Let $p, q \in \mathcal{P}^{\log}(\mathbb{R}^n)$ with $p^+, q^+ < \infty$ and $\boldsymbol{w} \in \mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$. Let $M > \alpha_2$ and*

$$\alpha_1 > \sigma_{p^-,q^-} \cdot \left[1 + \frac{\alpha}{n} \cdot \min\big(p^-, q^-\big) \right]. \tag{56}$$

Then

$$F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)=\big\{f\in L_{p(\cdot)}\big(\mathbb{R}^n\big):\big\|f|F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^{**}<\infty\big\}$$

and $\|\cdot|F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ *and* $\|\cdot|F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ *are equivalent on* $F^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

(ii) *Let* $p,q\in\mathcal{P}^{\log}(\mathbb{R}^n)$ *and* $\boldsymbol{w}\in\mathcal{W}^{\alpha}_{\alpha_1,\alpha_2}$. *Let* $M>\alpha_2$ *and*

$$\alpha_1>\sigma_{p^-}\cdot\left[1+\frac{c_{\log}(1/q)}{n}+\frac{\alpha}{n}\cdot p^-\right]. \tag{57}$$

Then

$$B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)=\big\{f\in L_{p(\cdot)}\big(\mathbb{R}^n\big):\big\|f|B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}\big(\mathbb{R}^n\big)\big\|^{**}<\infty\big\}$$

and $\|\cdot|B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|$ *and* $\|\cdot|B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)\|^{**}$ *are equivalent on* $B^{\boldsymbol{w}}_{p(\cdot),q(\cdot)}(\mathbb{R}^n)$.

Remark 10 Again, if $\min(p^-,q^-)\geq 1$ in the F-case, or $p^-\geq 1$ in the B-case, then the conditions (56) and (57) simplify to $\alpha_1>0$. In the case of constant exponents p,q we obtain similar results to [8] and [34].

4.6 Lemmas

The following lemma is a variant of Lemma 6.1 from [16].

Lemma 19 *Let* $s\in C^{\log}_{loc}(\mathbb{R}^n)$ *and let* $R\geq c_{\log}(s)$, *where* $c_{\log}(s)$ *is the constant from* (2) *for* $s(\cdot)$. *Then*

$$2^{\nu s(x)}\eta_{\nu,m+R}(x-y)\leq c\,2^{\nu s(y)}\eta_{\nu,m}(x-y)$$

holds for all $x,y\in\mathbb{R}^n$ *and* $m\in\mathbb{N}_0$.

Lemma 20 *Let* $r>0$, $\nu\geq 0$ *and* $m>n$. *Then there exists* $c>0$, *which depends only on* m,n *and* r, *such that for all* $g\in S'(\mathbb{R}^n)$ *with* $\operatorname{supp}\hat{g}\subset\{\xi\in\mathbb{R}^n:|\xi|\leq 2^{\nu+1}\}$, *we have*

$$\big|g(x)\big|\leq c\big(\eta_{\nu,m}*|g|^r(x)\big)^{1/r},\quad x\in\mathbb{R}^n.$$

The following lemma is the counterpart to Lemma 10 for Triebel-Lizorkin spaces.

Lemma 21 ([16], Theorem 3.2) *Let* $p,q\in\mathcal{P}^{\log}(\mathbb{R}^n)$ *with* $1<p^-\leq p^+<\infty$ *and* $1<q^-\leq q^+<\infty$. *Then the inequality*

$$\big\|(\eta_{\nu,m}*f)^{\infty}_{\nu=0}|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|\leq c\big\|(f_\nu)^{\infty}_{\nu=0}|L_{p(\cdot)}(\ell_{q(\cdot)})\big\|$$

holds for every sequence $(f_\nu)_{\nu\in\mathbb{N}_0}$ *of* $L^{loc}_1(\mathbb{R}^n)$ *functions and* $m>n$.

The following lemma is well known (cf. [51]). We sketch its proof for the sake of completeness.

Lemma 22 *Let $a, b > 0$, $M \in \mathbb{N}$ and $h \in \mathbb{R}^n$. Let $f \in S'(\mathbb{R}^n)$ with $\operatorname{supp} \hat{f} \subset \{\xi \in \mathbb{R}^n : |\xi| \le b\}$. Then there is a constant $C > 0$ independent of f, b and h, such that*

$$\left|\Delta_h^M f(x)\right| \le C \max\bigl(1, |bh|^a\bigr) \cdot \min\bigl(1, |bh|^M\bigr) P_{b,a} f(x)$$

holds for every $x \in \mathbb{R}^n$.

Proof The estimate

$$\begin{aligned}\left|f(x+jh)\right| &= \frac{|f(x+jh)|}{1+|jbh|^a} \cdot \bigl(1+|jbh|^a\bigr) \le \bigl(1+|Mbh|^a\bigr) \sup_{z\in\mathbb{R}^n} \frac{f(x-z)}{1+|bz|^a}\\ &\lesssim \max\bigl(1, |bh|^a\bigr) P_{b,a} f(x), \qquad j = 0, \dots, M,\end{aligned}$$

holds for all the admissible parameters even without the assumption on $\hat{f}$.

Hence we need to prove only

$$\left|\Delta_h^M f(x)\right| \le C \max\bigl(1, |bh|^a\bigr) \cdot |bh|^M \cdot P_{b,a} f(x). \tag{58}$$

Using the Taylor formula for the (analytic) function f, we obtain by direct calculation

$$\begin{aligned}\left|\Delta_h^M f(x)\right| &\le c|h|^M \sup_{|\alpha|=M} \sup_{|y|\le M|h|} \frac{|(D^\alpha f)(x-y)|}{1+|by|^a} \cdot \bigl(1+|by|^a\bigr)\\ &\le c'|h|^M \max\bigl(1, |bh|^a\bigr) \cdot \sup_{|\alpha|=M} \sup_{|y|\le M|h|} \frac{|(D^\alpha f)(x-y)|}{1+|by|^a}.\end{aligned}$$

If $\operatorname{supp} \hat{g} \subset \{\xi \in \mathbb{R}^n : |\xi| \le 1\}$, then this may be combined with the Nikol'skij inequality, cf. [51, Sect. 1.3.1], in the form

$$\sup_{|\alpha|=M} \sup_{z\in\mathbb{R}^n} \frac{|(D^\alpha g)(x-z)|}{1+|z|^a} \lesssim \sup_{z\in\mathbb{R}^n} \frac{|g(x-z)|}{1+|z|^a}$$

to obtain

$$\left|\Delta_h^M g(x)\right| \le c''|h|^M \max\bigl(1, |h|^a\bigr) \cdot \sup_{z\in\mathbb{R}^n} \frac{|g(x-z)|}{1+|z|^a}. \tag{59}$$

If $\operatorname{supp} \hat{f} \subset \{\xi \in \mathbb{R}^n : |\xi| \le b\}$, we define $g(x) = f(x/b)$, apply (59) together with $\Delta_h^M f(x) = \Delta_{bh}^M g(bx)$ and obtain

$$\left|\Delta_h^M f(x)\right| \lesssim |bh|^M \max\bigl(1, |bh|^a\bigr) \sup_{z\in\mathbb{R}^n} \frac{|g(bx-z)|}{1+|z|^a}.$$

From this (58) follows and the proof is then complete. □

The following lemma resembles Lemma A.3 of [16].

Lemma 23 *Let $k \in \mathbb{Z}$, $l \in \mathbb{N}_0$ and $m > n$. Then*

$$\eta_{k+l,m} * \bigl[2^{kn} \chi_{2^{-k}B}\bigr] \lesssim \eta_{k,m}.$$

Proof Using dilations, we may suppose that $k=0$. If $|x|\le 2$, then

$$\int_{\{y:|x-y|\le 1\}} 2^{nl}\big(1+2^l|y|\big)^{-m}dy \le \int_{y\in\mathbb{R}^n} 2^{nl}\big(1+2^l|y|\big)^{-m}dy \lesssim \big(1+|x|\big)^{-m}.$$

If $|x|>2$ and $|x-y|\le 1$, we obtain $1+2^l|y|\gtrsim 1+2^l|x|$ and $2^{nl}(1+2^l|x|)^{-m}\lesssim (1+|x|)^{-m}$. This immediately implies that

$$\int_{\{y:|x-y|\le 1\}} 2^{nl}\big(1+2^l|y|\big)^{-m}dy \lesssim \int_{\{y:|x-y|\le 1\}} \big(1+|x|\big)^{-m}dy \lesssim \big(1+|x|\big)^{-m}.$$

□

Remark 11 Another way, how to prove Lemma 23 is to use the inequality $\chi_B(x)\le 2^m\eta_{0,m}(x)$ and apply Lemma A.3 of [16].

The following Lemma is quite simple and we leave out its proof.

Lemma 24 *Let $0<q<\infty$, $\delta>0$ and let $(a_l)_{l\in\mathbb{N}}$ be a sequence of non-negative real numbers. Then*

$$\left(\sum_{l=1}^{\infty} 2^{-l\delta q}a_l\right)^{1/q} \lesssim \sum_{l=1}^{\infty} 2^{-l\delta/2}a_l^{1/q},$$

where the constant involved depends only on δ and q.

Finally, we shall need a certain version of Hölder's inequality for $\ell_{q(\cdot)}(L_{p(\cdot)})$ spaces.

Lemma 25 *Let $p,q\in\mathcal{P}(\mathbb{R}^n)$ and let $0<\lambda<1$. Then*

$$\big\| f_k\cdot g_k|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| \le 2^{1/q^-}\big\| f_k^{1/(1-\lambda)}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|^{1-\lambda}\cdot\big\| g_k^{1/\lambda}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\|^{\lambda} \tag{60}$$

holds for all sequences of non-negative functions $(f_k)_{k\in\mathbb{N}_0}$ and $(g_k)_{k\in\mathbb{N}_0}$.

Proof Due to the homogeneity, we may assume that

$$\big\| f_k^{1/(1-\lambda)}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| = \big\| g_k^{1/\lambda}|\ell_{q(\cdot)}(L_{p(\cdot)})\big\| = 1.$$

Then for every $\varepsilon>0$, there exist two sequences of positive real numbers $(\lambda_k)_{k\in\mathbb{N}_0}$ and $(\mu_k)_{k\in\mathbb{N}_0}$, such that

$$\sum_{k=0}^{\infty}\lambda_k<1+\varepsilon,\qquad \sum_{k=0}^{\infty}\mu_k<1+\varepsilon$$

and

$$\varrho_{p(\cdot)}\left(\frac{f_k^{1/(1-\lambda)}}{\lambda_k^{1/q(\cdot)}}\right)\le 1,\qquad \varrho_{p(\cdot)}\left(\frac{g_k^{1/\lambda}}{\mu_k^{1/q(\cdot)}}\right)\le 1.$$

We put

$$c := 2^{1/q^-} \quad \text{and} \quad \gamma_k := \frac{\lambda_k + \mu_k}{2} \geq \frac{\lambda_k + \mu_k}{c^{q(x)}}$$

and use the Young inequality in the form

$$\big[f_k(x) g_k(x)\big]^{p(x)} \leq (1-\lambda) f_k(x)^{p(x)/(1-\lambda)} + \lambda g_k(x)^{p(x)/\lambda}$$

to obtain

$$\begin{aligned}
&\int_{\mathbb{R}^n} \left(\frac{f_k(x) g_k(x)}{c\gamma_k^{1/q(\cdot)}} \right)^{p(x)} dx \\
&\quad \leq (1-\lambda) \int_{\mathbb{R}^n} \frac{f_k(x)^{p(x)/(1-\lambda)}}{c^{p(x)} \gamma_k^{p(x)/q(x)}} dx + \lambda \int_{\mathbb{R}^n} \frac{g_k(x)^{p(x)/\lambda}}{c^{p(x)} \gamma_k^{p(x)/q(x)}} dx \\
&\quad \leq (1-\lambda) \int_{\mathbb{R}^n} \frac{f_k(x)^{p(x)/(1-\lambda)}}{\lambda_k^{p(x)/q(x)}} dx + \lambda \int_{\mathbb{R}^n} \frac{g_k(x)^{p(x)/\lambda}}{\mu_k^{p(x)/q(x)}} dx \leq 1.
\end{aligned}$$

Furthermore, the estimate

$$\sum_{k=0}^{\infty} \gamma_k < \frac{2(1+\varepsilon)}{2} = 1 + \varepsilon$$

finishes the proof of (60) with the constant $c = 2^{1/q^-}$. □

Acknowledgements The first author acknowledges the financial support provided by the DFG project HA 2794/5-1 “Wavelets and function spaces on domains”. Furthermore, the first author thanks the RICAM for his hospitality and support during a short term visit in Linz.

The second author acknowledges the financial support provided by the FWF project Y 432-N15 START-Preis “Sparse Approximation and Optimization in High Dimensions”.

We thank the anonymous referee for pointing the reference [18] out to us.

References

1. Almeida, A., Hästö, P.: Besov spaces with variable smoothness and integrability. J. Funct. Anal. **258**(5), 1628–1655 (2010)
2. Almeida, A., Samko, S.: Characterization of Riesz and Bessel potentials on variable Lebesgue spaces. J. Funct. Spaces Appl. **4**(2), 113–144 (2006)
3. Almeida, A., Samko, S.: Pointwise inequalities in variable Sobolev spaces and applications. Z. Anal. Anwend. **26**(2), 179–193 (2007)
4. Almeida, A., Samko, S.: Embeddings of variable Hajłasz-Sobolev spaces into Hölder spaces of variable order. J. Math. Anal. Appl. **353**(2), 489–496 (2009)
5. Andersson, P.: Two-microlocal spaces, local norms and weighted spaces. Paper 2 in PhD Thesis, pp. 35–58 (1997)
6. Aoki, T.: Locally bounded linear topological spaces. Proc. Imp. Acad. (Tokyo) **18**, 588–594 (1942)
7. Beauzamy, B.: Espaces de Sobolev et de Besov d’ordre variable définis sur L^p. C. R. Math. Acad. Sci. Paris **274**, 1935–1938 (1972)
8. Besov, O.V.: Equivalent normings of spaces of functions of variable smoothness. Proc. Steklov Inst. Math. **243**(4), 80–88 (2003)

9. Bony, J.-M.: Second microlocalization and propagation of singularities for semi-linear hyperbolic equations. In: Taniguchi Symp. HERT, Katata, pp. 11–49 (1984)
10. Cobos, F., Fernandez, D.L.: Hardy-Sobolev spaces and Besov spaces with a function parameter. In: Proc. Lund Conf. 1986. Lect. Notes Math., vol. 1302, pp. 158–170. Springer, Berlin (1986)
11. Cruz-Uribe, D., Fiorenza, A., Neugebauer, C.J.: The maximal function on variable L^p spaces. Ann. Acad. Sci. Fenn., Ser. A 1 Math. **28**, 223–238 (2003)
12. Cruz-Uribe, D., Fiorenza, A., Martell, J.M., Pérez, C.: The boundedness of classical operators in variable L^p-spaces. Ann. Acad. Sci. Fenn., Ser. A 1 Math. **31**, 239–264 (2006)
13. Diening, L.: private communication
14. Diening, L.: Maximal function on generalized Lebesgue spaces $L^{p(\cdot)}$. Math. Inequal. Appl. **7**(2), 245–254 (2004)
15. Diening, L., Harjulehto, P., Hästö, P., Mizuta, Y., Shimomura, T.: Maximal functions in variable exponent spaces: limiting cases of the exponent. Ann. Acad. Sci. Fenn., Ser. A 1 Math. **34**(2), 503–522 (2009)
16. Diening, L., Hästö, P., Roudenko, S.: Function spaces of variable smoothness and integrability. J. Funct. Anal. **256**(6), 1731–1768 (2009)
17. Diening, L., Harjulehto, P., Hästö, P., Růžička, M.: Lebesgue and Sobolev Spaces with Variable Exponents. Lecture Notes in Mathematics, vol. 2017. Springer Berlin (2011)
18. Drihem, D.: Atomic decomposition of Besov spaces with variable smoothness and integrability. J. Math. Anal. Appl. **389**, 15–31 (2012)
19. Farkas, W., Leopold, H.-G.: Characterisations of function spaces of generalised smoothness. Ann. Mat. Pura Appl. **185**(1), 1–62 (2006)
20. Goldman, M.L.: A description of the traces of some function spaces. Tr. Mat. Inst. Steklova **150**, 99–127 (1979). English transl.: Proc. Steklov Inst. Math. **150**(4) (1981)
21. Goldman, M.L.: A method of coverings for describing general spaces of Besov type. Tr. Mat. Inst. Steklova **156**, 47–81 (1980). English transl.: Proc. Steklov Inst. Math. **156**(2) (1983)
22. Goldman, M.L.: Imbedding theorems for anisotropic Nikol'skij-Besov spaces with moduli of continuity of general type. Tr. Mat. Inst. Steklova **170**, 86–104 (1984). English transl.: Proc. Steklov Inst. Math. **170**(1) (1987)
23. Gurka, P., Harjulehto, P., Nekvinda, A.: Bessel potential spaces with variable exponent. Math. Inequal. Appl. **10**(3), 661–676 (2007)
24. Jaffard, S.: Pointwise smoothness, two-microlocalisation and wavelet coefficients. Publ. Math. **35**, 155–168 (1991)
25. Jaffard, S., Meyer, Y.: Wavelet Methods for Pointwise Regularity and Local Oscillations of Functions. Memoirs of the AMS, vol. 123 (1996)
26. Kalyabin, G.A.: Characterization of spaces of generalized Liouville differentiation. Mat. Sb. Nov. Ser. **104**, 42–48 (1977)
27. Kalyabin, G.A.: Description of functions in classes of Besov-Lizorkin-Triebel type. Tr. Mat. Inst. Steklova **156**, 82–109 (1980). English transl.: Proc. Steklov Institut Math. **156**(2) (1983)
28. Kalyabin, G.A.: Characterization of spaces of Besov-Lizorkin and Triebel type by means of generalized differences. Tr. Mat. Inst. Steklova **181**, 95–116 (1988). English transl.: Proc. Steklov Inst. Math. **181**(4) (1989)
29. Kalyabin, G.A., Lizorkin, P.I.: Spaces of functions of generalized smoothness. Math. Nachr. **133**, 7–32 (1987)
30. Kempka, H.: 2-microlocal Besov and Triebel-Lizorkin spaces of variable integrability. Rev. Mat. Complut. **22**(1), 227–251 (2009)
31. Kempka, H., Vybíral, J.: A note on the spaces of variable integrability and summability of Almeida and Hästö. Proc. Am. Math. Soc. (to appear)
32. Kováčik, O., Rákosník, J.: On spaces $L^{p(x)}$ and $W^{1,p(x)}$. Czechoslov. Math. J. **41**(4), 592–618 (1991)
33. Leopold, H.-G.: On function spaces of variable order of differentiation. Forum Math. **3**, 1–21 (1991)
34. Lévy Véhel, J., Seuret, S.: A time domain characterization of 2-microlocal spaces. J. Fourier Anal. Appl. **9**(5), 473–495 (2003)
35. Lévy Véhel, J., Seuret, S.: The 2-microlocal formalism. In: Fractal Geometry and Applications: A Jubilee of Benoit Mandelbrot, Proceedings of Symposia in Pure Mathematics, PSPUM, vol. 72, pp. 153–215 (2004). Part 2
36. Merucci, C.: Applications of interpolation with a function parameter to Lorentz Sobolev and Besov spaces. In: Proc. Lund Conf., 1983. Lect. Notes Math., vol. 1070, pp. 183–201. Springer, Berlin (1983)

37. Meyer, Y.: Wavelets, Vibrations and Scalings. CRM Monograph Series, vol. 9. AMS, Providence (1998)
38. Moritoh, S., Yamada, T.: Two-microlocal Besov spaces and wavelets. Rev. Mat. Iberoam. **20**, 277–283 (2004)
39. Moura, S.: Function spaces of generalised smoothness. Diss. Math. **398**, 1–87 (2001)
40. Nekvinda, A.: Hardy-Littlewood maximal operator on $L^{p(x)}(\mathbb{R}^n)$. Math. Inequal. Appl. **7**(2), 255–266 (2004)
41. Orlicz, W.: Über konjugierte Exponentenfolgen. Stud. Math. **3**, 200–212 (1931)
42. Peetre, J.: On spaces of Triebel-Lizorkin type. Ark. Math. **13**, 123–130 (1975)
43. Rolewicz, S.: On a certain class of linear metric spaces. Bull. Acad. Pol. Sci., Sér. Sci. Math. Astron. Phys. **5**, 471–473 (1957)
44. Ross, B., Samko, S.: Fractional integration operator of variable order in the spaces H^λ. Int. J. Math. Sci. **18**(4), 777–788 (1995)
45. Růžička, M.: Electrorheological Fluids: Modeling and Mathematical Theory. Lecture Notes in Mathematics, vol. 1748. Springer, Berlin (2000)
46. Rychkov, V.S.: On a theorem of Bui, Paluszynski and Taibleson. Proc. Steklov Inst. Math. **227**, 280–292 (1999)
47. Scharf, B.: Atomare Charakterisierungen vektorwertiger Funktionenräume. Diploma Thesis, Jena (2009)
48. Schneider, J.: Function spaces of varying smoothness I. Math. Nachr. **280**(16), 1801–1826 (2007)
49. Schneider, C.: On dilation operators in Besov spaces. Rev. Mat. Complut. **22**(1), 111–128 (2009)
50. Schneider, C., Vybíral, J.: On dilation operators in Triebel-Lizorkin spaces. Funct. Approx. Comment. Math. **41**, 139–162 (2009). Part 2
51. Triebel, H.: Theory of Function Spaces. Birkhäuser, Basel (1983)
52. Triebel, H.: Theory of Function Spaces II. Birkhäuser, Basel (1992)
53. Ullrich, T.: Function spaces with dominating mixed smoothness, characterization by differences. Technical report, Jenaer Schriften zur Math. und Inform., Math/Inf/05/06 (2006)
54. Ullrich, T.: Continuous characterizations of Besov-Lizorkin-Triebel spaces and new interpretations as coorbits. J. Funct. Spaces Appl. (2012). doi:10.1115/2012/163213. Article ID 163213, 47 pages
55. Ullrich, T., Rauhut, H.: Generalized coorbit space theory and inhomogeneous function spaces of Besov-Lizorkin-Triebel type. J. Funct. Anal. **260**(11), 3299–3362 (2011). doi:10.1016/j.jfa.2010.12.006
56. Unterberger, A., Bokobza, J.: Les opérateurs pseudodifférentiels d'ordre variable. C. R. Math. Acad. Sci. Paris **261**, 2271–2273 (1965)
57. Unterberger, A.: Sobolev spaces of variable order and problems of convexity for partial differential operators with constant coefficients. In: Astérisque 2 et 3, pp. 325–341. Soc. Math. France, Paris (1973)
58. Višik, M.I., Eskin, G.I.: Convolution equations of variable order (russ.). Tr. Mosk. Mat. Obsc. **16**, 26–49 (1967)
59. Vybíral, J.: Sobolev and Jawerth embeddings for spaces with variable smoothness and integrability. Ann. Acad. Sci. Fenn., Ser. A 1 Math. **34**(2), 529–544 (2009)
60. Xu, H.: Généralisation de la théorie des chirps à divers cadres fonctionnels et application à leur analyse par ondelettes. Ph.D. thesis, Université Paris IX Dauphine (1996)
61. Xu, J.-S.: Variable Besov and Triebel-Lizorkin spaces. Ann. Acad. Sci. Fenn., Ser. A 1 Math. **33**(2), 511–522 (2008)
62. Xu, J.-S.: An atomic decomposition of variable Besov and Triebel-Lizorkin spaces. Armenian J. Math. **2**(1), 1–12 (2009)
63. Yuan, W., Sickel, W., Yang, D.: Morrey and Campanato Meet Besov, Lizorkin and Triebel. Lecture Notes in Mathematics, vol. 2005. Springer, Berlin (2010)

Chapter 1
A Survey of Compressed Sensing

Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral

Abstract Compressed sensing was introduced some ten years ago as an effective way of acquiring signals, which possess a sparse or nearly sparse representation in a suitable basis or dictionary. Due to its solid mathematical backgrounds, it quickly attracted the attention of mathematicians from several different areas, so that the most important aspects of the theory are nowadays very well understood. In recent years, its applications started to spread out through applied mathematics, signal processing, and electrical engineering. The aim of this chapter is to provide an introduction into the basic concepts of compressed sensing. In the first part of this chapter, we present the basic mathematical concepts of compressed sensing, including the Null Space Property, Restricted Isometry Property, their connection to basis pursuit and sparse recovery, and construction of matrices with small restricted isometry constants. This presentation is easily accessible, largely self-contained, and includes proofs of the most important theorems. The second part gives an overview of the most important extensions of these ideas, including recovery of vectors with sparse representation in frames and dictionaries, discussion of (in)coherence and its implications for compressed sensing, and presentation of other algorithms of sparse recovery.

H. Boche (✉)
Technische Universität München, Theresienstr. 90/IV, München, Germany
e-mail: boche@tum.de

R. Calderbank
Duke University, 317 Gross Hall, Durham NC, USA
e-mail: robert.calderbank@duke.edu

G. Kutyniok
Technische Universität Berlin, Straße des 17. Juni 136, Berlin, Germany
e-mail: kutyniok@math.tu-berlin.de

J. Vybíral
Faculty of Mathematics and Physics, Charles University, Sokolovska 83,
186 00 Prague 8, Czech Republic
e-mail: vybiral@karlin.mff.cuni.cz

H. Boche et al. (eds.), *Compressed Sensing and its Applications*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-16042-9_1

1.1 Introduction

Compressed sensing is a novel method of signal processing, which was introduced in [25] and [15, 16] and which profited from its very beginning from fruitful interplay between mathematicians, applied mathematicians, and electrical engineers. The mathematical concepts are inspired by ideas from a number of different disciplines, including numerical analysis, stochastic, combinatorics, and functional analysis. On the other hand, the applications of compressed sensing range from image processing [29], medical imaging [52], and radar technology [5] to sampling theory [56, 69], and statistical learning.

The aim of this chapter is twofold. In Section 1.3 we collect the basic mathematical ideas from numerical analysis, stochastic, and functional analysis used in the area of compressed sensing to give an overview of basic notions, including the Null Space Property and the Restricted Isometry Property, and the relations between them. Most of the material in this section is presented with a self-contained proof, using only few simple notions from approximation theory and stochastic recalled in Section 1.2. We hope that this presentation will make the mathematical concepts of compressed sensing appealing and understandable both to applied mathematicians and electrical engineers. Although it can also be used as a basis for a lecture on compressed sensing for a wide variety of students, depending on circumstances, it would have to be complemented by other subjects of the lecturers choice to make a full one-semester course. Let us stress that the material presented in this section is by no means new or original, actually it is nowadays considered classical, or "common wisdom" throughout the community.

The second aim of this chapter is to give (without proof) an overview of the most important extensions (Section 1.4). In this part, we refer to original research papers or to more extensive summaries of compressed sensing [23, 35, 40] for more details and further references.

1.2 Preliminaries

As the mathematical concepts of compressed sensing rely on the interplay of ideas from linear algebra, numerical analysis, stochastic, and functional analysis, we start with an overview of basic notions from these fields. We shall restrict ourselves to the minimum needed in the sequel.

1.2.1 Norms and quasi-norms

In the most simple setting of discrete signals on finite domain, signals are modeled as (column) vectors in then n-dimensional Euclidean space, denoted by $\mathbb{R}^n$. We shall

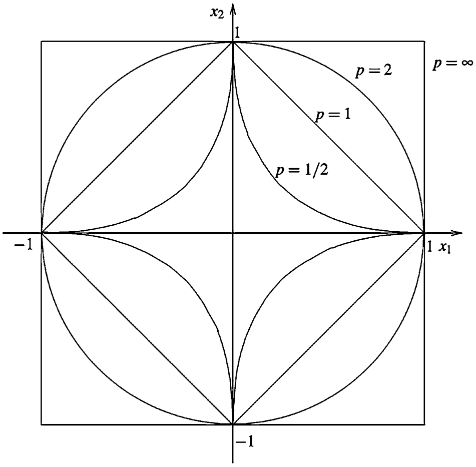


Fig. 1.1 Shape of the l_p^2 unit ball for $p = 1/2, p = 1, p = 2$, and $p = \infty$

use different ways how to measure the size of such a vector. The most typical way, however, is to consider its ℓ_p^n-norm, which is defined for $x = (x_1, \dots, x_n)^T$ and $p \in (0, \infty]$ as (Fig. 1.1)

$$\|x\|_p = \begin{cases} \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}, & p \in (0,\infty); \\ \max_{j=1,\dots,n} |x_j|, & p = \infty. \end{cases} \tag{1.1}$$

If $p < 1$, this expression does not satisfy the triangle inequality. Instead of that the following inequalities hold

$$\|x+z\|_p \le 2^{1/p-1}\big(\|x\|_p + \|z\|_p\big),$$
$$\|x+z\|_p^p \le \|x\|_p^p + \|z\|_p^p$$

for all $x \in \mathbb{R}^n$ and all $z \in \mathbb{R}^n$. If $p = 2$, ℓ_2^n is a (real) Hilbert space with the scalar product

$$\langle x, z\rangle = z^T x = \sum_{i=j}^n x_j z_j.$$

If $x \in \mathbb{R}^n$, we can always find a permutation $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$, such that the nonincreasing rearrangement $x^* \in [0, \infty)^n$ of x, defined by $x_j^* = |x_{\sigma(j)}|$ satisfies

$$x_1^* \ge x_2^* \ge \dots \ge x_n^* \ge 0.$$

If $T \subset \{1,\dots,n\}$ is a set of indices, we denote by $|T|$ the number of its elements. We shall complement this notation by denoting the size of the support of $x \in \mathbb{R}^n$ by

$$\|x\|_0 = |\mathrm{supp}\,(x)| = |\{j : x_j \neq 0\}|.$$

Note that this expression is not even a quasinorm. The notation is justified by the observation, that

$$\lim_{p\to 0} \|x\|_p^p = \|x\|_0 \quad \text{for all} \quad x \in \mathbb{R}^n.$$

Let k be a natural number at most equal to n. A vector $x \in \mathbb{R}^n$ is called k-sparse, if $\|x\|_0 \le k$ and the set of all k-sparse vectors is denoted by

$$\Sigma_k = \{x \in \mathbb{R}^n : \|x\|_0 \le k\}.$$

Finally, if $k < n$, the best k-term approximation $\sigma_k(x)_p$ of $x \in \mathbb{R}^n$ describes, how well can x be approximated by k-sparse vectors in the ℓ_p^n-norm. This can be expressed by the formula

$$\sigma_k(x)_p = \inf_{z\in\Sigma_k} \|x-z\|_p = \begin{cases} \left(\sum\limits_{j=k+1}^{n} (x_j^*)^p\right)^{1/p}, & p \in (0,\infty); \\ x_{k+1}^*, & p = \infty. \end{cases} \tag{1.2}$$

The notions introduced so far can be easily transferred to n-dimensional complex spaces. Especially, the scalar product of $x, y \in \mathbb{C}^n$ is defined by

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j \overline{y_j},$$

where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$.

Linear operators between finite-dimensional spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ can be represented with the help of matrices $A \in \mathbb{R}^{m\times n}$. The entries of A are denoted by a_{ij}, $i = 1,\dots,m$ and $j = 1,\dots,n$. The transpose of a matrix $A \in \mathbb{R}^{m\times n}$ is a matrix $A^T \in \mathbb{R}^{n\times m}$ with entries $(A^T)_{ij} = a_{ji}$. The identity matrix in $\mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$ will be denoted by I.

1.2.2 Random Variables

As several important constructions from the field of compressed sensing rely on randomness, we recall the basic notions from probability theory.

We denote by $(\Omega,\Sigma,\mathbb{P})$ a probability space. Here stands Ω for the sample space, Σ for a σ-algebra of subsets of Ω, and $\mathbb{P}$ is a probability measure on (Ω,Σ). The sets $B\in\Sigma$ are called events, and their probability is denoted by

$$\mathbb{P}(B)=\int_B d\,\mathbb{P}(\omega).$$

A random variable X is a measurable function $X:\Omega\to\mathbb{R}$ and we denote by

$$\mu=\mathbb{E}X=\int_\Omega X(\omega)d\mathbb{P}(\omega)$$

its expected value, or mean, and by $\sigma^2=\mathbb{E}[(X-\mu)^2]=\mathbb{E}(X^2)-(\mathbb{E}X)^2$ its variance. We recall Markov's inequality, which states

$$\mathbb{P}(|X|\geq t)\leq\frac{\mathbb{E}|X|}{t}\quad\text{for all } t>0. \tag{1.3}$$

A random variable X is called *normal* (or *Gaussian*), if it has a density function

$$f(t)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big),\quad t\in\mathbb{R}$$

for some real μ and positive σ^2, i.e. if $\mathbb{P}(a<X\leq b)=\int_a^b f(t)dt$ for all real $a<b$. In that case, the expected value of X is equal to μ and its variance to σ^2 and we often write $X\sim\mathcal{N}(\mu,\sigma^2)$. If $\mu=0$ and $\sigma^2=1$, the normal variable is called *standard* and its density function is

$$f(t)=\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{t^2}{2}\Big),\quad t\in\mathbb{R}.$$

A random variable X is called *Rademacher* if

$$\mathbb{P}(X=1)=\mathbb{P}(X=-1)=1/2. \tag{1.4}$$

Random variables $X_1,\dots,X_N$ are called *independent*, if for every real $t_1,\dots,t_N$ the following formula holds

$$\mathbb{P}(X_1\leq t_1,\dots,X_N\leq t_N)=\prod_{j=1}^N\mathbb{P}(X_j\leq t_j).$$

In that case,

$$\mathbb{E}\Big[\prod_{j=1}^N X_j\Big]=\prod_{j=1}^N\mathbb{E}(X_j). \tag{1.5}$$

If the random variables $X_1,\dots,X_N$ are independent and have the same distribution, we call them *independent identically distributed*, which is usually abbreviated as *i.i.d.*

1.3 Basic ideas of compressed sensing

There is a number of ways how to discover the landscape of compressed sensing. The point of view, which we shall follow in this section, is that we are looking for sparse solutions $x\in\mathbb{R}^n$ of a system of linear equations $Ax=y$, where $y\in\mathbb{R}^m$ and the $m\times n$ matrix A are known. We shall be interested in underdetermined systems, i.e. in the case $m\le n$. Intuitively, this corresponds to solving the following optimization problem

$$\min_z \|z\|_0 \quad \text{subject to} \quad y=Az. \tag{P_0}$$

We will first show that this problem is numerically intractable if m and n are getting larger. Then we introduce the basic notions of compressed sensing, showing that for specific matrices A and measurement vectors y, one can recover the solution of (P_0) in a much more effective way.

1.3.1 Basis pursuit

The minimization problem (P_0) can obviously be solved by considering first all index sets $T\subset\{1,\dots,n\}$ with one element and employing the methods of linear algebra to decide if there is a solution x to the system with support included in T. If this fails for all such index sets, we continue with all index sets with two, three, and more elements. The obvious drawback is the rapidly increasing number of these index sets. Indeed, there is $\binom{n}{k}$ index sets $T\subset\{1,\dots,n\}$ with k elements and this quantity grows (in some sense) exponentially with k and n.

We shall start our tour through compressed sensing by showing that even every other algorithm solving (P_0) suffers from this drawback. This will be formulated in the language of complexity theory as the statement, that the (P_0) problem is NP-hard. Before we come to that, we introduce the basic terms used in the sequel. We refer for example to [2] for an introduction to computational complexity.

The *P-class* ("polynomial time") consists of all decision problems that can be solved in polynomial time, i.e. with an algorithm, whose running time is bounded from above by a polynomial expression in the size of the input.

The *NP-class* ("nondeterministic polynomial time") consists of all decision problems, for which there is a polynomial-time algorithm V (called verifier), with the following property. If, given an input α, the right answer to the decision problem is “yes”, then there is a proof β, such that $V(\alpha,\beta) = \text{yes}$. Roughly speaking, when the answer to the decision problem is positive, then the proof of this statement can be verified with a polynomial-time algorithm.

Let us reformulate (P_0) as a decision problem. Namely, if the natural numbers k,m,n, $m\times n$ matrix A and $y\in\mathbb{R}^m$ are given, decide if there is a k-sparse solution x of the equation $Ax=y$. It is easy to see that this version of (P_0) is in the NP-class. Indeed, if the answer to the problem is “yes” and a certificate $x\in\mathbb{R}^n$ is given, then it can be verified in polynomial time if x is k-sparse and $Ax=y$.

A problem is called *NP-hard* if any of its solving algorithms can be transformed in polynomial time into a solving algorithm of any other NP-problem. We shall rely on a statement from complexity theory, that the following problem is both NP and NP-hard.

Exact cover problem
Given as the input a natural number m divisible by 3 and a system $\{T_j : j = 1,\dots,n\}$ of subsets of $\{1,\dots,m\}$ with $|T_j| = 3$ for all $j = 1,\dots,n$, decide, if there is a subsystem of mutually disjoint sets $\{T_j : j\in J\}$, such that $\bigcup_{j\in J} T_j = \{1,\dots,m\}$. Such a subsystem is frequently referred to as *exact cover*.

Let us observe that for any subsystem $\{T_j : j\in J\}$ it is easy to verify (in polynomial time) if it is an exact cover or not. So the problem is in the NP-class. The non-trivial statement from computational complexity is that this problem is also NP-hard. The exact formulation of (P_0) looks as follows.

ℓ_0-minimization problem
Given natural numbers m,n, an $m\times n$ matrix A and a vector $y\in\mathbb{R}^m$ as input, find the solution of

$$\min_z \|z\|_0 \quad \text{s.t.} \quad y = Az.$$

Theorem 1. *The ℓ_0-minimization problem is NP-hard.*

Proof. It is sufficient to show that any algorithm solving the ℓ_0-minimization problem can be transferred in polynomial time into an algorithm solving the exact cover problem. Let therefore $\{T_j : j = 1,\dots,n\}$ be a system of subsets of $\{1,\dots,m\}$ with $|T_j| = 3$ for all $j = 1,\dots,n$. Then we construct a matrix $A\in\mathbb{R}^{m\times n}$ by putting

$$a_{ij} := \begin{cases} 1 & \text{if } i\in T_j, \\ 0 & \text{if } i\not\in T_j, \end{cases}$$

i.e. the jth column of A is the indicator function of T_j (denoted by $\chi_{T_j} \in \{0,1\}^m$) and

$$Ax = \sum_{j=1}^{n} x_j \chi_{T_j}. \tag{1.6}$$

The construction of A can of course be done in polynomial time.

Let now x be the solution to the ℓ_0-minimization problem with the matrix A and the vector $y = (1,\dots,1)^T$. It follows by (1.6) that $m = \|y\|_0 = \|Ax\|_0 \le 3\|x\|_0$, i.e. that $\|x\|_0 \ge m/3$. We will show that the exact cover problem has a positive solution if, and only if, $\|x\|_0 = m/3$.

Indeed, if the exact cover problem has a positive solution, then there is a set $J \subset \{1,\dots,n\}$ with $|J| = m/3$ and

$$\chi_{\{1,\dots,m\}} = \sum_{j \in J} \chi_{T_j}.$$

Hence $y = Ax$ for $x = \chi_J$ and $\|x\|_0 = |J| = m/3$. If, on the other hand, $y = Ax$ and $\|x\|_0 = m/3$, then $\{T_j : j \in \operatorname{supp}(x)\}$ solves the exact cover problem. □

The ℓ_0-minimization problem is NP-hard, if all matrices A and all measurement vectors y are allowed as inputs. The theory of compressed sensing shows nevertheless that for special matrices A and for $y = Ax$ for some sparse x, the problem can be solved efficiently.

In general, we replace the $\|z\|_0$ in (P_0) by some $\|z\|_p$ for $p > 0$. To obtain a convex problem, we need to have $p \ge 1$. To obtain sparse solutions, $p \le 1$ is necessary, cf. Figure 1.2.

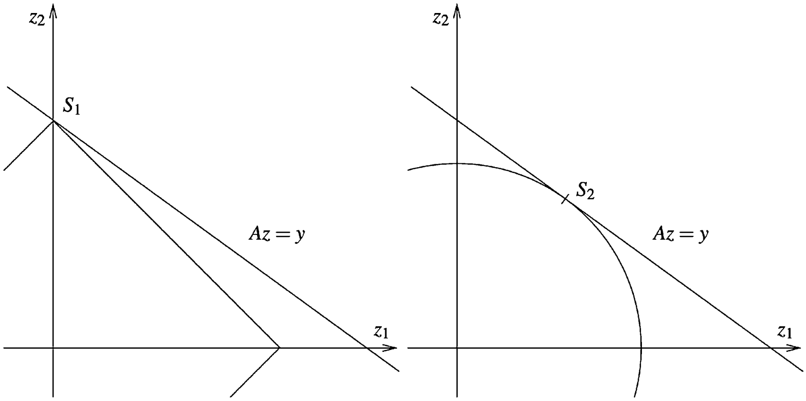


Fig. 1.2 Solution of $S_p = \underset{z \in \mathbb{R}^2}{\operatorname{argmin}} \|z\|_p \quad \text{s.t.} \quad y = Az$ for $p = 1$ and $p = 2$

We are therefore naturally led to discuss under which conditions the solution to (P_0) coincides with the solution of the following convex optimization problem called *basis pursuit*

$$\min_z \|z\|_1 \quad \text{s.t.} \quad y = Az, \tag{P_1}$$

which was introduced in [19]. But before we come to that, let us show that in the real case this problem may be reformulated as a linear optimization problem, i.e. as the search for the minimizer of a linear function over a set given by linear constraints, whose number depends polynomially on the dimension. We refer to [42] for an introduction to linear programming.

Indeed, let us assume that (P_1) has a unique solution, which we denote by $x \in \mathbb{R}^n$. Then the pair (u,v) with $u = x^+$ and $v = x^-$, i.e. with

$$u_j = \begin{cases} x_j, & x_j \geq 0, \\ 0, & x_j < 0, \end{cases} \quad \text{and} \quad v_j = \begin{cases} 0, & x_j \geq 0, \\ -x_j, & x_j < 0, \end{cases}$$

is the unique solution of

$$\min_{u,v \in \mathbb{R}^n} \sum_{j=1}^n (u_j + v_j) \text{ s.t. } Au - Av = y \text{ and } u_j \geq 0 \text{ and } v_j \geq 0 \text{ for all } j = 1, \dots, n. \tag{1.7}$$

If namely (u', v') is another pair of vectors admissible in (1.7), then $x' = u' - v'$ satisfies $Ax' = y$ and x' is therefore admissible in (P_1). As x is the solution of (P_1), we get

$$\sum_{j=1}^n (u_j + v_j) = \|x\|_1 < \|x'\|_1 = \sum_{j=1}^n |u_j' - v_j'| \leq \sum_{j=1}^n (u_j' + v_j').$$

If, on the other hand, the pair (u,v) is the unique solution of (1.7), then $x = u - v$ is the unique solution of (P_1). If namely z is another admissible vector in (P_1), then $u' = z^+$ and $v' = z^-$ are admissible in (1.7) and we obtain

$$\|x\|_1 = \sum_{j=1}^n |u_j - v_j| \leq \sum_{j=1}^n (u_j + v_j) < \sum_{j=1}^n (u_j' + v_j') = \|z\|_1.$$

Very similar argument works also in the case when (P_1) has multiple solutions.

1.3.2 Null Space Property

If $T \subset \{1,\dots,n\}$, then we denote by $T^c = \{1,\dots,n\} \setminus T$ the complement of T in $\{1,\dots,n\}$. If furthermore $v \in \mathbb{R}^n$, then we denote by v_T either the vector in $\mathbb{R}^{|T|}$, which contains the coordinates of v on T, or the vector in $\mathbb{R}^n$, which equals v on T and is zero on T^c. It will be always clear from the context, which notation is being used.

Finally, if $A \in \mathbb{R}^{m\times n}$ is a matrix, we denote by A_T the $m \times |T|$ sub-matrix containing the columns of A indexed by T. Let us observe that if $x \in \mathbb{R}^n$ with $T = \text{supp}\,(x)$, that $Ax = A_T x_T$.

We start the discussion of the properties of basis pursuit by introducing the notion of Null Space Property, which first appeared in [20].

Definition 1. Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then A is said to have the *Null Space Property* (NSP) of order k if

$$\|v_T\|_1 < \|v_{T^c}\|_1 \quad \text{for all } v \in \ker A \setminus \{0\} \text{ and all } T \subset \{1,\dots,n\} \text{ with } |T| \le k. \tag{1.8}$$

Remark 1. (i) The condition (1.8) states that vectors from the kernel of A are well spread, i.e. not supported on a set of small size. Indeed, if $v \in \mathbb{R}^n \setminus \{0\}$ is k-sparse and $T = \text{supp}\,(v)$, then (1.8) shows immediately, that v cannot lie in the kernel of A.

(ii) If we add $\|v_{T^c}\|_1$ to both sides of (1.8), we obtain $\|v\|_1 < 2\|v_{T^c}\|_1$. If then T are the indices of the k largest coordinates of v taken in the absolute value, this inequality becomes $\|v\|_1 < 2\sigma_k(v)_1$.

Theorem 2. *Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$ if, and only if, A has the NSP of order k.*

Proof. Let us assume that every k-sparse vector x is the unique solution of (P_1) with $y = Ax$. Let $v \in \ker A \setminus \{0\}$ and let $T \subset \{1,\dots,n\}$ with $|T| \le k$ be arbitrary. Then v_T is k-sparse, and is therefore the unique solution of

$$\min_z \|z\|_1, \quad \text{s.t.} \quad Az = Av_T. \tag{1.9}$$

As $A(-v_{T^c}) = A(v - v_{T^c}) = A(v_T)$, this gives especially $\|v_T\|_1 < \|v_{T^c}\|_1$ and A has the NSP of order k.

Let us, on the other hand, assume that A has the NSP of order k. Let $x \in \mathbb{R}^n$ be a k-sparse vector and let $T = \text{supp}\,(x)$. We have to show that $\|x\|_1 < \|z\|_1$ for every $z \in \mathbb{R}^n$ different from x with $Az = Ax$. But this follows easily by using (1.8) for the vector $(x - z) \in \ker A \setminus \{0\}$

$$\begin{aligned}\|x\|_1 &\le \|x - z_T\|_1 + \|z_T\|_1 = \|(x-z)_T\|_1 + \|z_T\|_1 < \|(x-z)_{T^c}\|_1 + \|z_T\|_1\\ &= \|z_{T^c}\|_1 + \|z_T\|_1 = \|z\|_1.\end{aligned}$$

□

Remark 2. Theorem 2 states that the solutions of (P_0) may be found by (P_1), if A has the NSP of order k and if $y \in \mathbb{R}^m$ is such that, there exists a k-sparse solution x of the equation $Ax = y$. Indeed, if in such a case, $\hat{x}$ is a solution of (P_0), then $\|\hat{x}\|_0 \leq \|x\|_0 \leq k$. Finally, it follows by Theorem 2 that $\hat{x}$ is also a solution of (P_1) and that $x = \hat{x}$.

In the language of complexity theory, if we restrict the inputs of the ℓ_0-minimization problem to matrices with the NSP of order k and to vectors y, for which there is a k-sparse solution of the equation $Ax = y$, the problem belongs to the P-class and the solving algorithm with polynomial running time is any standard algorithm solving (P_1), or the corresponding linear problem (1.7).

1.3.3 Restricted Isometry Property

Although the Null Space Property is equivalent to the recovery of sparse solutions of underdetermined linear systems by basis pursuit in the sense just described, it is somehow difficult to construct matrices satisfying this property. We shall therefore present a sufficient condition called Restricted Isometry Property, which was first introduced in [15], and which ensures that the Null Space Property is satisfied.

Definition 2. Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then the *restricted isometry constant* $\delta_k = \delta_k(A)$ of A of order k is the smallest $\delta \geq 0$, such that

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \quad \text{for all} \quad x \in \Sigma_k. \tag{1.10}$$

Furthermore, we say that A satisfies the *Restricted Isometry Property* (RIP) of order k with the constant δ_k if $\delta_k < 1$.

Remark 3. The condition (1.10) states that A acts nearly isometrically when restricted to vectors from Σ_k. Of course, the smaller the constant $\delta_k(A)$ is, the closer is the matrix A to isometry on Σ_k. We will be therefore later interested in constructing matrices with small RIP constants. Finally, the inequality $\delta_1(A) \leq \delta_2(A) \leq \dots \leq \delta_k(A)$ follows trivially.

The following theorem shows that RIP of sufficiently high order with a constant small enough is indeed a sufficient condition for NSP.

Theorem 3. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then A has the NSP of order k.*

Proof. Let $v \in \ker A$ and let $T \subset \{1,\dots,n\}$ with $|T| \leq k$. We shall show that

$$\|v_T\|_2 \leq \frac{\delta_{2k}}{1-\delta_k} \cdot \frac{\|v\|_1}{\sqrt{k}}. \tag{1.11}$$

If $\delta_k \leq \delta_{2k} < 1/3$, then Hölder's inequality gives immediately $\|v_T\|_1 \leq \sqrt{k}\|v_T\|_2 < \|v\|_1/2$ and the NSP of A of order k follows.

Before we come to the proof of (1.11), let us make the following observation. If $x,z\in\Sigma_k$ are two vectors with disjoint supports and $\|x\|_2=\|z\|_2=1$, then $x\pm z\in\Sigma_{2k}$ and $\|x\pm z\|_2^2=2$. If we now combine the RIP of A

$$2(1-\delta_{2k})\le\|A(x\pm z)\|_2^2\le 2(1+\delta_{2k})$$

with the polarization identity, we get

$$|\langle Ax,Az\rangle|=\frac{1}{4}\Big|\|Ax+Az\|_2^2-\|Ax-Az\|_2^2\Big|\le\delta_{2k}.$$

Using this formula for $x'=x/\|x\|_2$ and $z'=z/\|z\|_2$, we see that if A has the RIP of order $2k$ and $x,z\in\Sigma_k$ have disjoint supports, then

$$|\langle Ax,Az\rangle|\le\delta_{2k}\|x\|_2\|z\|_2. \tag{1.12}$$

To show (1.11), let us assume that $v\in\ker A$ is fixed. It is enough to consider $T=T_0$ the set of the k largest entries of v taken in the absolute value. Furthermore, we denote by T_1 the set of k largest entries of $v_{T_0^c}$ in the absolute value, by T_2 the set of k largest entries of $v_{(T_0\cup T_1)^c}$ in the absolute value, etc. Using $0=Av=A(v_{T_0}+v_{T_1}+v_{T_2}+\dots)$ and (1.12), we arrive at

$$\begin{aligned}\|v_{T_0}\|_2^2&\le\frac{1}{1-\delta_k}\|Av_{T_0}\|_2^2=\frac{1}{1-\delta_k}\langle Av_{T_0},A(-v_{T_1})+A(-v_{T_2})+\dots\rangle\\&\le\frac{1}{1-\delta_k}\sum_{j\ge1}|\langle Av_{T_0},Av_{T_j}\rangle|\le\frac{\delta_{2k}}{1-\delta_k}\sum_{j\ge1}\|v_{T_0}\|_2\cdot\|v_{T_j}\|_2.\end{aligned}$$

We divide this inequality by $\|v_{T_0}\|_2\neq0$ and obtain

$$\|v_{T_0}\|_2\le\frac{\delta_{2k}}{1-\delta_k}\sum_{j\ge1}\|v_{T_j}\|_2.$$

The proof is then completed by the following simple chain of inequalities, which involve only the definition of the sets T_j, $j\ge0$.

$$\begin{aligned}\sum_{j\ge1}\|v_{T_j}\|_2&=\sum_{j\ge1}\Big(\sum_{l\in T_j}|v_l|^2\Big)^{1/2}\le\sum_{j\ge1}\Big(k\max_{l\in T_j}|v_l|^2\Big)^{1/2}\\&=\sum_{j\ge1}\sqrt{k}\max_{l\in T_j}|v_l|\le\sum_{j\ge1}\sqrt{k}\min_{l\in T_{j-1}}|v_l|\le\sum_{j\ge1}\sqrt{k}\cdot\frac{\sum_{l\in T_{j-1}}|v_l|}{k}\\&=\sum_{j\ge1}\frac{\|v_{T_{j-1}}\|_1}{\sqrt{k}}=\frac{\|v\|_1}{\sqrt{k}}.\end{aligned}\tag{1.13}$$

□

Combining Theorems 2 and 3, we obtain immediately the following corollary.

Corollary 1. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \le n/2$. If $\delta_{2k}(A) < 1/3$, then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$.*

1.3.4 RIP for random matrices

From what was said up to now, we know that matrices with small restricted isometry constants fulfill the null space property, and sparse solutions of underdetermined linear equations involving such matrices can be found by ℓ_1-minimization (P_1). We discuss in this chapter a class of matrices with small RIP constants. It turns out that the most simple way is to construct these matrices by taking its entries to be independent standard normal variables.

We denote until the end of this section

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \dots & \omega_{mn} \end{pmatrix}, \tag{1.14}$$

where $\omega_{ij}, i = 1,\dots,m, j = 1,\dots,n$, are i.i.d. standard normal variables. We shall show that such a matrix satisfies the RIP with reasonably small constants with high probability.

1.3.4.1 Concentration inequalities

Before we come to the main result of this chapter, we need some properties of independent standard normal variables.

Lemma 1. *(i) Let ω be a standard normal variable. Then $\mathbb{E}\,(e^{\lambda\omega^2}) = 1/\sqrt{1-2\lambda}$ for $-\infty < \lambda < 1/2$.*

(ii) (2-stability of the normal distribution) Let $m \in \mathbb{N}$, let $\lambda = (\lambda_1,\dots,\lambda_m) \in \mathbb{R}^m$ and let $\omega_1,\dots,\omega_m$ be i.i.d. standard normal variables. Then $\lambda_1\omega_1 + \dots + \lambda_m\omega_m \sim (\sum_{i=1}^m \lambda_i^2)^{1/2} \cdot \mathcal{N}(0,1)$, i.e. it is equidistributed with a multiple of a standard normal variable.

Proof. The proof of (i) follows from the substitution $s := \sqrt{1-2\lambda}\cdot t$ in the following way.

$$\mathbb{E}\,(e^{\lambda\omega^2}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda t^2}\cdot e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{(\lambda-1/2)t^2}dt$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-s^2/2}\cdot\frac{ds}{\sqrt{1-2\lambda}} = \frac{1}{\sqrt{1-2\lambda}}.$$

Although the property (ii) is very well known (and there are several different ways to prove it), we provide a simple geometric proof for the sake of completeness. It is enough to consider the case $m = 2$. The general case then follows by induction.

Let therefore $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda \neq 0$, be fixed and let ω_1 and ω_2 be i.i.d. standard normal random variables. We put $S := \lambda_1\omega_1 + \lambda_2\omega_2$. Let $t \geq 0$ be an arbitrary non-negative real number. We calculate

$$\begin{aligned}\mathbb{P}(S \leq t) &= \frac{1}{2\pi}\int_{(u,v):\lambda_1 u+\lambda_2 v\leq t} e^{-(u^2+v^2)/2}dudv = \frac{1}{2\pi}\int_{u\leq c;v\in\mathbb{R}} e^{-(u^2+v^2)/2}dudv\\ &= \frac{1}{\sqrt{2\pi}}\int_{u\leq c} e^{-u^2/2}du.\end{aligned}$$

We have used the rotational invariance of the function $(u,v) \to e^{-(u^2+v^2)/2}$. The value of c is given by the distance of the origin from the line $\{(u,v) : \lambda_1 u + \lambda_2 v = t\}$. It follows by elementary geometry and Pythagorean theorem that (cf. $\Delta OAP \simeq \Delta BAO$ in Figure 1.3)

$$c = |OP| = |OB| \cdot \frac{|OA|}{|AB|} = \frac{t}{\sqrt{\lambda_1^2+\lambda_2^2}}.$$

We therefore get

$$\mathbb{P}(S \leq t) = \frac{1}{\sqrt{2\pi}}\int_{\sqrt{\lambda_1^2+\lambda_2^2}\cdot u\leq t} e^{-u^2/2}du = \mathbb{P}\left(\sqrt{\lambda_1^2+\lambda_2^2}\cdot\omega \leq t\right).$$

The same estimate holds for negative t's by symmetry and the proof is finished. □

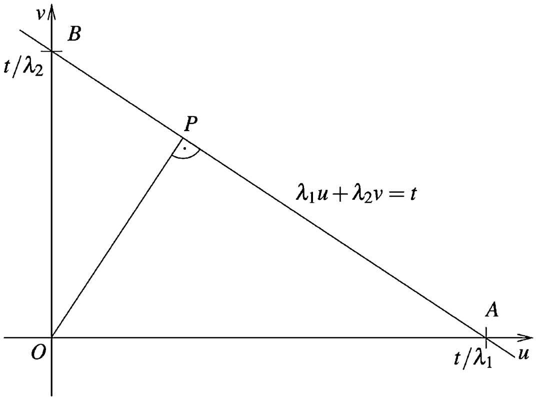


Fig. 1.3 Calculating $c = |OP|$ by elementary geometry for $\lambda_1, \lambda_2 > 0$

If $\omega_1,\dots,\omega_m$ are (possibly dependent) standard normal random variables, then $\mathbb{E}(\omega_1^2+\dots+\omega_m^2)=m$. If $\omega_1,\dots,\omega_m$ are even independent, then the value of $\omega_1^2+\dots+\omega_m^2$ concentrates very strongly around m. This effect is known as *concentration of measure*, cf. [49, 50, 55].

Lemma 2. *Let $m\in\mathbb{N}$ and let $\omega_1,\dots,\omega_m$ be i.i.d. standard normal variables. Let* $0<\varepsilon<1$. *Then*

$$\mathbb{P}(\omega_1^2+\dots+\omega_m^2\ge(1+\varepsilon)m)\le e^{-\frac{m}{2}[\varepsilon^2/2-\varepsilon^3/3]}$$

and

$$\mathbb{P}(\omega_1^2+\dots+\omega_m^2\le(1-\varepsilon)m)\le e^{-\frac{m}{2}[\varepsilon^2/2-\varepsilon^3/3]}.$$

Proof. We prove only the first inequality. The second one follows in exactly the same manner. Let us put $\beta:=1+\varepsilon>1$ and calculate

$$\begin{aligned}\mathbb{P}(\omega_1^2+\dots+\omega_m^2\ge\beta m)&=\mathbb{P}(\omega_1^2+\dots+\omega_m^2-\beta m\ge 0)\\&=\mathbb{P}(\lambda(\omega_1^2+\dots+\omega_m^2-\beta m)\ge 0)\\&=\mathbb{P}(\exp(\lambda(\omega_1^2+\dots+\omega_m^2-\beta m))\ge 1)\\&\le\mathbb{E}\exp(\lambda(\omega_1^2+\dots+\omega_m^2-\beta m)),\end{aligned}$$

where $\lambda>0$ is a positive real number, which shall be chosen later on. We have used the Markov's inequality (1.3) in the last step. Further we use the elementary properties of exponential function and (1.5) for the independent variables $\omega_1,\dots,\omega_m$. This leads to

$$\mathbb{E}\exp(\lambda(\omega_1^2+\dots+\omega_m^2-\beta m))=e^{-\lambda\beta m}\cdot\mathbb{E}\,e^{\lambda\omega_1^2}\cdots e^{\lambda\omega_m^2}=e^{-\lambda\beta m}\cdot(\mathbb{E}\,e^{\lambda\omega_1^2})^m$$

and with the help of Lemma 1 we get finally (for $0<\lambda<1/2$)

$$\mathbb{E}\exp(\lambda(\omega_1^2+\dots+\omega_m^2-\beta m))=e^{-\lambda\beta m}\cdot(1-2\lambda)^{-m/2}.$$

We now look for the value of $0<\lambda<1/2$, which would minimize the last expression. Therefore, we take the derivative of $e^{-\lambda\beta m}\cdot(1-2\lambda)^{-m/2}$ and put it equal to zero. After a straightforward calculation, we get

$$\lambda=\frac{1-1/\beta}{2},$$

which obviously satisfies also $0 < \lambda < 1/2$. Using this value of λ we obtain

$$\mathbb{P}(\omega_1^2 + \cdots + \omega_m^2 \geq \beta m) \leq e^{-\frac{1-1/\beta}{2}\cdot\beta m} \cdot (1-(1-1/\beta))^{-m/2} = e^{-\frac{\beta-1}{2}m} \cdot \beta^{m/2}$$
$$= e^{-\frac{\varepsilon m}{2}} \cdot e^{\frac{m}{2}\ln(1+\varepsilon)}.$$

The result then follows from the inequality

$$\ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}, \quad -1 < t < 1.$$

□

Using 2-stability of the normal distribution, Lemma 2 shows immediately that A defined as in (1.14) acts with high probability as isometry on one fixed $x \in \mathbb{R}^n$.

Theorem 4. *Let $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ and let A be as in* (1.14)*. Then*

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - 1\Big| \geq t\Big) \leq 2e^{-\frac{m}{2}[t^2/2 - t^3/3]} \leq 2e^{-Cmt^2} \tag{1.15}$$

for $0 < t < 1$ *with an absolute constant* $C > 0$.

Proof. Let $x = (x_1, x_2, \ldots, x_n)^T$. Then we get by the 2-stability of normal distribution and Lemma 2

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\|Ax\|_2^2 - 1\Big| \geq t\Big)\\
&= \mathbb{P}\Big(\Big|(\omega_{1,1}x_1 + \cdots + \omega_{1n}x_n)^2 + \cdots + (\omega_{m1}x_1 + \cdots + \omega_{mn}x_n)^2 - m\Big| \geq mt\Big)\\
&= \mathbb{P}\Big(\Big|\omega_1^2 + \cdots + \omega_m^2 - m\Big| \geq mt\Big)\\
&= \mathbb{P}\Big(\omega_1^2 + \cdots + \omega_m^2 \geq m(1+t)\Big) + \mathbb{P}\Big(\omega_1^2 + \cdots + \omega_m^2 \leq m(1-t)\Big)\\
&\leq 2e^{-\frac{m}{2}[t^2/2 - t^3/3]}.
\end{aligned}$$

This gives the first inequality in (1.15). The second one follows by simple algebraic manipulations (for $C = 1/12$). □

Remark 4. (i) Observe that (1.15) may be easily rescaled to

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - \|x\|_2^2\Big| \geq t\|x\|_2^2\Big) \leq 2e^{-Cmt^2}, \tag{1.16}$$

which is true for every $x \in \mathbb{R}^n$.

(ii) A slightly different proof of (1.15) is based on the rotational invariance of the distribution underlying the random structure of matrices defined by (1.14). Therefore, it is enough to prove (1.15) only for one fixed element $x\in\mathbb{R}^n$ with $\|x\|_2=1$. Taking $x=e_1=(1,0,\dots,0)^T$ to be the first canonical unit vector allows us to use Lemma 2 without the necessity of applying the 2-stability of normal distribution.

1.3.4.2 RIP for random Gaussian matrices

The proof of restricted isometry property of random matrices generated as in (1.14) is based on two main ingredients. The first is the concentration of measure phenomenon described in its most simple form in Lemma 2, and reformulated in Theorem 4. The second is the following entropy argument, which allows to extend Theorem 4 and (1.15) from one fixed $x\in\mathbb{R}^n$ to the set Σ_k of all k-sparse vectors.

Lemma 3. *Let $t>0$. Then there is a set $\mathcal{N}\subset\mathbb{S}^{n-1}=\{x\in\mathbb{R}^n:\|x\|_2=1\}$ with*

(i) $|\mathcal{N}|\le(1+2/t)^n$ and
(ii) for every $z\in\mathbb{S}^{n-1}$, there is a $x\in\mathcal{N}$ with $\|x-z\|_2\le t$.

Proof. Choose any $x^1\in\mathbb{S}^{n-1}$. If $x^1,\dots,x^j\in\mathbb{S}^{n-1}$ were already chosen, take $x^{j+1}\in\mathbb{S}^{n-1}$ arbitrarily with $\|x^{j+1}-x^l\|_2>t$ for all $l=1,\dots,j$. This process is then repeated as long as possible, i.e. until we obtain a set $\mathcal{N}=\{x^1,\dots,x^N\}\subset\mathbb{S}^{n-1}$, such that for every $z\in\mathbb{S}^{n-1}$ there is a $j\in\{1,\dots,N\}$ with $\|x^j-z\|_2\le t$. This gives the property (ii).

We will use volume arguments to prove (i). It follows by construction that $\|x^i-x^j\|_2>t$ for every $i,j\in\{1,\dots,N\}$ with $i\ne j$. By triangle inequality, the balls $B(x^j,t/2)$ are all disjoint and are all included in the ball with the center in the origin and radius $1+t/2$. By comparing the volumes we get

$$N\cdot(t/2)^n\cdot V\le(1+t/2)^n\cdot V,$$

where V is the volume of the unit ball in $\mathbb{R}^n$. Hence, we get $N=|\mathcal{N}|\le(1+2/t)^n$.

□

With all these tools at hand, we can now state the main theorem of this section, whose proof follows closely the arguments of [4].

Theorem 5. *Let $n\ge m\ge k\ge 1$ be natural numbers and let $0<\varepsilon<1$ and $0<\delta<1$ be real numbers with*

$$m\ge C\delta^{-2}\Big(k\ln(en/k)+\ln(2/\varepsilon)\Big), \tag{1.17}$$

where $C>0$ *is an absolute constant. Let* A *be again defined by* (1.14)*. Then*

$$\mathbb{P}\big(\delta_k(A)\le\delta\big)\ge 1-\varepsilon.$$

Proof. The proof follows by the concentration inequality of Theorem 4 and the entropy argument described in Lemma 3. By this lemma, there is a set

$$\mathscr{N}\subset Z:=\{z\in\mathbb{R}^n:\text{supp }(z)\subset\{1,\dots,k\},\|z\|_2=1\},$$

such that

(i) $|\mathscr{N}|\le 9^k$ and
(ii) $\min_{x\in\mathscr{N}}\|z-x\|_2\le 1/4$ for every $z\in Z$.

We show that if $\big|\|Ax\|_2^2-1\big|\le\delta/2$ for all $x\in\mathscr{N}$, then $\big|\|Az\|_2^2-1\big|\le\delta$ for all $z\in Z$.

We proceed by the following bootstrap argument. Let $\gamma>0$ be the smallest number, such that $\big|\|Az\|_2^2-1\big|\le\gamma$ for all $z\in Z$. Then $\big|\|Au\|_2^2-\|u\|_2^2\big|\le\gamma\|u\|_2^2$ for all $u\in\mathbb{R}^n$ with supp $(u)\subset\{1,\dots,k\}$. Let us now assume that $\|u\|_2=\|v\|_2=1$ with supp $(u)\cup$supp $(v)\subset\{1,\dots,k\}$. Then we get by polarization identity

$$\begin{aligned}|\langle Au,Av\rangle-\langle u,v\rangle|&=\frac{1}{4}\Big|(\|A(u+v)\|_2^2-\|A(u-v)\|_2^2)-(\|u+v\|_2^2-\|u-v\|_2^2)\Big|\\&\le\frac{1}{4}\Big|\|A(u+v)\|_2^2-\|u+v\|_2^2\Big|+\frac{1}{4}\Big|\|A(u-v)\|_2^2-\|u-v\|_2^2\Big|\\&\le\frac{\gamma}{4}\|u+v\|_2^2+\frac{\gamma}{4}\|u-v\|_2^2=\frac{\gamma}{2}(\|u\|_2^2+\|v\|_2^2)=\gamma.\end{aligned}$$

Applying this inequality to $u'=u/\|u\|_2$ and $v'=v/\|v\|_2$, we obtain

$$|\langle Au,Av\rangle-\langle u,v\rangle|\le\gamma\|u\|_2\|v\|_2 \tag{1.18}$$

for all $u,v\in\mathbb{R}^n$ with supp $(u)\cup$supp $(v)\subset\{1,\dots,k\}$.

Let now again $z\in Z$. Then there is an $x\in\mathscr{N}$, such that $\|z-x\|_2\le 1/4$. We obtain by triangle inequality and (1.18)

$$\begin{aligned}\big|\|Az\|_2^2-1\big|&=\big|\|Ax\|_2^2-1+\langle A(z+x),A(z-x)\rangle-\langle z+x,z-x\rangle\big|\\&\le\delta/2+\gamma\|z+x\|_2\|z-x\|_2\le\delta/2+\gamma/2.\end{aligned}$$

As the supremum of the left-hand side over all admissible z's is equal to γ, we obtain that $\gamma\le\delta$ and the statement follows.

Equipped with this tool, the rest of the proof follows by a simple union bound.

$$\begin{aligned}\mathbb{P}(\delta_k(A)>\delta) &\leq \sum_{\substack{T\subset\{1,\dots,n\}\\ |T|\leq k}} \mathbb{P}\Big(\exists z\in\mathbb{R}^n : \text{supp }(z)\subset T, \|z\|_2=1 \text{ and } \big|\|Az\|_2^2-1\big|>\delta\Big)\\ &= \binom{n}{k}\mathbb{P}\Big(\exists z\in Z \text{ with } \big|\|Az\|_2^2-1\big|>\delta\Big)\\ &\leq \binom{n}{k}\mathbb{P}\Big(\exists x\in\mathscr{N} : \big|\|Ax\|_2^2-1\big|>\delta/2\Big).\end{aligned}$$

By Theorem 4, the last probability may be estimated from above by $2e^{-C'm\delta^2}$. Hence we obtain

$$\mathbb{P}(\delta_k(A)>\delta)\leq 9^k\binom{n}{k}\cdot 2e^{-C'm\delta^2}$$

Hence it is enough to show that the last quantity is at most ε if (1.17) is satisfied. But this follows by straightforward algebraic manipulations and the well-known estimate

$$\binom{n}{k}\leq\frac{n^k}{k!}\leq\Big(\frac{en}{k}\Big)^k. \qquad \square$$

1.3.4.3 Lemma of Johnson and Lindenstrauss

Concentration inequalities similar to (1.15) play an important role in several areas of mathematics. We shall present their connection to the famous result from functional analysis called Johnson–Lindenstrauss lemma, cf. [1, 22, 46, 54]. The lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that the mutual distances between the points are nearly preserved. The connection between this classical result and compressed sensing was first highlighted in [4], cf. also [47].

Lemma 4. *Let* $0<\varepsilon<1$ *and let* m, N *and* n *be natural numbers with*

$$m\geq 4(\varepsilon^2/2-\varepsilon^3/3)^{-1}\ln N.$$

Then for every set $\{x^1,\dots,x^N\}\subset\mathbb{R}^n$ *there exists a mapping* $f:\mathbb{R}^n\to\mathbb{R}^m$*, such that*

$$(1-\varepsilon)\|x^i-x^j\|_2^2\leq\|f(x^i)-f(x^j)\|_2^2\leq(1+\varepsilon)\|x^i-x^j\|_2^2,\qquad i,j\in\{1,\dots,N\}. \tag{1.19}$$

Proof. We put $f(x) = Ax$, where again

$$Ax = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \dots & \omega_{mn} \end{pmatrix} x,$$

and $\omega_{ij}, i = 1, \dots, m, j = 1, \dots, n$ are i.i.d. standard normal variables. We show that with this choice f satisfies (1.19) with positive probability. This proves the existence of such a mapping.

Let $i, j \in \{1, \dots, N\}$ arbitrary with $x^i \neq x^j$. Then we put $z = \frac{x^i - x^j}{\|x^i - x^j\|_2}$ and evaluate the probability that the right-hand side inequality in (1.19) does not hold. Theorem 4 then implies

$$\mathbb{P}\Big(\Big|\|f(x^i) - f(x^j)\|_2^2 - \|x^i - x^j\|_2^2\Big| > \varepsilon\|x^i - x^j\|_2^2\Big) = \mathbb{P}\Big(\Big|\|Az\|^2 - 1\Big| > \varepsilon\Big)$$
$$\leq 2e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}.$$

The same estimate is also true for all $\binom{N}{2}$ pairs $\{i, j\} \subset \{1, \dots, N\}$ with $i \neq j$. The probability that one of the inequalities in (1.19) is not satisfied is therefore at most

$$2 \cdot \binom{N}{2} \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} < N^2 \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} = \exp\Big(2\ln N - \frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]\Big) \leq e^0 = 1$$

for $m \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln N$. Therefore, the probability that (1.19) holds for all $i, j \in \{1, \dots, N\}$ is positive and the result follows. □

1.3.5 Stability and Robustness

The ability to recover sparse solutions of underdetermined linear systems by quick recovery algorithms as ℓ_1-minimization is surely a very promising result. On the other hand, two additional features are obviously necessary to extend this results to real-life applications, namely

- Stability: We want to be able to recover (or at least approximate) also vectors $x \in \mathbb{R}^n$, which are not exactly sparse. Such vectors are called *compressible* and mathematically they are characterized by the assumption that their best k-term approximation decays rapidly with k. Intuitively, the faster the decay of the best k-term approximation of $x \in \mathbb{R}^n$ is, the better we should be able to approximate x.
- Robustness: Equally important, we want to recover sparse or compressible vectors from noisy measurements. The basic model here is the assumptions that the

measurement vector y is given by $y = Ax + e$, where e is small (in some sense). Again, the smaller the error e is, the better we should be able to recover an approximation of x.

We shall show that the methods of compressed sensing can be extended also to this kind of scenario. There is a number of different estimates in the literature, which show that the technique of compressed sensing is stable and robust. We will present only one of them (with more to come in Section 1.4.3). Its proof is a modification of the proof of Theorem 3, and follows closely [11].

Inspired by the form of the noisy measurements just described, we will concentrate on the recovery properties of the following slight modification of (P_1). Namely, let $\eta \geq 0$, then we consider the convex optimization problem

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az-y\|_2 \leq \eta. \tag{$P_{1,\eta}$}$$

If $\eta = 0$, $(P_{1,\eta})$ reduces back to (P_1).

Theorem 6. *Let $\delta_{2k} < \sqrt{2}-1$ and $\|e\|_2 \leq \eta$. Then the solution $\hat{x}$ of $(P_{1,\eta})$ satisfies*

$$\|x-\hat{x}\|_2 \leq \frac{C\sigma_k(x)_1}{\sqrt{k}} + D\eta, \tag{1.20}$$

where $C, D > 0$ are two universal positive constants.

Proof. First, let us recall that if A has RIP of order $2k$ and $u, v \in \Sigma_k$ are two vectors with disjoint supports, then we have by (1.12)

$$|\langle Au, Av\rangle| \leq \delta_{2k}\|u\|_2\|v\|_2. \tag{1.21}$$

Let us put $h = \hat{x} - x$ and let us define the index set $T_0 \subset \{1,\dots,n\}$ as the locations of k largest entries of x taken in the absolute value. Furthermore, we define $T_1 \subset T_0^c$ to be the indices of k largest absolute entries of $h_{T_0^c}$, T_2 the indices of k largest absolute entries of $h_{(T_0\cup T_1)^c}$, etc. As $\hat{x}$ is an admissible point in $(P_{1,\eta})$, the triangle inequality gives

$$\|Ah\|_2 = \|A(x-\hat{x})\|_2 \leq \|Ax-y\|_2 + \|y-A\hat{x}\|_2 \leq 2\eta. \tag{1.22}$$

As $\hat{x}$ is the minimizer of $(P_{1,\eta})$, we get $\|\hat{x}\|_1 = \|x+h\|_1 \leq \|x\|_1$, which we use to show that h must be small outside of T_0. Indeed, we obtain

$$\begin{aligned}\|h_{T_0^c}\|_1 &= \|(x+h)_{T_0^c} - x_{T_0^c}\|_1 + \|(x+h)_{T_0} - h_{T_0}\|_1 - \|x_{T_0}\|_1\\ &\leq \|(x+h)_{T_0^c}\|_1 + \|x_{T_0^c}\|_1 + \|(x+h)_{T_0}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1\\ &= \|x+h\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1\end{aligned}$$

$$\begin{aligned}&\leq \|x\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1\\ &= \|h_{T_0}\|_1 + 2\|x_{T_0^c}\| \leq k^{1/2}\|h_{T_0}\|_2 + 2\sigma_k(x)_1.\end{aligned}$$

Using this together with the approach applied already in (1.13), we derive

$$\sum_{j\geq 2}\|h_{T_j}\|_2 \leq k^{-1/2}\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_2 + 2k^{-1/2}\sigma_k(x)_1. \tag{1.23}$$

We use the RIP property of A, (1.21), (1.22), (1.23) and the simple inequality $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \leq \sqrt{2}\|h_{T_0\cup T_1}\|_2$ and get

$$\begin{aligned}(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2^2 &\leq \|Ah_{T_0\cup T_1}\|_2^2 = \langle Ah_{T_0\cup T_1}, Ah\rangle - \langle Ah_{T_0\cup T_1}, \sum_{j\geq 2} Ah_{T_j}\rangle\\ &\leq \|Ah_{T_0\cup T_1}\|_2\|Ah\|_2 + \sum_{j\geq 2}|\langle Ah_{T_0}, Ah_{T_j}\rangle| + \sum_{j\geq 2}|\langle Ah_{T_1}, Ah_{T_j}\rangle|\\ &\leq 2\eta\sqrt{1+\delta_{2k}}\|h_{T_0\cup T_1}\|_2 + \delta_{2k}(\|h_{T_0}\|_2 + \|h_{T_1}\|_2)\sum_{j\geq 2}\|h_{T_j}\|_2\\ &\leq \|h_{T_0\cup T_1}\|_2\Big(2\eta\sqrt{1+\delta_{2k}} + \sqrt{2}\delta_{2k}\|h_{T_0}\|_2 + 2\sqrt{2}\delta_{2k}k^{-1/2}\sigma_k(x)_1\Big).\end{aligned}$$

We divide this inequality with $(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2$, replace $\|h_{T_0}\|_2$ with the larger quantity $\|h_{T_0\cup T_1}\|_2$ and subtract $\sqrt{2}\delta_{2k}/(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2$ to arrive at

$$\|h_{T_0\cup T_1}\|_2 \leq (1-\rho)^{-1}(\alpha\eta + 2\rho k^{-1/2}\sigma_k(x)_1), \tag{1.24}$$

where

$$\alpha = \frac{2\sqrt{1+\delta_{2k}}}{1-\delta_{2k}} \quad\text{and}\quad \rho = \frac{\sqrt{2}\delta_{2k}}{1-\delta_{2k}}. \tag{1.25}$$

We conclude the proof by using this estimate and (1.23)

$$\begin{aligned}\|h\|_2 \leq \|h_{(T_0\cup T_1)^c}\|_2 + \|h_{T_0\cup T_1}\|_2 &\leq \sum_{j\geq 2}\|h_{T_j}\|_2 + \|h_{T_0\cup T_1}\|_2\\ &\leq 2\|h_{T_0\cup T_1}\|_2 + 2k^{-1/2}\sigma_k(x)_1 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} + D\eta\end{aligned}$$

with $C = 2(1-\rho)^{-1}\alpha$ and $D = 2(1+\rho)(1-\rho)^{-1}$.

We shall give more details on stability and robustness of compressed sensing in Section 1.4.3.

1.3.6 Optimality of bounds

When recovering k-sparse vectors one obviously needs at least $m \geq k$ linear measurements. Even when the support of the unknown vector would be known, this number of measurements would be necessary to identify the value of the non-zero coordinates. Therefore, the dependence of the bound (1.17) on k can possibly only be improved in the logarithmic factor. We shall show that even that is not possible and that this dependence is already optimal as soon as a stable recovery of k-sparse vectors is requested. The approach presented here is essentially taken over from [40].

The proof is based on the following combinatorial lemma.

Lemma 5. *Let $k \leq n$ be two natural numbers. Then there are N subsets $T_1, \dots, T_N$ of $\{1, \dots, n\}$, such that*

(i) $N \geq \left(\frac{n}{4k}\right)^{k/2}$,
(ii) $|T_i| = k$ *for all* $i = 1, \dots, N$ *and*
(iii) $|T_i \cap T_j| < k/2$ *for all* $i \neq j$.

Proof. We may assume that $k \leq n/4$, otherwise one can take $N = 1$ and the statement becomes trivial. The main idea of the proof is straightforward (and similar to the proof of Lemma 3). We choose the sets $T_1, T_2, \dots$ inductively one after another as long as possible, satisfying (ii) and (iii) on the way, and then we show that this process will run for at least N steps with N fulfilling (i).

Let $T_1 \subset \{1, \dots, n\}$ be any set with k elements. The number of subsets of $\{1, \dots, n\}$ with exactly k elements, whose intersection with T_1 has at least $k/2$ elements is bounded by the product of 2^k (i.e., the number of all subsets of T_1) and $\binom{n-k}{\lfloor k/2 \rfloor}$, which is the number of all subsets of T_1^c with at most $k/2$ elements. Therefore there are at least

$$\binom{n}{k} - 2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

sets $T \subset \{1, \dots, n\}$ with k elements and $|T \cap T_1| < k/2$. We select T_2 to be any of them. After the jth step, we have selected sets $T_1, \dots, T_j$ with (ii) and (iii) and there are still

$$\binom{n}{k} - j2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

to choose from. The process stops if this quantity is not positive any more, i.e. after at least

$$N \geq \frac{\binom{n}{k}}{2^k\binom{n-k}{\lfloor k/2\rfloor}} \geq 2^{-k}\frac{\binom{n}{k}}{\binom{n-\lceil k/2\rceil}{\lfloor k/2\rfloor}} = 2^{-k}\frac{n!}{(n-k)!k!}\cdot\frac{(\lfloor k/2\rfloor)!(n-k)!}{(n-\lceil k/2\rceil)!}$$
$$= 2^{-k}\frac{n(n-1)\dots(n-\lceil k/2\rceil+1)}{k(k-1)\dots(k-\lceil k/2\rceil+1)} \geq 2^{-k}\Big(\frac{n}{k}\Big)^{\lceil k/2\rceil} \geq \Big(\frac{n}{4k}\Big)^{k/2}$$

steps.

The following theorem shows that any stable recovery of sparse solutions requires at least m number of measurements, where m is of the order $k\ln(en/k)$.

Theorem 7. *Let $k \leq m \leq n$ be natural numbers, let $A \in \mathbb{R}^{m\times n}$ be a measurement matrix, and let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ be an arbitrary recovery map such that for some constant $C > 0$*

$$\|x-\Delta(Ax)\|_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} \quad \textit{for all} \quad x\in\mathbb{R}^n. \tag{1.26}$$

Then

$$m \geq C'k\ln(en/k) \tag{1.27}$$

with some other constant C' depending only on C.

Proof. We may assume that $C \geq 1$. Furthermore, if k is proportional to n (say $k \geq n/8$), then (1.27) becomes trivial. Hence we may also assume that $k \leq n/8$.

By Lemma 5, there exist index sets $T_1,\dots,T_N$ with $N \geq (n/4k)^{k/2}$, $|T_i| = k$ and $|T_i\cap T_j| < k/2$ if $i\neq j$. We put $x_i = \chi_{T_i}/\sqrt{k}$. Then $\|x_i\|_2 = 1$, $\|x_i\|_1 = \sqrt{k}$ and $\|x_i - x_j\|_2 > 1$ for $i \neq j$.

Let

$$\mathscr{B} = \Big\{z\in\mathbb{R}^n : \|z\|_1 \leq \frac{\sqrt{k}}{4C} \quad \text{and} \quad \|z\|_2 \leq 1/4\Big\}.$$

Then $x_i \in 4C\cdot\mathscr{B}$ for all $i = 1,\dots,N$.

We claim that the sets $A(x_i+\mathscr{B})$ are mutually disjoint. Indeed, let us assume that this is not the case. Then there is a pair of indices $i,j\in\{1,\dots,n\}$ and $z,z'\in\mathscr{B}$ with $i\neq j$ and $A(x_i+z) = A(x_j+z')$. It follows that $\Delta(A(x_i+z)) = \Delta(A(x_j+z'))$ and we get a contradiction by

$$1 < \|x_i-x_j\|_2 = \|(x_i+z-\Delta(A(x_i+z)) - (x_j+z'-\Delta(A(x_j+z')) - z + z')\|_2$$
$$\leq \|(x_i+z-\Delta(A(x_i+z))\|_2 + \|x_j+z'-\Delta(A(x_j+z'))\|_2 + \|z\|_2 + \|z'\|_2$$

$$\leq C\frac{\sigma_k(x_i+z)_1}{\sqrt{k}}+C\frac{\sigma_k(x_j+z')_1}{\sqrt{k}}+\|z\|_2+\|z'\|_2$$
$$\leq C\frac{\|z\|_1}{\sqrt{k}}+C\frac{\|z'\|_1}{\sqrt{k}}+\|z\|_2+\|z'\|_2\leq 1.$$

Furthermore,

$$A(x_i+\mathscr{B})\subset A((4C+1)\mathscr{B}),\quad i=1,\dots,N$$

Let $d\leq m$ be the dimension of the range of A. We denote by $V\neq 0$ the d-dimensional volume of $A(\mathscr{B})$ and compare the volumes

$$\sum_{j=1}^{N}\mathrm{vol}\big(A(x_j+\mathscr{B})\big)\leq \mathrm{vol}\big(A((4C+1)\mathscr{B})\big).$$

Using linearity of A, we obtain

$$\left(\frac{n}{4k}\right)^{k/2}V\leq N\cdot V\leq (4C+1)^dV\leq (4C+1)^mV.$$

We divide by V and take the logarithm to arrive at

$$\frac{k}{2}\ln\left(\frac{n}{4k}\right)\leq m\ln(4C+1). \tag{1.28}$$

If $k\leq n/8$, then it is easy to check that there is a constant $c'>0$, such that

$$\ln\left(\frac{n}{4k}\right)\geq c'\ln\left(\frac{en}{k}\right).$$

Putting this into (1.28) finishes the proof. □

1.4 Extensions

Section 1.3 gives a detailed overview of the most important features of compressed sensing. On the other hand, inspired by many questions coming from application driven research, various additional aspects of the theory were studied in the literature. We present here few selected extensions of the ideas of compressed sensing, which turned out to be the most useful in practice. To keep the presentation reasonable short, we do not give any proofs, and only refer to relevant sources.

1.4.1 Frames and Dictionaries

We have considered in Section 1.3 vectors $x \in \mathbb{R}^n$, which are sparse with respect to the natural canonical basis $\{e_j\}_{j=1}^n$ of $\mathbb{R}^n$. In practice, however, the signal has a sparse representation with respect to a basis (or, more general, with respect to a frame or dictionary). Let us first recall some terminology.

A set of vectors $\{\phi_j\}_{j=1}^n$ in $\mathbb{R}^n$, which is linearly independent and which spans the whole space $\mathbb{R}^n$ is called a basis. It follows easily that such a set necessarily has n elements. Furthermore, every $x \in \mathbb{R}^n$ can be expressed uniquely as a linear combination of the basis vectors, i.e. there is a unique $c = (c_1, \dots, c_n)^T \in \mathbb{R}^n$, such that

$$x = \sum_{j=1}^{n} c_j \phi_j. \tag{1.29}$$

A basis is called orthonormal, if it satisfies the orthogonality relations

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1.30}$$

If $\{\phi\}_{j=1}^n$ is an orthonormal basis and $x \in \mathbb{R}^n$, then the decomposition coefficients c_j in (1.29) are given by $c_j = \langle x, \phi_j \rangle$. Furthermore, the relation

$$\|x\|_2^2 = \sum_{j=1}^{n} |c_j|^2 \tag{1.31}$$

holds true.

Equations (1.29)–(1.30) can be written also in matrix notation. If $\boldsymbol{\Phi}$ is an $n \times n$ matrix with j-th column equal to ϕ_j, then (1.29) becomes $x = \boldsymbol{\Phi} c$ and (1.30) reads $\boldsymbol{\Phi}^T \boldsymbol{\Phi} = I$, where I denoted the $n \times n$ identity matrix. As a consequence, $c = \boldsymbol{\Phi}^T x$. We shall say that x has sparse or compressible representation with respect to the basis $\{\phi_j\}_{j=1}^n$ if the vector $c \in \mathbb{R}^n$ is sparse or compressible, respectively.

To allow for more flexibility in representation of signals, it is often useful to drop the condition of linear independence of the set $\{\phi_j\}_{j=1}^N \subset \mathbb{R}^n$. As before, we represent such a system of vectors by an $n \times N$ matrix $\boldsymbol{\Phi}$. We say that $\{\phi_j\}_{j=1}^N$ is a frame, if there are two positive finite constants $0 < A \leq B$, such that

$$A\|x\|_2^2 \leq \sum_{j=1}^{N} |\langle x, \phi_j \rangle|^2 \leq B\|x\|_2^2. \tag{1.32}$$

From $A > 0$, it follows that the span of the frame vectors is the whole $\mathbb{R}^n$ and, therefore, that $N \geq n$. If one can choose $A = B$ in (1.32), then the frame is called tight. Dual frame of $\boldsymbol{\Phi}$ is any other frame $\tilde{\boldsymbol{\Phi}}$ with

$$\boldsymbol{\Phi}\tilde{\boldsymbol{\Phi}}^T = \tilde{\boldsymbol{\Phi}}\boldsymbol{\Phi}^T = I. \tag{1.33}$$

In general, for a given signal $x \in \mathbb{R}^n$ we can find infinitely many coefficients c, such that $x = \boldsymbol{\Phi}c$. Actually, if $\tilde{\boldsymbol{\Phi}}$ is a dual frame to $\boldsymbol{\Phi}$, one can take $c = \tilde{\boldsymbol{\Phi}}^T x$. One is often interested in finding a vector of coefficients c with $x = \boldsymbol{\Phi}c$, which is optimal in some sense. Especially, we shall say that x has a sparse or compressible representation with respect to the frame $\{\phi_j\}_{j=1}^N$ if c can be chosen sparse or compressible, cf. [33].

It can be shown that the smallest coefficient sequence in the ℓ_2^N sense is obtained by the choice $c = \boldsymbol{\Phi}^\dagger x$, where $\boldsymbol{\Phi}^\dagger$ is the Penrose pseudoinverse. In this context, $\boldsymbol{\Phi}^\dagger$ is also called the canonical dual frame. Finally, let us note that (1.33) implies that

$$\sum_{j=1}^N \langle x, \phi_j \rangle \tilde{\phi}_j = \sum_{j=1}^N \langle x, \tilde{\phi}_j \rangle \phi_j = x$$

for every $x \in \mathbb{R}^n$.

The theory of compressed sensing was extended to the setting of sparse representations with respect to frames and dictionaries in [60]. The measurements now take the form $y = Ax = A\boldsymbol{\Phi}c$, where c is sparse. Essentially, it turns out that if A satisfies the concentration inequalities from Section 1.3.4 and the dictionary $\boldsymbol{\Phi}$ has small coherence, then the matrix $A\boldsymbol{\Phi}$ has small RIP constants, and the methods of compressed sensing can be applied.

1.4.2 Coherence

We have provided in Section 1.3.4 a simple recipe how to construct matrices with small RIP constants - namely to choose each entry independently at random with respect to a correctly normalized standard distribution. On the other hand, if the matrix A is given beforehand, it is quite difficult to check if this matrix really satisfies the RIP, or to calculate its RIP constants. Another property of A, which is easily verifiable and which also ensures good recovery guarantees, is the coherence of A.

Definition 3. Let A be an $m \times n$ matrix and let $a_1, \dots, a_n \in \mathbb{R}^m$ be its columns. Then the coherence of A is the number $\mu(A)$ defined as

$$\mu(A) = \max_{1 \leq i < j \leq n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}. \tag{1.34}$$

Due to Cauchy–Schwartz inequality, $\mu(A) \leq 1$ is always true. If $m \leq n$, then there is a lower bound (known as the Welch bound [71]) on the coherence given by $\mu(A) \geq \sqrt{\frac{n-m}{m(n-1)}}$. We give a particulary elegant proof of this bound, which has recently appeared in [45]. Without loss of generality, we may assume that the vectors $a_1, \dots, a_n$ (which may be even complex) have unit norm and that $\mu = \max_{1 \leq i < j \leq n} |\langle a_i, a_j \rangle|$. Using the notion of the *trace* of a square matrix (which is just the sum of its diagonal entries) and some of its basic and very well-known properties, we obtain

$$\begin{aligned} 0 \leq \mathrm{tr}\Big[(AA^* - \frac{n}{m}I)^2\Big] &= \mathrm{tr}[(A^*A)^2] - \frac{n^2}{m} \\ &= \sum_{k,l=1}^{n} |\langle a_k, a_l \rangle|^2 - \frac{n^2}{m} \leq n + n(n-1)\mu^2 - \frac{n^2}{m}. \end{aligned}$$

Solving this inequality for μ gives the Welch bound.

Let us observe that if $n \gg m$, then this bound reduces to approximately $\mu(A) \geq 1/\sqrt{m}$. There is a lot of possible ways how to construct matrices with small coherence. Not surprisingly, one possible option is to consider random matrices A with each entry generated independently at random, cf. [58, Chapter 11]. Nevertheless the construction of matrices achieving the Welch bound exactly is still an active area of research, making use of ideas from algebra and number theory. On the other hand, it is easy to show that the Welch bound can not be achieved if n is much larger than m. It can be done only if $n \leq m(m+1)/2$ in the real case, and if $n \leq m^2$ in the complex case.

The connection of coherence to RIP is given by the following Lemma.

Lemma 6. *If A has unit-norm columns and coherence $\mu(A)$, then it satisfies the RIP of order k with $\delta_k(A) \leq (k-1)\mu(A)$ for all $k < 1/\mu(A)$.*

Combining this with Theorem 5, it gives recovery guarantees for the number of measurements m growing quadratically in the sparsity k.

1.4.3 Stability and Robustness

Basic discussion of stability and robustness of the methods of compressed sensing was given already in Section 1.3.5 with Theorem 6 being the most important representative of the variety of noise-aware estimates in the area. Its proof follows closely the presentation of [11]. The proof can be easily transformed to the spirit of Section 1.3.2 and 1.3.3 using the following modification of the Null Space Property.

Definition 4. We say that $A \in \mathbb{R}^{m\times n}$ satisfies the ℓ_2-*Robust Null Space Property* of order k with constants $0<\rho<1$ and $\tau>0$ if

$$\|v_T\|_2 \le \frac{\rho\|v_{T^c}\|_1}{\sqrt{k}} + \tau\|Av\|_2 \tag{1.35}$$

for all $v\in\mathbb{R}^n$ and all sets $T\subset\{1,\dots,n\}$ with $|T|\le k$.

The following theorem (which goes essentially back to [14]) is then the noise-aware replacement of Theorem 2.

Theorem 8. *Let $A \in \mathbb{R}^{m\times n}$ with ℓ_2-Robust Null Space Property of order k with constants $0<\rho<1$ and $\tau>0$. Then for any $x\in\mathbb{R}^n$ the solution $\hat{x}$ of $(P_{1,\eta})$ with $y=Ax+e$ and $\|e\|_2\le\eta$ satisfies*

$$\|x-\hat{x}\|_2 \le \frac{C}{\sqrt{k}}\sigma_k(x)_1 + D\eta \tag{1.36}$$

with constants $C,D>0$ depending only on ρ and τ.

Finally, it turns out that the Restricted Isometry Property is also sufficient to guarantee the ℓ_2-Robust Null Space Property and Theorem 3 can be extended to

Theorem 9. *Let $A\in\mathbb{R}^{m\times n}$ and let k be a natural number with $k\le n/2$. If $\delta_{2k}(A)<1/3$, then A satisfies the ℓ_2-Robust Null Space Property of order k with constants $0<\rho<1$ and $\tau>0$ depending only on $\delta_{2k}(A)$.*

Let us only point out, that the constant $1/3$ is by no means optimal, and that the same result (with more technical analysis) holds also if $\delta_{2k}(A)<4/\sqrt{41}$, cf. [9, 10, 38, 39].

Theorems 6 and 8 are sufficient to analyze the situation, when the noise is bounded in the ℓ_2-norm, no matter what the structure of the noise is. Unfortunately, it is not optimal for the analysis of measurements perturbed by Gaussian noise. To demonstrate this, let us assume that $e=(e_1,\dots,e_m)^T$, where e_i's are independent normal variables with variance σ^2, and that

$$y = Ax + e, \tag{1.37}$$

where the entries of $A\in\mathbb{R}^{m\times n}$ are independent standard normal variables. We divide this equation by $\sqrt{m}$ and use that $A'=A/\sqrt{m}$ satisfies the RIP of order k with high probability for $m\ge Ck\ln(eN/k)$. As $\|e/\sqrt{m}\|_2\le 2\sigma$ with high probability, (1.36) becomes for a k-sparse $x\in\mathbb{R}^n$

$$\|x-\hat{x}\|_2 \le D'\sigma. \tag{1.38}$$

We observe that increasing the number of (properly normalised) measurements does not lead to any decay of the approximation error.

To deal with this issue, the following recovery algorithm, called *Dantzig selector*

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|A^T(Az-y)\|_\infty \le \tau, \tag{1.39}$$

was proposed and analyzed in [17]. It deals with the case, when $\|A^Te\|_\infty$ is small.

Theorem 10. *Let $A\in\mathbb{R}^{m\times n}$ be a matrix with RIP of order $2k$ and $\delta_{2k}<\sqrt{2}-1$. Let the measurements y take the form $y=Ax+e$, where $\|A^Te\|_\infty\le\tau$. Then the solution $\hat{x}$ of* (1.39) *satisfies*

$$\|\hat{x}-x\|_2 \le \frac{C}{\sqrt{k}}\sigma_k(x)_1 + D\sqrt{k}\tau, \tag{1.40}$$

where $C,D>0$ depend only on $\delta_{2k}(A)$.

To see how this is related to measurements corrupted with Gaussian noise, let us assume again that the components of $e\in\mathbb{R}^m$ are i.i.d. normal variables with variance σ^2. If the entries of A are again independent standard normal variables, then the 2-stability of normal variables gives that the coordinates of A^Te are independent normal variables with mean zero and variance $\|e\|_2^2$. By simple union bound, we then obtain

$$\mathbb{P}(\|A^Te\|_\infty \ge t\|e\|_2) \le 2n\exp(-t^2/2).$$

Combining this with the fact that $\mathbb{P}(\|e\|_2\ge 2\sigma\sqrt{m})\le\exp(-m/2)$ and choosing $t=2\sqrt{\ln(2n)}$, we finally get

$$\mathbb{P}\big(\|A^Te\|_\infty \ge 4\sigma\sqrt{m\ln(2n)}\big) \le \exp(-m/2)+2n\exp(-2\ln(2n)) \le \frac{1}{n}. \tag{1.41}$$

Dividing (1.37) by $\sqrt{m}$ again and applying Theorem 10, we obtain for the case of a k sparse vector $x\in\mathbb{R}^n$

$$\|x-\hat{x}\|_2 \le D'\sigma\sqrt{\frac{k\ln(2n)}{m}} \tag{1.42}$$

if $m\ge Ck\ln(2n)$. The advantage of (1.42) over (1.38) is that (once $m\ge Ck\ln(2n)$) it decreases with m, i.e. taking more noisy measurements decreases the approximation error.

1.4.4 Recovery algorithms

Although we concentrated on ℓ_1-minimization in the first part of this chapter, there is a number of different algorithms solving the problem of sparse signal recovery. Similarly to ℓ_1-minimization, which was used successfully in machine learning much before the advent of compressed sensing, many of these algorithms also predate the field of compressed sensing. We give an overview of some of these algorithms and refer to [40] for more extensive treatment.

1.4.4.1 ℓ_1-minimization

The ℓ_1-minimization problems (P_1) or ($P_{1,\eta}$) presented before form a backbone of the theory of compressed sensing. Their geometrical background allows for theoretical recovery guarantees, including corresponding stability and robustness extensions. They are formulated as convex optimization problems, which can be solved effectively by any general purpose numerical solver. Furthermore, several implementations dealing with the specific setting of compressed sensing are available nowadays.

Sometimes, it is more convenient to work with some of the equivalent reformulations of ($P_{1,\eta}$). Let us discuss two most important of them. Let $\eta \geq 0$ be given and let $\hat{x}$ be a solution of the optimization problem ($P_{1,\eta}$)

$$\hat{x} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az-y\|_2 \leq \eta. \tag{$P_{1,\eta}$}$$

Then there is a $\lambda \geq 0$, such that $\hat{x}$ is also a solution of the non-constrained convex problem

$$\hat{x} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \frac{1}{2}\|Az-y\|_2^2 + \lambda\|z\|_1. \tag{1.43}$$

This version of ℓ_1-minimization is probably the mostly studied one, see, for example, [34, 41, 51, 73]. On the other hand, if $\lambda > 0$ is given and $\hat{x}$ is a solution to (1.43), then there is an $\eta > 0$, such that $\hat{x}$ is also a solution of ($P_{1,\eta}$). In the same sense, ($P_{1,\eta}$) and (1.43) is also equivalent to *Lasso* (least absolute shrinkage and selection operator, cf. [64])

$$\hat{x} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \|Az-y\|_2^2 \quad \text{s.t.} \quad \|z\|_1 \leq \tau. \tag{1.44}$$

Unfortunately, the values of λ and $\tau > 0$ making these problems equivalent are a-priori unknown.

The last prominent example of an optimization problem, which takes a form of ℓ_1-minimization is the Dantzig selector (1.39). Let us also point out that [7] provides solvers for a variety of ℓ_1-minimization problems.

1.4.4.2 Greedy algorithms

Another approach to sparse recovery is based on iterative identification/approximation of the support of the unknown vector x and of its components. For example, one adds in each step of the algorithm one index to the support to minimize the mismatch to the measured data as much as possible. Therefore, such algorithms are usually referred to as greedy algorithms. For many of them, remarkable theoretical guarantees are available in the literature, sometimes even optimal in the sense of the lower bounds discussed above. Nevertheless, the techniques necessary to achieve these results are usually completely different from those needed to analyze ℓ_1-minimization. We will discuss three of these algorithms, *Orthogonal Matching Pursuit*, *Compressive Sampling Matching Pursuit*, and *Iterative Hard Thresholding*.

Orthogonal Matching Pursuit (OMP)

Orthogonal Matching Pursuit [53, 65, 67] adds in each iteration exactly one entry into the support of $\hat{x}$. After k iterations, it therefore outputs a k-sparse vector $\hat{x}$.

The algorithm finds in each step the column of A most correlated with the residual of the measurements. Its index is then added to the support. Finally, it updates the target vector $\hat{x}_i$ as the vector supported on T_i that best fits the measurements, i.e. which minimizes $\|y - Az\|_2$ among all $z \in \mathbb{R}^n$ with supp $(z) \subset T_i$. It is well known that this vector is given as the product of the Penrose pseudoinverse $A^\dagger$ of A and y.

The formal transcription of this algorithm is given as follows.

```
Orthogonal Matching Pursuit (OMP)
Input: Compressed sensing matrix $A$, measurement vector $y$
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until stopping criterion is met
    $i := i + 1$
    $T_i \leftarrow T_{i-1} \cup \operatorname{supp} H_1(A^T r)$      add largest residual entry to the support
    $\hat{x}_i|_{T_i} \leftarrow A^\dagger_{T_i} y$                    update the estimate of the signal
    $r \leftarrow y - A\hat{x}_i$                          update the residual of the measurements
Output: $\hat{x}_i$
```

It makes use of the hard thresholding operator $H_k(x)$. If $x \in \mathbb{R}^n$ and $k \in \{0, 1, \ldots, n\}$, then $H_k : x \to H_k(x)$ associates with x a vector $H_k(x) \in \mathbb{R}^n$, which is equal to x on the k entries of x with largest magnitude and zero otherwise. The stopping criteria can either limit the overall number of iterations (limiting also the size of the support of the output vector $\hat{x}$) or ensure that the distance between y and $A\hat{x}$ is small in some norm.

The simplicity of OMP is unfortunately connected with one of its weak points. If an incorrect index is added to the support in some step (which can happen in general and depends on the properties of the input parameters), it cannot be removed any more, and stays there until the end of OMP. We refer also to [26] for another variant of OMP.

Compressive Sampling Matching Pursuit (CoSaMP)

One attempt to overcome this drawback is presented in the following algorithm called *Compressive Sampling Matching Pursuit* [57]. It assumes that an additional input is given - namely the expected sparsity of the output. At each step it again enlarges the support, but in contrast to OMP, it will add at least k new entries. Afterwards, it again uses the Penrose pseudo-inverse to find the minimizer of $\|Az-y\|_2$ among all $z \in \mathbb{R}^n$ with $\mathrm{supp}\,(z) \subset T_i$, but this time only the k largest of coordinates of this minimizer are stored.

The formal description is given by the following scheme.

```
Compressive Sampling Matching Pursuit (CoSaMP)
Input: Compressed sensing matrix $A$, measurement vector $y$, sparsity level $k$
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until stopping criterion is met
    $i := i+1$
    $T_i \leftarrow \mathrm{supp}\,(\hat{x}_{i-1}) \cup \mathrm{supp}\, H_{2k}(A^T r)$    update the support
    $\hat{x}_i|_{T_i} \leftarrow H_k(A^{\dagger}_{T_i} y)$    update the estimate of the signal
    $r \leftarrow y - A\hat{x}_i$    update the residual
Output: $\hat{x}_i$
```

Iterative Hard Thresholding (IHT)

The last algorithm [8] we shall discuss is also making use of the hard thresholding operator H_k. The equation $Az = y$ is transformed into $A^T Az = A^T y$, which again can be interpreted as looking for the fixed point of the mapping $z \to (I - A^T A)z + A^T y$. Classical approach is then to iterate this mapping and to put $\hat{x}_i = (I - A^T A)\hat{x}_{i-1} + A^T y = \hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1})$. Iterative Hard Thresholding algorithm is doing exactly this, only combined with the hard thresholding operator H_k.

```
Iterative Hard Thresholding (IHT)
Input: Compressed sensing matrix $A$, measurement vector $y$, sparsity level $k$
Initial values: $\hat{x}_0 = 0, i = 0$
Iteration step: Repeat until stopping criterion is met
    $i := i+1$
    $\hat{x}_i = H_k(\hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1}))$    update the estimate of the signal
Output: $\hat{x}_i$
```

1.4.4.3 Combinatorial algorithms

The last class of algorithms for sparse recovery we shall review were developed mainly in the context of theoretical computer science and they are based on classical ideas from this field, which usually pre-date the area of compressed sensing. Nevertheless, they were successfully adapted to the setting of compressed sensing.

Let us present the basic idea on the example of Group Testing, which was introduced by Robert Dorfman [27] in 1943. One task of United States Public Health Service during the Second World War was to identify all syphilitic soldiers. However, syphilis test in that time was expensive and the naive approach of testing every soldier independently would have been very costly.

If the portion of infected soldiers would be large (say above 50 percent), then the method of individual testing would be reasonable (and nearly optimal). A realistic assumption however is that only a tiny fraction of all the soldiers is infected, say one in thousand, or one in ten thousand. The main idea of the area of Group Testing in this setting is that we can combine blood samples and test a combined sample to check if at least one soldier in the group has syphilis. Another example of this technique is the false coin problem from recreational mathematics, in which one is supposed to identify in a group of n coins a false coin weighting less than a real coin. We refer to [28] to an overview of the methods of Group Testing.

To relate this problem to compressed sensing, let us consider a vector $x = (x_1,\dots,x_n) \in \{0,1\}^n$, where n is the number of soldiers, with $x_i = 0$ if the ith soldier is healthy, or $x_i = 1$ if he has syphilis. The grouping is then represented by an $m \times n$ matrix $A = (a_{ij})$, where $a_{ij} = 1$, if the blood sample of jth soldier was added to ith combined sample. The methods of Group Testing then allow to design efficient matrices A, such that the recovery of x can be done in a surprisingly small number of steps - even linear in the length of the sparse representation of x, i.e. in its sparsity k, cf. [43, 44].

1.4.5 Structured sparsity

In many applications, one has much more prior knowledge about the signal x, than just assuming that it possesses a sparse representation with respect to certain basis, frame, or dictionary.

For example, the image coder JPEG2000 exploits not only the fact that natural images have compressible representation in the wavelet basis (i.e., that most of their wavelet coefficients are small) but it also uses the fact that the values and locations of the large coefficients have a special structure. It turns out that they tend to cluster into a connected subtree inside the wavelet parent–child tree. Using this additional information can of course help to improve the properties of the coder and provide better compression rates [30, 31, 48].

Another model appearing frequently in practice is the model of block-sparse (or joint-sparse) signals. Assume that we want to recover N correlated signals

$x^1,\dots,x^N \in \mathbb{R}^n$ with (nearly) the same locations of their most significant elements. A simple example of such a situation are the three color channels of a natural RGB image, where we intuitively expect the important wavelet coefficients in all three channels to be on nearly the same locations. Furthermore, the same model often appears in the study of DNA microarrays, magnetoencephalography, sensor networks, and MIMO communication [6, 32, 63, 70]. It is usually convenient to represent the signals as columns of an $n\times N$ matrix $X=[x^1\dots x^N]$. The recovery algorithms are then based on mixed matrix norms, which are defined for such an X as

$$\|X\|_{(p,q)}=\Big(\sum_{i=1}^n \|\tilde{x}^i\|_p^q\Big)^{1/q},$$

where $p,q\geq 1$ are real numbers and $\tilde{x}^i$, $i=1,\dots,n$, are the rows of the matrix X. If A is again the sensing matrix and $Y=AX$ are the measurements, then the analogue of (P_1) in this setting is then

$$\hat{X}=\underset{Z\in\mathbb{R}^{n\times N}}{\operatorname{argmin}}\ \|Z\|_{(p,q)} \quad \text{s. t.} \quad Y=AZ$$

for a suitable choice of p and q, typically $(p,q)=(2,1)$. We refer, for example, to [36, 66, 68] for further results.

Finally, let us point out that *model-based compressive sensing* [3] provides a general framework for many different kinds of structured sparsity.

1.4.6 Compressed Learning

In this last part, we will discuss applications of compressed sensing to a classical task of approximation theory, namely to learning of an unknown function f from a limited number of its samples $f(x^1),\dots,f(x^m)$. In its most simple form, treated already in [13] and elaborated in [59], one assumes that the function f is known to be a sparse combination of trigonometric polynomials of maximal order q in dimension d, i.e. that

$$f(x)=\sum_{l\in\{-q,-q+1,\dots,q-1,q\}^d} c_l e^{il\cdot x}$$

and $\|c\|_0\leq k$, where $k\in\mathbb{N}$ is the level of sparsity. Theorem 2.1 of [59] then shows that, with probability at least $1-\varepsilon$, f can be exactly recovered from samples $f(x^1),\dots,f(x^m)$, where $m\geq Ck\ln((2q+1)^d/\varepsilon)$ and $x^1,\dots,x^m$ are uniformly and independently distributed in $[0,2\pi]^d$. The recovery algorithm is given by

$$\underset{c}{\operatorname{argmin}}\|c\|_1 \quad \text{s. t.} \quad \sum_l c_l e^{il\cdot x^j}=f(x^j),\ j=1,\dots,m.$$

We refer to [12, 61] for further results and to [40, Chapter 12] for an overview on random sampling of functions with sparse representation in a bounded orthonormal system.

In another line of study, compressed sensing was used to approximate functions $f:[0,1]^d\to\mathbb{R}$, which depend only on $k\ll d$ (unknown) *active variables* $i_1,\dots,i_k$, i.e.

$$f(x)=f(x_1,\dots,x_d)=g(x_{i_1},\dots,x_{i_k}),\quad x\in[0,1]^d.$$

In [24] and [72], the authors presented sophisticated combinatorial (adaptive and non-adaptive) constructions of sets of sampling points, which allowed for recovery of f to a precision of $1/L$ using only $C(k)(L+1)^k\ln d$ points. Observe that $(L+1)^k$ points would be necessary even if the location of the active coordinates would be known. The use of compressed sensing in this setting was then discussed in [62]. The algorithm developed there was based on approximation of directional derivatives of f at random points $\{x^1,\dots,x^{m_X}\}$ and random directions $\{\varphi^1,\dots,\varphi^{m_\Phi}\}$. Denoting the $m_\Phi\times m_X$ matrix of first order differences as Y and the $m_\Phi\times d$ matrix of random directions by Φ, it was possible to use direct estimates of probability concentrations to ensure that the k largest rows of $\Phi^T Y$ correspond to the k active coordinates of f with high probability. Again, only an additional $\ln d$ factor is paid for identifying the unknown active coordinates.

Finally, the paper [21] initiated a study of approximation of ridge functions of the type

$$f(x)=g(\langle a,x\rangle),\quad x\in[0,1]^d,\tag{1.45}$$

where both the direction $a\in\mathbb{R}^d\setminus\{0\}$ and the univariate function g are unknown. Due to the assumption $a_j\ge 0$ for all $j=1,\dots,d$, posed in [21], it was first possible to approximate g by sampling on grid points along the diagonal $\{\frac{i}{L}(1,\dots,1)^T, i=0,\dots,L\}$. Afterwards, the methods of compressed sensing were used in connection with the first order differences to identify the vector a. The importance of derivatives of f in connection with the assumption (1.45) is best seen from the simple formula

$$\nabla f(x)=g'(\langle a,x\rangle)\cdot a.\tag{1.46}$$

Hence, approximating the gradient of f at a point x gives actually also a scalar multiple of a.

Another algorithm to approximate the ridge functions was proposed in [37]. Similarly to [62], it was based on (1.46) and on approximation of the first order derivatives by first order differences. In contrary to [21], first the ridge direction a was recovered, and only afterwards the ridge profile g was approximated by any standard one-dimensional sampling scheme. Furthermore, no assumptions on signs of a was needed and it was possible to generalize the approach also for recovery of k-ridge functions of the type $f(x)=g(Ax)$, where $A\in\mathbb{R}^{k\times d}$ and g is a function of k variables. We refer also to [18] for further results.

Acknowledgements We would like to thank Sandra Keiper and Holger Rauhut for careful reading of a previous version of this Introduction and for useful comments. The last author was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**, 671–687 (2003)
2. Arora, S., Barak, B.: Computational Complexity: A Modern Approach. Cambridge University Press, Cambridge (2009)
3. Baraniuk, R., Cevher, V., Duarte, M.F., Hegde, C.: Model-based compressive sensing, IEEE Trans. Inf. Theory **56**, 1982–2001 (2010)
4. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**, 253–263 (2008)
5. Baraniuk, R., Steeghs, P.: Compressive radar imaging. In: Proc. IEEE Radar Conf., Boston, pp. 128–133 (2007)
6. Baron, D., Duarte, M.F., Sarvotham, S., Wakin, M.B., Baraniuk, R.: Distributed compressed sensing of jointly sparse signals. In: Proc. Asilomar Conf. Signals, Systems, and Computers, Pacic Grove (2005)
7. Becker, S., Candés, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Math. Program. Comput. **3**, 165–218 (2010)
8. Blumensath, T., Davies, M.: Iterative hard thresholding for compressive sensing. Appl. Comput. Harmon. Anal. **27**, 265–274 (2009)
9. Cai, T., Wang, L., Xu, G.: New bounds for restricted isometry constants. IEEE Trans. Inf. Theory **56**, 4388–4394 (2010)
10. Cai, T., Wang, L., Xu, G.: Shifting inequality and recovery of sparse vectors. IEEE Trans. Signal Process. **58**, 1300–1308 (2010)
11. Candés, E.J.: The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci., Paris, Ser. I **346**, 589–592 (2008)
12. Candés, E.J., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. IEEE Trans. Inf. Theory **57**, 7235–7254 (2011)
13. Candés, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
14. Candés, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**, 1207–1223 (2006)
15. Candés, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 4203–4215 (2005)
16. Candés, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**, 5406–5425 (2006)
17. Candés, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)
18. Cevher, V., Tyagi, H.: Active learning of multi-index function models. In: Proc. NIPS (The Neural Information Processing Systems), Lake Tahoe, Reno (2012)
19. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
20. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**, 211–231 (2009)
21. Cohen, A., Daubechies, I., DeVore, R., Kerkyacharian, G., Picard, D.: Capturing ridge functions in high dimensions from point queries. Constr. Approx. **35**, 225–243 (2012)
22. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithm. **22**, 60–65 (2003)

23. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing. Compressed sensing, pp. 1–64. Cambridge University Press, Cambridge (2012)
24. DeVore, R., Petrova, G., Wojtaszczyk, P.: Approximation of functions of few variables in high dimensions. Constr. Approx. **33**, 125–143 (2011)
25. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
26. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.-L.: Sparse solution of underdetermined systems of linear equations by stagewise Orthogonal Matching Pursuit. IEEE Trans. Inf. Theory **58**, 1094–1121 (2012)
27. Dorfman, R.: The detection of defective members of large populations. Ann. Math. Stat. **14**, 436–440 (1943)
28. Du, D., Hwang, F.: Combinatorial group testing and its applications. World Scientic, Singapore (2000)
29. Duarte, M., Davenport, M., Takhar, D., Laska, J., Ting, S., Kelly, K., Baraniuk R.: Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. **25**, 83–91 (2008)
30. Duarte, M., Wakin, M., Baraniuk, R.: Fast reconstruction of piecewise smooth signals from random projections. In: Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS), Rennes (2005)
31. Duarte, M., Wakin, M., Baraniuk, R.: Wavelet-domain compressive signal reconstruction using a hidden Markov tree model. In: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Las Vegas (2008)
32. Eldar, Y., Mishali, M.: Robust recovery of signals from a structured union of subspaces. IEEE Trans. Inf. Theory **55**, 5302–5316 (2009)
33. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, New York (2010)
34. Figueiredo, M., Nowak, R., Wright, S.: Gradient projections for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Select. Top. Signal Process. **1**, 586–597 (2007)
35. Fornasier, M., Rauhut, H.: Compressive sensing. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, Heidelberg (2011)
36. Fornasier, M., Rauhut, H.: Recovery algorithms for vector valued data with joint sparsity constraints. SIAM J. Numer. Anal. **46**, 577–613 (2008)
37. Fornasier, M., Schnass, K., Vybíral, J.: Learning functions of few arbitrary linear parameters in high dimensions. Found. Comput. Math. **12**, 229–262 (2012)
38. Foucart, S.: A note on guaranteed sparse recovery via 1-minimization. Appl. Comput. Harmon. Anal. **29**, 97–103 (2010)
39. Foucart, S., Lai M.: Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. **26**, 395–407 (2009)
40. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser/Springer, New York (2013)
41. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)
42. Gärtner, B., Matoušek, J.: Understanding and Using Linear Programming. Springer, Berlin (2006)
43. Gilbert, A., Li, Y., Porat, E., and Strauss, M.: Approximate sparse recovery: optimizaing time and measurements. In: Proc. ACM Symp. Theory of Comput., Cambridge (2010)
44. Gilbert, A., Strauss, M., Tropp, J., Vershynin, R.: One sketch for all: fast algorithms for compressed sensing. In: Proc. ACM Symp. Theory of Comput., San Diego (2007)
45. Jasper, J., Mixon, D.G., Fickus M.: Kirkman equiangular tight frames and codes. IEEE Trans. Inf. Theory **60**, 170–181 (2014)
46. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conf. in Modern Analysis and Probability, pp. 189–206 (1984)
47. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**, 1269–1281 (2011)
48. La, C., Do, M.N.: Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In: IEEE Int. Conf. Image Processing (ICIP), Atlanta (2006)

49. Ledoux, M.: The Concentration of Measure Phenomenon. American Mathematical Society, Providence (2001)
50. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Isoperimetry and Processes. Springer, Berlin (1991)
51. Loris, I.: On the performance of algorithms for the minimization of ℓ_1-penalized functions. Inverse Prob. **25**, 035008 (2009)
52. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**, 1182–1195 (2007)
53. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**, 3397–3415 (1993)
54. Matoušek, J.: On variants of the Johnson-Lindenstrauss lemma. Rand. Struct. Algorithm. **33**, 142–156 (2008)
55. Milman, V.D., Schechtman, G.: Asymptotic theory of finite-dimensional normed spaces. Springer, Berlin (1986)
56. Mishali, M., Eldar, Y.: From theory to practice: Sub-nyquist sampling of sparse wideband analog signals. IEEE J. Sel. Top. Signal Process. **4**, 375–391 (2010)
57. Needell, D., Tropp, J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**, 301–321 (2009)
58. Pietsch, A.: Operator Ideals. North-Holland, Amsterdam (1980)
59. Rauhut, H.: Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**, 16–42 (2007)
60. Rauhut, H., Schnass, K., Vandergheynst, P.: Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theor. **54**, 2210–2219 (2008)
61. Rauhut, H., Ward, R.: Sparse Legendre expansions via ℓ_1-minimization. J. Approx. Theory **164**, 517–533 (2012)
62. Schnass, K., Vybíral, J.: Compressed learning of high-dimensional sparse functions. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 3924–3927 (2011)
63. Stojnic, M., Parvaresh, F., Hassibi, B.: On the reconstruction of block-sparse signals with an optimal number of measurements. IEEE Trans. Signal Process. **57**, 3075–3085 (2009)
64. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)
65. Tropp, J.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theor. **50**, 2231–2242 (2004)
66. Tropp, J.: Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. Signal Process. **86**, 589–602 (2006)
67. Tropp, J., Gilbert, A.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory **53**, 4655–4666 (2007)
68. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Process. **86**, 572–588 (2006)
69. Tropp, J., Laska, J., Duarte, M., Romberg, J., Baraniuk, R.: Beyond Nyquist: efficient sampling of sparse bandlimited signals. IEEE Trans. Inf. Theor. **56**, 520–544 (2010)
70. Wakin, M.B., Sarvotham, S., Duarte, M.F., Baron, D., Baraniuk, R.: Recovery of jointly sparse signals from few random projections. In: Proc. Workshop on Neural Info. Proc. Sys. (NIPS), Vancouver (2005)
71. Welch, L.: Lower bounds on the maximum cross correlation of signals. IEEE Trans. Inf. Theory **20**, 397–399 (1974)
72. Wojtaszczyk, P.: Complexity of approximation of functions of few variables in high dimensions. J. Complex. **27**, 141–150 (2011)
73. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imag. Sci. **1**, 143–168 (2008)

*Johnson-Lindenstrauss Lemma for Circulant Matrices**

Aicke Hinrichs,[1] Jan Vybíral[2]
[1] *Department of Mathematics, Universität Jena, 07740 Jena, Germany; e-mail: a.hinrichs@uni-jena.de*
[2] *Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, A-4040 Linz, Austria; e-mail: jan.vybiral@oeaw.ac.at*

Received 27 January 2010; accepted 23 August 2010
Published online 21 January 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/rsa.20360

ABSTRACT: We prove a variant of a Johnson-Lindenstrauss lemma for matrices with circulant structure. This approach allows to minimize the randomness used, is easy to implement and provides good running times. The price to be paid is the higher dimension of the target space $k = O(\varepsilon^{-2} \log^3 n)$ instead of the classical bound $k = O(\varepsilon^{-2} \log n)$. © 2011 Wiley Periodicals, Inc. Random Struct. Alg., 39, 391–398, 2011

Keywords: Johnson-Lindenstrauss lemma; circulant matrices; decoupling lemma

1. INTRODUCTION

The classical Johnson-Lindenstrauss lemma may be formulated as follows.

Theorem 1.1. *Let $\varepsilon \in (0, \frac{1}{2})$ and let $x_1, \dots, x_n \in \mathbb{R}^d$ be arbitrary points. Let $k = O(\varepsilon^{-2} \log n)$ be a natural number. Then there exists a (linear) mapping $f : \mathbb{R}^d \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|f(x_i) - f(x_j)\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$

for all $i,j \in \{1, \dots, n\}$. Here $\|\cdot\|_2$ stands for the Euclidean norm in $\mathbb{R}^d$ or $\mathbb{R}^k$, respectively.

The original proof of Johnson and Lindenstrauss [14] uses (up to a scaling factor) an orthogonal projection onto a random k-dimensional subspace of $\mathbb{R}^d$. We refer also to [8]

Correspondence to: J. Vybíral

*Supported by DFG Heisenberg (HI 584/3-2); FWF START-Preis Sparse Approximation and Optimization in High Dimensions (Y 432-N15).

for a beautiful and self-contained proof. Later on, this lemma found many applications, especially in design of algorithms, where it sometimes allows to reduce the dimension of the underlying problem essentially and break the so-called “curse of dimension”, cf. [12] or [13].

The evaluation of $f(x)$, where f is a projection onto a random k dimensional subspace, is a very time-consuming operation. Therefore, a significant effort was devoted to

- minimize the running time of $f(x)$,
- minimize the memory used,
- minimize the number of random bits used,
- simplify the algorithm to allow an easy implementation.

Achlioptas observed in [1], that the mapping may also be realised by a matrix, where each component is selected independently at random with a fixed distribution. This decreases the time for evaluation of $f(x)$ essentially.

An important breakthrough was achieved by Ailon and Chazelle in [3]. Let us briefly describe their *Fast Johnson-Lindenstrauss transform* (FJLT). The FJLT is the product of three matrices $f(x) = PHDx$, where

- P is a $k \times d$ matrix, where each component is generated independently at random. In particular, $P_{i,j} \approx N(0,1)$ with probability
$$q = \min\left\{\Theta\left(\frac{\log^2 n}{d}\right), 1\right\}$$
and $P_{i,j} = 0$ with probability $1 - q$,
- H is the $d \times d$ normalised Hadamard matrix,
- D is a random $d \times d$ diagonal matrix, with each $D_{i,i}$ drawn independently from $\{-1, 1\}$ with probability 1/2.

It follows, that with high probability, $f(x)$ may be calculated in time $O(d \log d + qd\varepsilon^{-2} \log n)$.

We refer to [17] for a historical overview as well as for an extensive description of the present “state of the art.” The ultimate goal, namely to find a fast Johnson-Lindenstrauss transform for *all* admissible parameters k, n, d and ε with optimal bound on k, remains open.

In this note we propose another direction to approach the Johnson-Lindenstrauss lemma, namely we investigate the possibility of taking a partial circulant matrix for f combined with a random ±1 diagonal matrix, see the next section for exact definitions.

This transform may be implemented using the Fast Fourier Transform, cf. [9, Section 4.7.7], and has therefore a running time of $O(d \log d)$. It requires $2d$ random bits (instead of kd used in [1] or $d + O(k \log^2 n \log(kd))$ used in [2,3]) and allows a simpler implementation.

Unfortunately, up to now, we were only able to prove the statement with $k = O(\varepsilon^{-2} \log^3 n)$, compared to the standard value $k = O(\varepsilon^{-2} \log n)$. We leave the possible improvements of this bound open for further investigations.

2. CIRCULANT MATRICES

We study the question (which to our knowledge has not been addressed in the literature before), whether$f$ in the Johnson-Lindenstrauss lemma may be chosen as a circulant matrix. Let us give the necessary notation.

Let $a = (a_0, \dots, a_{d-1})$ be independent identically distributed random variables. We denote by $M_{a,k}$ the partial circulant matrix

$$M_{a,k} = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{d-1} \\ a_{d-1} & a_0 & a_1 & \dots & a_{d-2} \\ a_{d-2} & a_{d-1} & a_0 & \dots & a_{d-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{d-k+1} & a_{d-k+2} & a_{d-k+3} & \dots & a_{d-k} \end{pmatrix}.$$

Furthermore, if $\varkappa = (\varkappa_0, \dots, \varkappa_{d-1})$ are independent Bernoulli variables, we put

$$D_\varkappa = \begin{pmatrix} \varkappa_0 & 0 & \dots & 0 \\ 0 & \varkappa_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varkappa_{d-1} \end{pmatrix}.$$

Theorem 2.1. *Let $x_1, \dots, x_n$ be arbitrary points in $\mathbb{R}^d$, let $\varepsilon \in (0, \frac{1}{2})$ and let $k = O(\varepsilon^{-2} \log^3 n)$ be a natural number. Let $a = (a_0, \dots, a_{d-1})$ be independent Bernoulli variables or independent normally distributed variables. Let $M_{a,k}$ and $D_\varkappa$ be as above and put $f(x) = \frac{1}{\sqrt{k}} M_{a,k} D_\varkappa x$.*

Then with probability at least 2/3 the following holds

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|f(x_i) - f(x_j)\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2, \qquad i,j = 1, \dots, n.$$

The preconditioning of x using $D_\varkappa$ seems to be necessary and we shall comment on this point later on. Its role may be compared with the use of the random Fourier transform in [3].

In contrast to the above mentioned variants of the Johnson-Lindenstrauss lemma, the coordinates of$f(x)$ are now no longer independent random variables. Our approach "decouples" the dependence caused by the circulant structure. It resembles in some aspects the methods used recently in compressed sensing, cf. [5,6,18].

First, we recall the Lemma 1 from Section 4.1 of [16] (cf. also Lemma 2.2 of [17]), which shall be useful later on.

Lemma 2.2. *Let*

$$Z = \sum_{i=1}^{D} \alpha_i \big(a_i^2 - 1\big),$$

where a_i are i.i.d. normal variables and α_i are nonnegative real numbers. Then for any $t > 0$

$$\mathbb{P}(Z \ge 2\|\alpha\|_2\sqrt{t} + 2\|\alpha\|_\infty t) \le \exp(-t),$$
$$\mathbb{P}(Z \le -2\|\alpha\|_2\sqrt{t}) \le \exp(-t).$$

Furthermore, we shall use the decoupling lemma of [7, Proposition 1.9].

Lemma 2.3. *Let $\xi_0, \dots, \xi_{d-1}$ be independent random variables with $\mathbb{E}\,\xi_0 = \cdots = \mathbb{E}\,\xi_{d-1} = 0$ and let $\{x_{i,j}\}_{i,j=0}^{d-1}$ be a double sequence of real numbers. Then for $1 \le p < \infty$*

$$\mathbb{E}\left|\sum_{i\neq j} x_{i,j}\xi_i\xi_j\right|^p \le 4^p\mathbb{E}\left|\sum_{i\neq j} x_{i,j}\xi_i\xi_j'\right|^p,$$

where $(\xi_0', \dots, \xi_{d-1}')$ denotes an independent copy of $(\xi_0, \dots, \xi_{d-1})$.

The key role in the proof of the Johnson-Lindenstrauss lemma is played by the following estimates.

Lemma 2.4. *Let $k \le d$ be natural numbers and let $\varepsilon \in (0, \frac{1}{2})$. Let $a = (a_0, \dots, a_{d-1})$, $M_{a,k}$ and $D_\varkappa$ be as in Theorem 2.1 and let $x \in \mathbb{R}^d$ be a unit vector. Put $f(x) = M_{a,k}D_\varkappa x$.*

Then there is a constant c, independent on k, d, ε and x, such that

$$\mathbb{P}_{a,\varkappa}\big(\|f(x)\|_2^2 \ge (1+\varepsilon)k\big) \le \exp(-c(k\varepsilon^2)^{1/3})$$

and

$$\mathbb{P}_{a,\varkappa}\big(\|f(x)\|_2^2 \le (1-\varepsilon)k\big) \le \exp(-c(k\varepsilon^2)^{1/3}).$$

Proof. Let $S : \mathbb{R}^d \to \mathbb{R}^d$ denote the shift operator

$$S(x_0, x_1, \dots, x_{d-1}) = (x_{d-1}, x_0, x_1, \dots, x_{d-2}), \quad x \in \mathbb{R}^d.$$

Then

$$\|f(x)\|_2^2 = \|M_{a,k}D_\varkappa x\|_2^2 = \sum_{j=0}^{k-1} |\langle S^j a, D_\varkappa x\rangle|^2 = \sum_{j=0}^{k-1}\left(\sum_{i=0}^{d-1} a_i\varkappa_{j+i}x_{j+i}\right)^2 = I_1 + I_2,$$

where

$$I_1 = \sum_{i=0}^{d-1} a_i^2 \cdot \sum_{j=0}^{k-1} x_{j+i}^2$$

and

$$I_2 = \sum_{j=0}^{k-1}\sum_{i\neq i'} a_i a_{i'}\varkappa_{j+i}\varkappa_{j+i'}x_{j+i}x_{j+i'}.$$

Here (and any time later) the summation in the index is to be understood modulo d.

The decoupling of the circulant matrix is based on

$$\mathbb{P}_{a,\varkappa}\big(\|M_{a,k}D_\varkappa x\|_2^2 \ge (1+\varepsilon)k\big) \le \mathbb{P}_a(I_1 \ge (1+\varepsilon/2)k) + \mathbb{P}_{a,\varkappa}(I_2 \ge \varepsilon k/2) \tag{2.1}$$

and

$$\mathbb{P}_{a,\varkappa}\big(\|M_{a,k}D_\varkappa x\|_2^2 \le (1-\varepsilon)k\big) \le \mathbb{P}_a(I_1 \le (1-\varepsilon/2)k) + \mathbb{P}_{a,\varkappa}(I_2 \le -\varepsilon k/2). \tag{2.2}$$

We use Lemma 2.2 to estimate the diagonal term I_1.

We choose $\alpha_i = \sum_{j=0}^{k-1} x_{j+i}^2$ and get $\|\alpha\|_1 = k$, $\|\alpha\|_\infty \le 1$ and hence $\|\alpha\|_2 \le \sqrt{k}$. This leads to

$$\mathbb{P}_a(I_1 \le k - 2\sqrt{kt}) \le \exp(-t) \tag{2.3}$$

and

$$\mathbb{P}_a(I_1 \ge k + 2\sqrt{kt} + 2t) \le \exp(-t). \tag{2.4}$$

We set $\varepsilon k/2 = 2\sqrt{kt}$, i.e. $t = \varepsilon^2 k/16$, in (2.3) and obtain

$$\mathbb{P}_a(I_1 \le (1-\varepsilon/2)k) \le \exp(-\varepsilon^2 k/16). \tag{2.5}$$

On the other hand, if $c_1 = 5/2 - \sqrt{6} > 1/20$, then $\sqrt{c_1} + c_1/2 = 1/4$ and

$$2\sqrt{kt} + 2t \le \varepsilon k/2$$

for $t = c_1\varepsilon^2 k$, which finally gives

$$\mathbb{P}_a(I_1 \ge (1+\varepsilon/2)k) \le \exp(-c_1\varepsilon^2 k). \tag{2.6}$$

Next, we estimate the moments of the off-diagonal part I_2. We use Lemma 2.3 twice, which gives

$$\mathbb{E}_{a,\varkappa}|I_2|^p \le 16^p \mathbb{E}_{a,a',\varkappa,\varkappa'}\left|I_2'\right|^p := 16^p \mathbb{E}_{a,a',\varkappa,\varkappa'}\left|\sum_{j=0}^{k-1}\sum_{i\ne i'} a_i a'_{i'} \varkappa_{j+i}\varkappa'_{j+i'} x_{j+i} x_{j+i'}\right|^p,$$

where a' and $\varkappa'$ are independent copies of a and $\varkappa$, respectively.

First, we make a substitution $v = j+i, v' = j+i'$ and use the Khintchine inequality with the optimal constant $C_p \le \sqrt{p}$, cf. [10], and the random variable $\varkappa$ to obtain

$$\begin{aligned}
\mathbb{E}_{\varkappa}\left|\sum_{j=0}^{k-1}\sum_{i\ne i'} a_i a'_{i'} \varkappa_{j+i}\varkappa'_{j+i'} x_{j+i} x_{j+i'}\right|^p &= \mathbb{E}_{\varkappa}\left|\sum_{v=0}^{d-1} \varkappa_v x_v \sum_{v'\ne v} \varkappa'_{v'} x_{v'} \sum_{j=0}^{k-1} a_{v-j} a'_{v'-j}\right|^p \\
&\le C_p^p \left(\sum_{v=0}^{d-1} x_v^2 \left(\sum_{v'\ne v} \varkappa'_{v'} x_{v'} \sum_{j=0}^{k-1} a_{v-j} a'_{v'-j}\right)^2\right)^{p/2}.
\end{aligned}$$

Next, we involve Minkowski's inequality with respect to $p/2 \ge 1$ and Khintchine's inequality for the random variable $\varkappa'$.

$$\begin{aligned}
\mathbb{E}_{\varkappa,\varkappa'}\left|I_2'\right|^p &\le C_p^p\, \mathbb{E}_{\varkappa'} \left(\sum_{v=0}^{d-1} x_v^2 \left(\sum_{v'\ne v} \varkappa_{v'} x_{v'} \sum_{j=0}^{k-1} a_{v-j} a'_{v'-j}\right)^2\right)^{p/2} \\
&\le C_p^p \left(\sum_{v=0}^{d-1} x_v^2 \left(\mathbb{E}_{\varkappa'}\left|\sum_{v'\ne v} \varkappa_{v'} x_{v'} \sum_{j=0}^{k-1} a_{v-j} a'_{v'-j}\right|^p\right)^{2/p}\right)^{p/2} \\
&\le C_p^{2p} \left(\sum_{v\ne v'} x_v^2 x_{v'}^2 \left(\sum_{j=0}^{k-1} a_{v-j} a'_{v'-j}\right)^2\right)^{p/2}.
\end{aligned}$$

Furthermore, the Minkowski inequality for a and a' gives

$$\mathbb{E}_{a,a',\varkappa,\varkappa'}\left|I_2'\right|^p \le C_p^{2p}\left(\sum_{\nu\neq\nu'} x_\nu^2 x_{\nu'}^2\left(\mathbb{E}_{a,a'}\left|\sum_{j=0}^{k-1} a_{\nu-j}a'_{\nu'-j}\right|^p\right)^{2/p}\right)^{p/2}.$$

If $a_0,\dots,a_{d-1}$ are Bernoulli variables, then Khintchine's inequality gives

$$\left(\mathbb{E}_{a,a'}\left|\sum_{j=0}^{k-1} a_{\nu-j}a'_{\nu'-j}\right|^p\right)^{1/p} \le \sqrt{kp},$$

as the product of two independent Bernoulli variables is again of this type.

For normal variables, we use first Khintchine's inequality and spherical coordinates to obtain

$$\begin{aligned}\mathbb{E}_{a,a'}\left|\sum_{j=0}^{k-1} a_{\nu-j}a'_{\nu'-j}\right|^p &= \mathbb{E}_{a,a'}\left|\sum_{j=0}^{k-1} a_j a'_j\right|^p \le C_p^p\mathbb{E}_a\left(\sum_{j=0}^{k-1}|a_j|^2\right)^{p/2}\\ &= C_p^p\mathbb{E}_a\|a\|_2^p = \frac{C_p^p}{(2\pi)^{k/2}}\int_{\mathbb{R}^k} e^{-\|a\|_2^2/2}\|a\|_2^p da\\ &= \frac{C_p^p}{(2\pi)^{k/2}}\cdot A_k\cdot\int_0^\infty e^{-r^2/2}r^{p+k-1}dr, \qquad (2.7)\end{aligned}$$

where

$$A_k = \frac{2\pi^{k/2}}{\Gamma(k/2)}$$

is the area of the unit ball in $\mathbb{R}^k$.

We combine (2.7) with Stirling's inequality and obtain

$$\left(\mathbb{E}_{a,a'}\left|\sum_{j=0}^{k-1} a_{\nu-j}a'_{\nu'-j}\right|^p\right)^{1/p} \le \sqrt{2}C_p\left[\frac{\Gamma((k+p)/2)}{\Gamma(k/2)}\right]^{1/p} \le c_2\sqrt{p(k+p)}.$$

Hence, if $a_0,\dots,a_{d-1}$ are independent Bernoulli or normally distributed variables, we may estimate

$$\left(\mathbb{E}_{a,a',\varkappa,\varkappa'}\left|I_2'\right|^p\right)^{1/p} \le c_2 p\cdot\sqrt{(k+p)p}\cdot\|x\|^2 = c_2p^{3/2}\sqrt{k+p}. \qquad (2.8)$$

Markov's inequality then gives

$$\mathbb{P}_{a,a',\varkappa,\varkappa'}\left(\left|I_2'\right| > k\varepsilon/2\right) = \mathbb{P}_{a,a',\varkappa,\varkappa'}\left(\frac{2^p\left|I_2'\right|^p}{k^p\varepsilon^p} > 1\right) \le \frac{2^p\mathbb{E}_{a,a',\varkappa,\varkappa'}\left|I_2'\right|^p}{k^p\varepsilon^p} \le \left(\frac{2c_2p^{3/2}\sqrt{k+p}}{k\varepsilon}\right)^p.$$

We choose p by the condition $\frac{\sqrt{8}c_2p^{3/2}}{\sqrt{k}\varepsilon} = e^{-1}$. We may assume $c_2 \ge 1$, which ensures that $p \le k$ and $\frac{\sqrt{k+p}}{k} \le \frac{\sqrt{2}}{\sqrt{k}}$, which leads to

$$\mathbb{P}_{a,a',\varkappa,\varkappa'}\left(\left|I_2'\right| > k\varepsilon/2\right) \le \exp(-c_3(k\varepsilon^2)^{1/3}). \qquad (2.9)$$

The proof then follows by (2.1) and (2.2) combined with (2.5), (2.6) and (2.9). ■

The proof of Theorem 2.1 follows from Lemma 2.4 by the union bound over all $\binom{n}{2}$ pairs of points.

Remark 2.5. i. We note that (2.8) follows directly by very well known estimates of moments of Gaussian chaos, cf. [11, 15]. We preferred to give a simple and direct proof.

ii. Let us also mention that Lemma 2.4 fails if the multiplication with $D_\varkappa$ is omitted. Namely, let $k \le d$ be natural numbers, let $a_0, \dots, a_{d-1}$ be independent normal variables and let $x = \frac{1}{\sqrt{d}}(1, \dots, 1)$. If $f(x) = M_{a,k}x$, then

$$\|f(x)\|_2^2 = k\left(\sum_{j=0}^{d-1} \frac{a_j}{\sqrt{d}}\right)^2.$$

Due to the 2-stability of the normal distribution, the variable

$$b := \sum_{j=0}^{d-1} \frac{a_j}{\sqrt{d}}$$

is again normally distributed, i.e. $b \approx N(0, 1)$. Hence

$$\mathbb{P}_a\big(\|f(x)\|_2^2 > (1+\varepsilon)k\big) = \mathbb{P}_b(b^2 > (1+\varepsilon))$$

depends neither on k nor on d and Lemma 2.4 cannot hold.

iii. The statement of Theorem 2.1 holds also for matrices with Toeplitz structure. The proof is literally the same, only notational changes are necessary.

Note added in proof: Interesting new work of Ailon and Liberty [4] appeared during the review process of this paper. Their transformation is a composition of a random sign matrix with a random selection of a suitable number k of rows from a Fourier matrix. Their bound on k, namely $k = O(\varepsilon^{-4} \cdot \log n \cdot \text{polylog } d)$, is optimal up to the polylog d factor. Depending on d and n, this may be better than our bound.

ACKNOWLEDGMENTS

We thank Albrecht Böttcher for his comments to the topic and to the anonymous referees for careful reading of the manuscript and valuable remarks, which helped to improve the paper.

REFERENCES

[1] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, J Comput Syst Sci 66 (2003), 671–687.

[2] N. Ailon and B. Chazelle, Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform, In Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, Washington, 2006.

[3] N. Ailon and B. Chazelle, The fast Johnson-Lindenstrauss transform and approximate nearest neighbors, SIAM J Comput 39 (2009), 302–322.

[4] N. Ailon and E. Liberty, Almost optimal unrestricted fast Johnson-Lindenstrauss transform, Available at: http://arxiv.org/abs/1005.5513, Accessed on May 30, 2010.

[5] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, Toeplitz-structured compressed sensing matrices, In IEEE Workshop SSP, Madison, Wisconsin, 2007.

[6] W. U. Bajwa, J. Haupt, G. Raz, and R. Nowak, Compressed channel sensing, In Proceedings of the CISS08, Princeton, 2008.

[7] J. Bourgain and L. Tzafriri, Invertibility of large submatrices with applications to the geometry of Banach spaces and harmonic analysis, Israel J Math 57 (1987), 137–224.

[8] S. Dasgupta and A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Struct Algorithms 22 (2003), 60–65.

[9] G. Golub and C. F. van Loan, Matrix computations, 3rd edition, The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[10] U. Haagerup, The best constants in the Khintchine inequality, Studia Math 70 (1982), 231–283.

[11] D. L. Hanson and F. T. Wright, A bound on tail probabilities for quadratic forms in independent random variables, Ann Math Statist 42 (1971), 1079–1083.

[12] P. Indyk and R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, In Proceedings of the 30th Annual ACM Symposium on Theory of Computing, Dallas, Texas, 1998, pp. 604–613.

[13] P. Indyk and A. Naor, Nearest neighbor preserving embeddings, ACM Trans Algorithms 3 (2007), Article no. 31.

[14] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, Contem Math 26 (1984), 189–206.

[15] R. Latała, Estimates of moments and tails of Gaussian chaoses, Ann Prob 34 (2006), 2315–2331.

[16] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann Statist 28 (2000), 1302–1338.

[17] J. Matoušek, On variants of the Johnson-Lindenstrauss lemma, Random Struct Algorithms 33 (2008), 142–156.

[18] H. Rauhut, Circulant and Toeplitz matrices in compressed sensing, In Proceeding of the SPARS'09, Saint-Malo, France, 2009.

Available online at www.sciencedirect.com

ScienceDirect

Journal of Functional Analysis 260 (2011) 1096–1105

JOURNAL OF
Functional
Analysis

www.elsevier.com/locate/jfa

A variant of the Johnson–Lindenstrauss lemma for circulant matrices

Jan Vybíral

Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria

Received 15 February 2010; accepted 26 November 2010

Available online 4 December 2010

Communicated by K. Ball

Abstract

We continue our study of the Johnson–Lindenstrauss lemma and its connection to circulant matrices started in Hinrichs and Vybíral (in press) [7]. We reduce the bound on k from $k = \Omega(\varepsilon^{-2}\log^3 n)$ proven there to $k = \Omega(\varepsilon^{-2}\log^2 n)$. Our technique differs essentially from the one used in Hinrichs and Vybíral (in press) [7]. We employ the discrete Fourier transform and singular value decomposition to deal with the dependency caused by the circulant structure.

Keywords: Johnson–Lindenstrauss lemma; Circulant matrix; Discrete Fourier transform; Singular value decomposition

1. Introduction

Let $x^1, \dots, x^n \in \mathbb{R}^d$ be n points in the d-dimensional Euclidean space $\mathbb{R}^d$. The classical Johnson–Lindenstrauss lemma tells that, for a given $\varepsilon \in (0, \frac{1}{2})$ and a natural number $k = \Omega(\varepsilon^{-2}\log n)$, there exists a linear map $f : \mathbb{R}^d \to \mathbb{R}^k$, such that

$$(1-\varepsilon)\|x^j\|_2^2 \leqslant \|f(x^j)\|_2^2 \leqslant (1+\varepsilon)\|x^j\|_2^2$$

for all $j \in \{1, \dots, n\}$.

E-mail address: jan.vybiral@oeaw.ac.at.

0022-1236/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jfa.2010.11.014

Here $\|\cdot\|_2$ stands for the Euclidean norm in $\mathbb{R}^d$ or $\mathbb{R}^k$, respectively. Furthermore, here and any time later, the condition $k = \Omega(\varepsilon^{-2}\log n)$ means, that there is an absolute constant $C > 0$, such that the statement holds for all natural numbers k with $k \geqslant C\varepsilon^{-2}\log n$. We shall also always assume, that $k \leqslant d$. Otherwise, the statement becomes trivial.

The original proof of this fact was given by Johnson and Lindenstrauss in [9]. We refer to [6] for a beautiful and self-contained proof. Since then, it has found many applications for example in algorithm design. These applications inspired numerous variants and improvements of the Johnson–Lindenstrauss lemma, which try to minimize the computational costs of $f(x)$, the memory used, the number of random bits used and to simplify the algorithm to allow an easy implementation. We refer to [8,1–3,12] for details and to [12] for a nice description of the history and the actual "state of the art".

All the known proofs of the Johnson–Lindenstrauss lemma work with random matrices and proceed more or less in the following way. One considers a probability measure $\mathbb{P}$ on a some subset $\mathcal{P}$ of all $k \times d$ matrices (i.e. all linear mappings $\mathbb{R}^d \to \mathbb{R}^k$). The proof of the Johnson–Lindenstrauss lemma then emerges by some variant of the following two estimates

$$\mathbb{P}\bigl(f \in \mathcal{P}\colon \ \bigl\|f(x)\bigr\|_2^2 \geqslant 1+\varepsilon\bigr) < 1 - \frac{1}{2n}$$

and

$$\mathbb{P}\bigl(f \in \mathcal{P}\colon \ \bigl\|f(x)\bigr\|_2^2 \leqslant 1-\varepsilon\bigr) < 1 - \frac{1}{2n},$$

which have to be proven for all unit vectors $x \in \mathbb{R}^d$, and a simple union bound over all points $x^j/\|x^j\|_2$, $j = 1, \ldots, n$. Here and later on we assume, without loss of generality, that $x^j \neq 0$ for all $j = 1, \ldots, n$.

The biggest breakthrough in the attempts to minimize the running time of f was achieved by Ailon and Chazelle in [2] (with improvements by Matoušek [12] and Ailon and Liberty [4]). The mapping f is given in [2] as the composition of a sparse matrix, a certain random Fourier matrix and a random diagonal matrix. The value $f(x)$ can be computed with high probability very efficiently, i.e. using $O(d\log d + \min\{d\varepsilon^{-2}\log n, \varepsilon^{-2}\log^3 n\})$ operations. This was later further improved by Ailon and Liberty to $O(d\log k)$ for $k = O(d^{1/2-\delta})$, for any arbitrary small fixed $\delta > 0$.

In [7], we studied a different construction of f, namely the possibility of a composition of a random circulant matrix with a random diagonal matrix. As a multiple of a circulant matrix may be implemented with the help of a discrete Fourier transform, it provides the running time of $O(d\log d)$, requires very few random bits (only $2d$ random bits in the case of Bernoulli variables) and allows a very simple implementation, as the Fast Fourier Transform is a part of every standard mathematical software package.

The main difference between this approach and the usual constructions available in the literature is that the components of $f(x)$ are now no longer independent random variables. Decoupling this dependence, we were able to prove in [7] the Johnson–Lindenstrauss lemma for composition of a random circulant matrix and a random diagonal matrix, but only for $k = \Omega(\varepsilon^{-2}\log^3 n)$. It is the main aim of this note to improve this bound to $k = \Omega(\varepsilon^{-2}\log^2 n)$. This comes essentially closer to the standard bound $k = \Omega(\varepsilon^{-2}\log n)$. Reaching this optimal bound (and keeping the control of the constants involved) remains an open problem and a subject of a challenging research.

We use a completely different technique here. We use the discrete Fourier transform and the singular value decomposition of circulant matrices. That is the reason, why we found it more instructive to state and prove our variant of Johnson–Lindenstrauss lemma for complex vectors and Gaussian random variables. As a corollary, we obtain of course a corresponding real version.

Before we state our main result, we give the necessary definitions.

Definition 1.1. Let α and β be independent real Gaussian random variables with

$$\mathbb{E}\alpha = \mathbb{E}\beta = 0 \quad \text{and} \quad \mathbb{E}|\alpha|^2 = \mathbb{E}|\beta|^2 = 1.$$

Then we call

$$a = \alpha + i\beta$$

a complex Gaussian variable.

Let us note, that if a is a complex Gaussian variable, then

$$\mathbb{E}a = \mathbb{E}\alpha + i\mathbb{E}\beta = 0 \quad \text{and} \quad \mathbb{E}|a|^2 = \mathbb{E}\alpha^2 + \mathbb{E}\beta^2 = 2.$$

Definition 1.2. (i) Let $k \leqslant d$ be natural numbers. Let $a = (a_0, \dots, a_{d-1}) \in \mathbb{C}^d$ be a fixed complex vector. We denote by $M_{a,k}$ the partial circulant matrix

$$M_{a,k} = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{d-1} \\ a_{d-1} & a_0 & a_1 & \cdots & a_{d-2} \\ a_{d-2} & a_{d-1} & a_0 & \cdots & a_{d-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{d-k+1} & a_{d-k+2} & a_{d-k+3} & \cdots & a_{d-k} \end{pmatrix} \in \mathbb{C}^{k\times d}.$$

If $k = d$, we denote by $M_a = M_{a,d}$ the full circulant matrix. This notation extends naturally to the case, when $a = (a_0, \dots, a_{d-1})$ are independent complex Gaussian variables.

(ii) If $\varkappa = (\varkappa_0, \dots, \varkappa_{d-1})$ are independent Bernoulli variables, we put

$$D_\varkappa = \operatorname{diag}(\varkappa) := \begin{pmatrix} \varkappa_0 & 0 & \cdots & 0 \\ 0 & \varkappa_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varkappa_{d-1} \end{pmatrix} \in \mathbb{R}^{d\times d}.$$

Of course, $D_\varkappa : \mathbb{C}^d \to \mathbb{C}^d$ is an isomorphism.

Theorem 1.3. *Let $\varepsilon \in (0, \frac{1}{2})$, $n \geqslant d$ be natural numbers, and let $x^1, \dots, x^n \in \mathbb{C}^d$ be n arbitrary points in $\mathbb{C}^d$. Let $a = (a_0, \dots, a_{d-1})$ be d independent complex Gaussian variables and let $\varkappa = (\varkappa_0, \dots, \varkappa_{d-1})$ be independent Bernoulli variables.*

If $k = \Omega(\varepsilon^{-2} \log^2 n)$ is a natural number, then the mapping $f : \mathbb{C}^d \to \mathbb{C}^d$ given by $f(x) = \frac{1}{\sqrt{2k}} M_{a,k} D_\varkappa x$ satisfies

$$(1-\varepsilon)\|x^j\|_2^2 \leqslant \|f(x^j)\|_2^2 \leqslant (1+\varepsilon)\|x^j\|_2^2$$

for all $j \in \{1,\dots,n\}$ with probability at least $2/3$. Here $\|\cdot\|_2$ stands for the ℓ_2-norm in $\mathbb{C}^d$ or $\mathbb{C}^k$, respectively.

For reader's convenience, we formulate also a variant of Theorem 1.3, which deals with real Euclidean spaces.

Corollary 1.4. *Let $\varepsilon \in (0, \frac{1}{2})$, $n \geqslant d$ be natural numbers, and let $x^1,\dots,x^n \in \mathbb{R}^{2d}$ be n arbitrary points in $\mathbb{R}^{2d}$. Let $\alpha_0,\dots,\alpha_{d-1},\beta_0,\dots,\beta_{d-1}$ be $2d$ independent real Gaussian variables and let $\varkappa = (\varkappa_0,\dots,\varkappa_{d-1})$ be independent Bernoulli variables.*

If $k = \Omega(\varepsilon^{-2}\log^2 n)$ is a natural number, then the mapping $f : \mathbb{R}^{2d} \to \mathbb{R}^{2k}$ given by

$$f(x) = \frac{1}{\sqrt{2k}} \begin{pmatrix} M_{\alpha,k} & -M_{\beta,k} \\ M_{\beta,k} & M_{\alpha,k} \end{pmatrix} \begin{pmatrix} D_\varkappa & 0 \\ 0 & D_\varkappa \end{pmatrix} x$$

satisfies

$$(1-\varepsilon)\|x^j\|_2^2 \leqslant \|f(x^j)\|_2^2 \leqslant (1+\varepsilon)\|x^j\|_2^2$$

for all $j \in \{1,\dots,n\}$ with probability at least $2/3$. Here $\|\cdot\|_2$ stands for the ℓ_2-norm in $\mathbb{R}^{2d}$ or $\mathbb{R}^{2k}$, respectively.

The proof follows trivially from Theorem 1.3 by considering complex Gaussian variables $a = (\alpha_0 + i\beta_0,\dots,\alpha_{d-1} + i\beta_{d-1})$ and complex vectors $y^j = (x_0^j + ix_d^j,\dots,x_{d-1}^j + ix_{2d-1}^j) \in \mathbb{C}^d$, $j = 1,\dots,n$.

2. Used techniques

We give an overview of the techniques used in the proof of Theorem 1.3.

2.1. Discrete Fourier transform

Our main tool in this note is the discrete Fourier transform. If d is a natural number, then the discrete Fourier transform $\mathcal{F}_d : \mathbb{C}^d \to \mathbb{C}^d$ is defined by

$$(\mathcal{F}_d x)(\xi) = \frac{1}{\sqrt{d}} \sum_{u=0}^{d-1} x_u \exp\left(-\frac{2\pi i u\xi}{d}\right).$$

With this normalization, $\mathcal{F}_d$ is an isomorphism of $\mathbb{C}^d$ onto itself. The inverse discrete Fourier transform is given by

$$\left(\mathcal{F}_d^{-1} x\right)(\xi) = \frac{1}{\sqrt{d}} \sum_{u=0}^{d-1} x_u \exp\left(\frac{2\pi i u\xi}{d}\right).$$

Observe, that the matrix representation of $\mathcal{F}_d^{-1}$ is the conjugate transpose of the matrix representation of $\mathcal{F}_d$, i.e. $\mathcal{F}_d^{-1} = \mathcal{F}_d^*$.

The fundamental connection between discrete Fourier transform and circulant matrices is given by

$$M_a = \mathcal{F}_d \operatorname{diag}(\sqrt{d}\mathcal{F}_d a)\mathcal{F}_d^{-1}, \tag{2.1}$$

which may be verified by direct calculation. Hence every circulant matrix may be diagonalized with the use of a discrete Fourier transform, its inverse and a multiple of the discrete Fourier transform of its first row.

2.2. *Singular value decomposition*

The last tool needed in the proof is the singular value decomposition. Let $M : \mathbb{C}^d \to \mathbb{C}^k$ be a $k \times d$ complex matrix with $k \leqslant d$. Then there exists a decomposition

$$M = U\Sigma V^*,$$

where U is a $k \times k$ unitary complex matrix, Σ is a $k \times k$ diagonal matrix with nonnegative entries on the diagonal, V is a $d \times k$ complex matrix with k orthonormal columns and V^* denotes the conjugate transpose of V. Hence V^* has k orthonormal rows. The entries of Σ are the singular values of M, namely the square roots of the eigenvalues of MM^*.

If $a = (a_0, \ldots, a_{d-1}) \in \mathbb{C}^d$ is a complex vector and M_a is the corresponding circulant matrix, then its singular values may be calculated using (2.1). We obtain

$$\begin{aligned} M_a M_a^* &= \mathcal{F}_d \operatorname{diag}(\sqrt{d}\mathcal{F}_d a)\mathcal{F}_d^{-1}\big[\mathcal{F}_d \operatorname{diag}(\sqrt{d}\mathcal{F}_d a)\mathcal{F}_d^{-1}\big]^* \\ &= \mathcal{F}_d \operatorname{diag}(\sqrt{d}\mathcal{F}_d a) \operatorname{diag}(\overline{\sqrt{d}\mathcal{F}_d a})\mathcal{F}_d^{-1} \\ &= \mathcal{F}_d \operatorname{diag}\big(d|\mathcal{F}_d a|^2\big)\mathcal{F}_d^{-1}. \end{aligned}$$

Hence, the singular values of M_a are $\{\sqrt{d}|(\mathcal{F}_d a)(\xi)|\}_{\xi=0}^{d-1}$.

The action of an arbitrary projection onto a vector of independent real Gaussian variables is very well known. It may be described as follows.

Lemma 2.1. *Let $a = (a_0, \ldots, a_{d-1})$ be independent real Gaussian variables. Let $k \leqslant d$ be a natural number and let $x^1, \ldots, x^k$ be mutually orthogonal unit vectors in $\mathbb{R}^d$. Then*

$$\big\{\langle a, x^j\rangle\big\}_{j=1}^k$$

is equidistributed with a k-dimensional vector of independent real Gaussian variables.

A direct calculation shows, that Lemma 2.1 holds also for complex vectors a and $x^1, \ldots, x^k$. We present the following formulation of this fact.

Lemma 2.2. *Let $a = (a_0, \ldots, a_{d-1})$ be independent complex Gaussian variables. Let W be a $k \times d$ matrix with k orthonormal rows. Then Wa is equidistributed with a k-dimensional vector of independent complex Gaussian variables.*

3. Proof of Theorem 1.3

We shall need the following statement, which describes the preconditioning role of the diagonal matrix $D_\varkappa$. A similar fact has been used also in [2]. Nevertheless, using discrete Fourier transform instead of a Hadamard matrix does not pose any restrictions on the underlying dimension d. Without repeating the details, we point out, that we discussed briefly in [7, Remark 2.5], why this preconditioning may not be omitted.

Lemma 3.1. *Let $n \geqslant d$ be natural numbers and let $x^1, \dots, x^n \in \mathbb{C}^d$ be complex vectors. Let $\varkappa = (\varkappa_0, \dots, \varkappa_{d-1})$ be independent Bernoulli variables. Then there is an absolute constant $C > 0$, such that with probability at least* $5/6$,

$$\left\|\mathcal{F}_d D_\varkappa(x^j)\right\|_\infty \leqslant \frac{C\sqrt{\log n}}{\sqrt{d}} \cdot \left\|x^j\right\|_2 \tag{3.1}$$

holds for all $j \in \{1, \dots, n\}$.

Proof. Let $x = \alpha + i\beta$ be a unit complex vector in $\mathbb{C}^d$. We put $y = (y_0, \dots, y_{d-1}) = \mathcal{F}_d D_\varkappa(x)$. Combining the inclusion

$$\{z \in \mathbb{C}\colon |z| > s\} = \{z \in \mathbb{C}\colon (\Re z)^2 + (\Im z)^2 > s^2\} \subset \left\{z \in \mathbb{C}\colon |\Re z| > \frac{s}{\sqrt{2}}\right\}$$
$$\cup \left\{z \in \mathbb{C}\colon |\Im z| > \frac{s}{\sqrt{2}}\right\}$$

with

$$\mathbb{P}_\varkappa\left(|\Re y_l| > \frac{s}{\sqrt{2}}\right) = 2\mathbb{P}_\varkappa\left(\Re y_l > \frac{s}{\sqrt{2}}\right),$$

we may estimate

$$\mathbb{P}_\varkappa\big(|y_l| > s\big) \leqslant 2\mathbb{P}_\varkappa\left(\Re y_l > \frac{s}{\sqrt{2}}\right) + 2\mathbb{P}_\varkappa\left(\Im y_l > \frac{s}{\sqrt{2}}\right), \quad l = 0, \dots, d-1, \tag{3.2}$$

where

$$\Re y_l = \frac{1}{\sqrt{d}} \sum_{u=0}^{d-1} \varkappa_u\big[\alpha_u \cos(2\pi lu/d) + \beta_u \sin(2\pi lu/d)\big]$$

and

$$\Im y_l = \frac{1}{\sqrt{d}} \sum_{u=0}^{d-1} \varkappa_u\big[\beta_u \cos(2\pi lu/d) - \alpha_u \sin(2\pi lu/d)\big]$$

are the real and the imaginary part of y_l, respectively.

Let $t > 0$ be a real parameter to be chosen later. Using Markov's inequality we may proceed in a standard way:

$$\begin{aligned}
\mathbb{P}_\varkappa\left(\Re y_l > \frac{s}{\sqrt{2}}\right) &= \mathbb{P}_\varkappa\left(\exp\left(t\Re y_l - \frac{st}{\sqrt{2}}\right) > 1\right)\\
&\leqslant \exp\left(-\frac{st}{\sqrt{2}}\right)\mathbb{E}_\varkappa \exp(t\Re y_l)\\
&= \exp\left(-\frac{st}{\sqrt{2}}\right)\prod_{u=0}^{d-1}\cosh\left[\frac{t}{\sqrt{d}}\left[\alpha_u\cos(2\pi lu/d)+\beta_u\sin(2\pi lu/d)\right]\right]\\
&\leqslant \exp\left(-\frac{st}{\sqrt{2}}\right)\prod_{u=0}^{d-1}\exp\left(\frac{t^2}{2d}\left[\alpha_u\cos(2\pi lu/d)+\beta_u\sin(2\pi lu/d)\right]^2\right)\\
&\leqslant \exp\left(-\frac{st}{\sqrt{2}}\right)\prod_{u=0}^{d-1}\exp\left(\frac{t^2}{2d}\left[\alpha_u^2+\beta_u^2\right]\right) = \exp\left(-\frac{st}{\sqrt{2}}+\frac{t^2}{2d}\right).
\end{aligned}$$

We have used the inequality $\cosh(v) \leqslant \exp(v^2/2)$, which holds for all $v \in \mathbb{R}$, and the inequality between geometric and quadratic means. For the optimal $t = \frac{sd}{\sqrt{2}}$, this is equal to $\exp(-\frac{s^2d}{4})$.

As the second summand in (3.2) may be estimated in the same way, we obtain

$$\mathbb{P}_\varkappa\big(|y_l| > s\big) \leqslant 4\exp\left(-\frac{s^2d}{4}\right), \quad l = 0,\dots,d-1. \tag{3.3}$$

Choosing $s = \Omega(d^{-1/2}\sqrt{\log n})$ and applying the union bound over all $nd \leqslant n^2$ components of $\{\mathcal{F}_d D_\varkappa(x^j/\|x^j\|_2)\}_{j=1}^n$, we obtain the result. □

Proof of Theorem 1.3. Let us choose a vector $\varkappa = (\varkappa_0,\dots,\varkappa_{d-1}) \in \{-1,+1\}^d$, such that (3.1) holds. According to Lemma 3.1 this happens with probability at least 5/6.

Let us take $\tilde{x} = \frac{x^j}{\|x^j\|_2}$ for any fixed $j = 1,\dots,n$. We show, that there is an absolute constant $c > 0$, such that

$$\mathbb{P}_a\big(\|M_{a,k}D_\varkappa\tilde{x}\|_2^2 \geqslant 2(1+\varepsilon)k\big) \leqslant \exp\left(-\frac{ck\varepsilon^2}{\log n}\right) \tag{3.4}$$

and

$$\mathbb{P}_a\big(\|M_{a,k}D_\varkappa\tilde{x}\|_2^2 \leqslant 2(1-\varepsilon)k\big) \leqslant \exp\left(-\frac{ck\varepsilon^2}{\log n}\right) \tag{3.5}$$

hold. From (3.4) and (3.5), Theorem 1.3 follows again by a union bound over all $j = 1,\dots,n$.

Let $y^j = S^j(D_\varkappa\tilde{x}) \in \mathbb{C}^d$, $j = 0,\dots,k-1$, where S is the shift operator defined by

$$S: \mathbb{C}^d \to \mathbb{C}^d, \qquad S(z_0,\dots,z_{d-1}) = (z_1,\dots,z_{d-1},z_0).$$

We denote by Y the $k \times d$ matrix with rows $y^0,\dots,y^{k-1}$.

Then it holds

$$\|M_{a,k}D_\varkappa\tilde{x}\|_2^2=\sum_{j=0}^{k-1}\left|\sum_{u=0}^{d-1}a_{(u-j)\bmod d}\varkappa_u\tilde{x}_u\right|^2=\sum_{j=0}^{k-1}\left|\sum_{u=0}^{d-1}y_u^j a_u\right|^2=\|Ya\|_2^2.$$

Let $Y=U\Sigma V^*$ be the singular value decomposition of Y. As mentioned above, $b:=V^*a$ is a k-dimensional vector of independent complex Gaussian variables. Hence,

$$\begin{aligned}\mathbb{P}_a\big(\|Ya\|_2^2>\tau\big)&=\mathbb{P}_a\big(\big\|U\Sigma V^*a\big\|_2^2>\tau\big)=\mathbb{P}_b\big(\|U\Sigma b\|_2^2>\tau\big)\\&=\mathbb{P}_b\big(\|\Sigma b\|_2^2>\tau\big)=\mathbb{P}_b\left(\sum_{j=0}^{k-1}\lambda_j^2|b_j|^2>\tau\right),\end{aligned}$$

holds for every $\tau>0$. Here, λ_j, $j=0,\dots,k-1$, are the singular values of Y. Let us denote $\mu_j=\lambda_j^2$. Then

$$\|\mu\|_1=\sum_{j=0}^{k-1}\lambda_j^2=\|Y\|_F^2=k,$$

where $\|Y\|_F$ is the Frobenius norm of Y.

Moreover,

$$\begin{aligned}\|\mu\|_\infty=\|\lambda\|_\infty^2&=\sup_{z\in\mathbb{C}^d,\|z\|_2\leqslant 1}\|Yz\|_2^2\\&\leqslant\sup_{z\in\mathbb{C}^d,\|z\|_2\leqslant 1}\|M_{D_\varkappa\tilde{x}}z\|_2^2=d\big\|\mathcal{F}_dD_\varkappa(\tilde{x})\big\|_\infty^2\leqslant C^2\log n,\end{aligned}\tag{3.6}$$

where $M_{D_\varkappa\tilde{x}}$ stands for the $d\times d$ complex circulant matrix with the first row equal to $D_\varkappa\tilde{x}$.

This leads finally also to

$$\|\mu\|_2\leqslant\sqrt{\|\mu\|_1\cdot\|\mu\|_\infty}\leqslant C\sqrt{k\log n}.\tag{3.7}$$

Then

$$\mathbb{P}_a\big(\|Ya\|_2^2>2(1+\varepsilon)k\big)=\mathbb{P}_b\left(\sum_{j=0}^{k-1}\mu_j\big(|b_j|^2-2\big)>2\varepsilon k\right).$$

We denote

$$Z:=\sum_{j=0}^{k-1}\mu_j\big(|b_j|^2-2\big).$$

The complex version of Lemma 1 from Section 4.1 of [11] (cf. also Lemma 2.2 of [12]) states that

$$\mathbb{P}_b\big(Z \geqslant 2\sqrt{2}\|\mu\|_2\sqrt{t} + 2\|\mu\|_\infty t\big) \leqslant \exp(-t). \tag{3.8}$$

Using (3.6) and (3.7), we arrive at

$$\mathbb{P}_b\big(Z \geqslant 2\sqrt{2}C\sqrt{tk\log n} + 2C^2 t\log n\big) \leqslant \exp(-t).$$

Choosing $t = \frac{c'k\varepsilon^2}{C^2\log n}$ for $c' > 0$ small enough, we get

$$\mathbb{P}_b(Z \geqslant 2\varepsilon k) \leqslant \exp\left(-\frac{ck\varepsilon^2}{\log n}\right).$$

This finishes the proof of (3.4). Let us note, that (3.5) follows in the same manner with (3.8) replaced by

$$\mathbb{P}_b\big(Z \leqslant -2\sqrt{2}\|\mu\|_2\sqrt{t}\big) \leqslant \exp(-t),$$

which may be again found in Lemma 1, Section 4.1 of [11]. □

Remark 3.2. The statement and the proof of Theorem 1.3 do not change, if we replace the partial circulant matrix $M_{a,k}$ with any $k \times d$ submatrix of M_a.

Note added in proof

Interesting new work of Ailon and Liberty [5] appeared during the review process of this paper. Their transformation is the composition of a random sign matrix with a random selection of a suitable number k of rows from a Fourier matrix. Their bound on k, namely $k = \Omega(\varepsilon^{-4}\log n \cdot \operatorname{polylog} d)$, is optimal up to the $\operatorname{polylog} d$ factor. Depending on d and n, this may be better than our bound.

In another very recent preprint [10], Krahmer and Ward applied the RIP bounds of [13] to prove that partial circulant matrices satisfy the Johnson–Lindenstrauss lemma if

$$k = \Omega\big(\max\big(\varepsilon^{-1}\log^{3/2} n \cdot \log^{3/2} d, \varepsilon^{-2}\log n \cdot \log^4 d\big)\big).$$

Acknowledgments

The author would like to thank Aicke Hinrichs for valuable comments and the anonymous referee for the careful reading of the manuscript. The author also acknowledges the financial support provided by the FWF project Y 432-N15 START-Preis "Sparse Approximation and Optimization in High Dimensions".

References

[1] D. Achlioptas, Database-friendly random projections: Johnson–Lindenstrauss with binary coins, J. Comput. System Sci. 66 (4) (2003) 671–687.

[2] N. Ailon, B. Chazelle, Approximate nearest neighbors and the fast Johnson–Lindenstrauss transform, in: Proc. 38th Annual ACM Symposium on Theory of Computing, 2006.

[3] N. Ailon, B. Chazelle, The fast Johnson–Lindenstrauss transform and approximate nearest neighbors, SIAM J. Comput. 39 (1) (2009) 302–322.

[4] N. Ailon, E. Liberty, Fast dimension reduction using Rademacher series on dual BCH codes, Discrete Comput. Geom. 42 (4) (2009) 615–630.

[5] N. Ailon, E. Liberty, Almost optimal unrestricted fast Johnson–Lindenstrauss transform, http://arxiv.org/abs/1005.5513.

[6] S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Structures Algorithms 22 (2003) 60–65.

[7] A. Hinrichs, J. Vybíral, Johnson–Lindenstrauss lemma for circulant matrices, Random Structures Algorithms, in press.

[8] P. Indyk, R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, in: Proc. 30th Annual ACM Symposium on Theory of Computing, 1998, pp. 604–613.

[9] W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz Mappings into a Hilbert Space, Contemp. Math., vol. 26, 1984, pp. 189–206.

[10] F. Krahmer, R. Ward, New and improved Johnson–Lindenstrauss embeddings via the Restricted Isometry Property, http://arxiv.org/abs/1009.0744.

[11] B. Laurent, P. Massart, Adaptive estimation of a quadratic functional by model selection, Ann. Statist. 28 (5) (2000) 1302–1338.

[12] J. Matoušek, On variants of the Johnson–Lindenstrauss lemma, Random Structures Algorithms 33 (2) (2008) 142–156.

[13] H. Rauhut, J. Romberg, J. Tropp, Restricted isometries for partial random circulant matrices, http://arxiv.org/abs/1010.1847.

Constr Approx (2012) 36:83–115
DOI 10.1007/s00365-012-9163-1

CONSTRUCTIVE
APPROXIMATION

Average Best m-term Approximation

Jan Vybíral

Received: 2 November 2010 / Revised: 8 November 2011 / Accepted: 12 January 2012 /
Published online: 12 April 2012

Abstract We introduce the concept of average best m-term approximation widths with respect to a probability measure on the unit ball or the unit sphere of ℓ_p^n. We estimate these quantities for the embedding $id : \ell_p^n \to \ell_q^n$ with $0 < p \le q \le \infty$ for the normalized cone and surface measure. Furthermore, we consider certain tensor product weights and show that a typical vector with respect to such a measure exhibits a strong compressible (i.e., nearly sparse) structure. This measure may therefore be used as a random model for sparse signals.

Keywords Nonlinear approximation · Best m-term approximation · Average widths · Random sparse vectors · Cone measure · Surface measure

Mathematics Subject Classification Primary: 41A46 · Secondary: 52A20 · 60B11 · 94A12

1 Introduction

1.1 Best m-term Approximation

Let $m \in \mathbb{N}_0$, and let Σ_m be the set of all sequences $x = \{x_j\}_{j=1}^\infty$ with

$$\|x\|_0 := \#\operatorname{supp} x = \#\{n \in \mathbb{N} : x_n \neq 0\} \le m.$$

Communicated by Albert Cohen.

J. Vybíral (✉)
Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: jan.vybiral@ricam.oeaw.ac.at

Here $\#A$ stands for the number of elements of a set A. The elements of Σ_m are said to be *m-sparse*. Observe that Σ_m is a nonlinear subset of every $\ell_q := \{x = \{x_j\}_{j=1}^\infty : \|x\|_q < \infty\}$, where

$$\|x\|_q := \begin{cases} (\sum_{j=1}^\infty |x_j|^q)^{1/q}, & 0 < q < \infty, \\ \sup_{j\in\mathbb{N}} |x_j|, & q = \infty. \end{cases}$$

For every $x \in \ell_q$, we define its *best m-term approximation error* by

$$\sigma_m(x)_q := \inf_{y\in\Sigma_m} \|x - y\|_q.$$

Moreover, for $0 < p \le q \le \infty$, we introduce the *best m-term approximation widths*

$$\sigma_m^{p,q} := \sup_{x:\|x\|_p\le 1} \sigma_m(x)_q.$$

The use of this concept goes back to Schmidt [44], and after the work of Oskolkov [39], it was widely used in approximation theory, cf. [15, 18, 45]. In fact, it is the main prototype of nonlinear approximation [17]. It is well known that

$$2^{-1/p}(m+1)^{1/q-1/p} \le \sigma_m^{p,q} \le (m+1)^{1/q-1/p}, \quad m = 0, 1, 2, \dots. \tag{1}$$

The proof of (1) is based on the simple fact that (roughly speaking) the best m-term approximation error of $x \in \ell_p$ is realized by subtracting the m largest coefficients taken in absolute value. Hence,

$$\sigma_m(x)_q = \begin{cases} (\sum_{j=m+1}^\infty (x_j^*)^q)^{1/q}, & 0 < q < \infty, \\ x_{m+1}^* = \sup_{j\ge m+1} x_j^*, & q = \infty, \end{cases}$$

where $x^* = (x_1^*, x_2^*, \dots)$ denotes the so-called *nonincreasing rearrangement* [6] of the vector $(|x_1|, |x_2|, |x_3|, \dots)$.

Let us recall the proof of (1) in the simplest case, namely $q = \infty$. The estimate from above then follows by

$$\sigma_m(x)_\infty = \sup_{j\ge m+1} x_j^* = x_{m+1}^* \le \left((m+1)^{-1}\sum_{j=1}^{m+1}(x_j^*)^p\right)^{1/p} \le (m+1)^{-1/p}\|x\|_p. \tag{2}$$

The lower estimate is supplied by taking

$$x = (m+1)^{-1/p}\sum_{j=1}^{m+1} e_j, \tag{3}$$

where $\{e_j\}_{j=1}^\infty$ are the canonical unit vectors.

For general q, the estimate from above in (1) may be obtained from (2) and Hölder's inequality

$$\|x\|_q \le \|x\|_p^\theta \cdot \|x\|_\infty^{1-\theta}, \quad \text{where } \frac{1}{q} = \frac{\theta}{p}. \tag{4}$$

The estimate from below follows for all q's by simple modification of (3).

The discussion above exhibits two effects:

(i) Best m-term approximation works particularly well when $1/p - 1/q$ is large, i.e., if $p < 1$ and $q = \infty$.
(ii) The elements used in the estimate from below (and hence the elements where the best m-term approximation performs worst) enjoy a very special structure.

Therefore, there is a reasonable hope that the best m-term approximation could behave better when considered in a certain average case. But first we point out two different interesting points of view on the subject.

1.2 Connection to Compressed Sensing

The interest in ℓ_p spaces (and especially in their finite-dimensional counterparts ℓ_p^n) with $0 < p < 1$ was recently stimulated by the impressive success of the novel and vastly growing area of *compressed sensing* as introduced in [9–11, 19]. Without going much into the details, we only note that the techniques of compressed sensing allow for the reconstruction of a vector from an incomplete set of measurements utilizing the prior knowledge that it is sparse, i.e., $\|x\|_0$ is small. Furthermore, this approach may be applied [14] also to vectors which are *compressible*, i.e., $\|x\|_p$ is small for (preferably small) $0 < p < 1$. Indeed, (1) tells us that such a vector x may be very well approximated by sparse vectors. We point to [8, 24, 25, 42] for the current state of the art of this field and for further references.

This leads in a very natural way to a question that stands in the background of this paper, namely:

What does a typical vector of the ℓ_p^n unit ball look like?

or, posed in an exact way:

Let μ be a probability measure on the unit ball of ℓ_p^n. What is the mean value of $\sigma_m(x)_q$ with respect to this measure?

Of course, the choice of μ plays a crucial role. There are several standard probability measures that are connected to the unit ball of ℓ_p^n in a natural way, namely (cf. Definitions 2 and 9):

(i) the normalized Lebesgue measure,
(ii) the $n-1$ dimensional Hausdorff measure restricted to the surface of the unit ball of ℓ_p^n and correspondingly normalized,
(iii) the so-called normalized cone measure.

Unfortunately, it turns out that all three of these measures are "bad"—a typical vector with respect to any of them does not involve much structure and corresponds to noise rather than to signal (in the sense described below). Therefore, we are looking for a new type of measures (cf. Definition 13) that would behave better in this respect.

1.3 Random Models of Noise and Signals

Random vectors play an important role in the area of signal processing. For example, if $n \in \mathbb{N}$ is a natural number, $\omega = (\omega_1, \dots, \omega_n)$ is a vector of independent Gaussian variables, and $\varepsilon > 0$ is a real number, then $\varepsilon\omega$ is a classical model of noise, namely the *white noise*. This model is used in the theory but also in the real life applications of signal processing.

The random generation of a structured signal seems to be a more complicated task. Probably the most common random model to generate sparse vectors, cf. [7, 13, 30, 40], is the so-called *Bernoulli–Gaussian model*. Again let $n \in \mathbb{N}$ be a natural number and $\varepsilon > 0$ be a real number. Also let $\omega = (\omega_1, \dots, \omega_n)$ stand for a vector of independent Gaussian variables. Furthermore, let $0 < p < 1$ be a real number, and let $\varrho = (\varrho_1, \dots, \varrho_n)$ be a vector of independent Bernoulli variables defined as

$$\varrho_i = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p. \end{cases}$$

The components of the random *Bernoulli–Gaussian vector* $x = (x_1, \dots, x_n)$ are then defined through

$$x_i = \varepsilon \varrho_i \cdot \omega_i, \quad i = 1, \dots, n. \tag{5}$$

Obviously, the average number of nonzero components of x is $k := pn$. Unfortunately, if k is much smaller than n, then the concentration of the number of nonzero components of x around k is not very strong. This improves if k gets larger. But in that case, the model (5) resembles more and more the model of white noise. In some sense, (5) represents a randomly filtered white noise rather than a structured signal. It is one of the main aims of this paper to find a new measure such that a random vector with respect to this measure would show a nearly sparse structure without the need for random filtering.

1.4 Unit Sphere

Let us describe the situation in the most prominent case, when $p = 2$, $m = 0$, and $\mu = \mu_2$ is the normalized surface measure on the unit sphere $\mathbb{S}^{n-1}$ of ℓ_2^n. Furthermore, we denote by γ_n the standard Gaussian measure on $\mathbb{R}^n$ with the density

$$\frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}, \quad x \in \mathbb{R}^n.$$

We use polar coordinates to calculate

$$\begin{aligned} \int_{\mathbb{R}^n} \max_{j=1,\dots,n} |x_j| \, d\gamma_n(x) &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \max_{j=1,\dots,n} |x_j| \cdot e^{-\|x\|_2^2/2} \, dx \\ &= \frac{\Omega_n}{(2\pi)^{n/2}} \int_0^\infty r^{n-1} \int_{\mathbb{S}^{n-1}} \max_{j=1,\dots,n} |rx_j| e^{-\|rx\|_2^2/2} \, d\mu_2(x) \, dr \\ &= \frac{\Omega_n}{(2\pi)^{n/2}} \int_0^\infty r^n e^{-r^2/2} \, dr \cdot \int_{\mathbb{S}^{n-1}} \max_{j=1,\dots,n} |x_j| \, d\mu_2(x) \\ &= \frac{\Omega_n}{(2\pi)^{n/2}} \int_0^\infty r^n e^{-r^2/2} \, dr \cdot \int_{\mathbb{S}^{n-1}} \sigma_0(x)_\infty \, d\mu_2(x), \end{aligned} \tag{6}$$

where Ω_n denotes the area of $\mathbb{S}^{n-1}$. This formula connects the expected value of $\sigma_0(x)_\infty$ with the expected value of a maximum of n independent Gaussian variables. Using that this quantity is known to be equivalent to $\sqrt{\log(n+1)}$, cf. [33, (3.14)],

$$\int_0^\infty r^n e^{-r^2/2}\,dr = 2^{(n-1)/2}\Gamma\bigl((n+1)/2\bigr) \quad \text{and} \quad \Omega_n = \frac{2\pi^{n/2}}{\Gamma(n/2)},$$

one obtains

$$\int_{\mathbb{S}^{n-1}} \sigma_0(x)_\infty\, d\mu_2(x) \approx \sqrt{\frac{\log(n+1)}{n}}, \quad n \in \mathbb{N}. \tag{7}$$

Several comments on (6) and (7) are necessary.

(i) Quantities similar to the left-hand side of (7) have been used in the study of geometry of Banach spaces and local theory of Banach spaces for many years and are treated in detail in the work of Milman [23, 35, 36]. Especially, if $\|\cdot\|_K$ is a norm in $\mathbb{R}^n$ and $K := \{x \in \mathbb{R}^n : \|x\|_K \le 1\}$ denotes the corresponding unit ball, then the quantity

$$A_K = \int_{\mathbb{S}^{n-1}} \|x\|_K\, d\mu_2(x)$$

(and the closely connected median M_K of $\|x\|_K$ over $\mathbb{S}^{n-1}$) plays a crucial role in the Dvoretzky theorem [20, 22, 35] and, in general, in the study of Euclidean sections of K, cf. [36, Sect. 5]. Furthermore, it is known that the case of $K = [-1,1]^n$, when

$$A_K = \int_{\mathbb{S}^{n-1}} \max_{j=1,\dots,n} |x_j|\, d\mu_2(x) = \int_{\mathbb{S}^{n-1}} \sigma_0(x)_\infty\, d\mu_2(x),$$

is extremal, cf. [35].

(ii) The connection between the estimated value of a maximum of independent Gaussian variables and the estimated value of the largest coordinate of a random vector on $\mathbb{S}^{n-1}$ is given just by integration in polar coordinates and is one of the standard techniques in the local theory of Banach spaces. Due to the result of [43], this holds true also for other values of p, even for $p < 1$, with Gaussian variables replaced by variables with the density $c_p e^{-|t|^p}$. This approach is nowadays classical in the study of the geometry and concentration of measure phenomenon on the ℓ_p^n-balls, cf. [2–5, 37, 38, 41].

(iii) For every $x \in \mathbb{S}^{n-1}$, we obtain easily that $\max_{j=1,\dots,n} |x_j| \ge (\frac{1}{n}\sum_{j=1}^n x_j^2)^{1/2} = 1/\sqrt{n}$. Estimate (7) shows that the average value of $\max_{j=1,\dots,n} |x_j|$ over $\mathbb{S}^{n-1}$ is asymptotically larger only by a logarithmic factor. The detailed study of the concentration of $\max_{j=1,\dots,n} |x_j|$ around its estimated value (or its mean value) is known as *concentration of measure phenomena* [32, 33, 36] and gives more accurate information then the one included in (7). As our main interest lies in estimates of *average best m-term widths*, cf. Definition 1, we do not investigate the concentration properties in this paper and leave this subject to further research.

(iv) The calculation (6) is based on the use of polar coordinates. For $p \neq 2$, the normalized cone measure is exactly that measure for which a similar formula holds, cf. (13). The estimates for $n-1$ dimensional surface measure are later obtained using its density with respect to the cone measure, cf. Lemma 10.

(v) As we want to keep the paper self-contained as much as possible and to make it readable also for readers without (almost) any stochastic background, we prefer to use simple and direct techniques. For example, we use the simple estimates in Lemma 5 rather than any of their sophisticated improvements available in the literature.

(vi) The connection to random Gaussian variables explains why a random point of $\mathbb{S}^{n-1}$ is sometimes referred to as *white* (*or Gaussian*) *noise*. It is usually not associated with any reasonable (i.e., structured) signal; rather, it represents a good model for random noise.

1.5 Basic Definitions and Main Results

1.5.1 Definition of Average Best m-term Widths

Having described the context of our work, we shall now present the definition of the so-called *average best m-term widths*, which are the main subject of our study.

First, we observe that

$$\sigma_m\big((x_1,\dots,x_n)\big)_q = \sigma_m\big((\varepsilon_1 x_1,\dots,\varepsilon_n x_n)\big)_q = \sigma_m\big((|x_1|,\dots,|x_n|)\big)_q$$

holds for every $x \in \mathbb{R}^n$ and $\varepsilon \in \{-1,+1\}^n$. Also, all the measures we shall consider are invariant under any of the mappings

$$(x_1,\dots,x_n) \to (\varepsilon_1 x_1,\dots,\varepsilon_n x_n), \quad \varepsilon \in \{-1,+1\}^n,$$

and therefore we restrict our attention only to $\mathbb{R}^n_+$ in the following definition.

Definition 1 Let $0 < p \le q \le \infty$, and let $n \ge 2$ and $0 \le m \le n-1$ be natural numbers.

(i) We set

$$\Delta^n_p = \begin{cases} \{(t_1,\dots,t_n) \in \mathbb{R}^n_+ : \sum_{j=1}^n t_j^p = 1\}, & p < \infty, \\ \{(t_1,\dots,t_n) \in \mathbb{R}^n_+ : \max_{j=1,\dots,n} t_j = 1\}, & p = \infty. \end{cases}$$

(ii) Let μ be a Borel probability measure on Δ^n_p. Then

$$\sigma^{p,q}_m(\mu) = \int_{\Delta^n_p} \sigma_m(x)_q \, d\mu(x)$$

is called *average surface best m-term width of $id : \ell^n_p \to \ell^n_q$ with respect to μ*.

(iii) Let ν be a Borel probability measure on $[0,1]\cdot\Delta_p^n$. Then

$$\sigma_m^{p,q}(\nu)=\int_{[0,1]\cdot\Delta_p^n}\sigma_m(x)_q\,d\nu(x)$$

is called *average volume best m-term width of $id:\ell_p^n\to\ell_q^n$ with respect to ν.*

Let us observe that the estimates

$$\sigma_m^{p,q}(\mu)\le\sigma_m^{p,q}\quad\text{and}\quad\sigma_m^{p,q}(\nu)\le\sigma_m^{p,q}$$

follow trivially by Definition 1. Furthermore, the mapping $x\to\sigma_m(x)_q$ is continuous and, therefore, measurable with respect to the Borel measure μ.

1.5.2 Main Results

After introducing the new notion of average best m-term width in Definition 1, we study its behavior for the measures on Δ_p^n that are widely used in the literature. A prominent role among them is played by the so-called *normalized cone measure* given by

$$\mu_p(\mathcal{A})=\frac{\lambda([0,1]\cdot\mathcal{A})}{\lambda([0,1]\cdot\Delta_p^n)},\quad\mathcal{A}\subset\Delta_p^n.$$

In Theorem 7 and Proposition 8, we provide basic estimates of $\sigma_m^{p,q}(\mu_p)$ for $q=\infty$ and $q<\infty$, respectively. Surprisingly enough, it turns out that (7) has its direct counterpart for all $0<p<\infty$. This means (as described above) that the coordinates of a "typical" element of the surface of the ℓ_p^n unit ball are well concentrated around the value $n^{-1/p}$. So, roughly speaking, it is only ℓ_p-normalized noise.

Another well-known probability measure on Δ_p^n is the *normalized surface measure ϱ_p*, cf. Definition 9. We calculate in Lemma 10 the density of ϱ_p with respect to μ_p to be equal to

$$\frac{d\varrho_p}{d\mu_p}(x)=c_{p,n}^{-1}\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2},$$

where

$$c_{p,n}=\int_{\Delta_p^n}\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2}d\mu_p(x)$$

is the normalizing constant. This result (which is a generalization of the work of Naor and Romik [38] to the nonconvex case $0<p<1$) might be of independent interest for the study of the geometry of ℓ_p^n spheres. One observes immediately that if $p<1$ and one or more coordinates of x_i are going to zero, then this density has a polynomial singularity and, therefore, gives more weight to areas closed to coordinate hyperplanes.

We then obtain in Theorem 12 an estimate of $\sigma_0^{p,\infty}(\varrho_p)$ from above. Although the measure ϱ_p concentrates around coordinate hyperplanes, it turns out that the estimate from above of $\sigma_0^{p,\infty}(\mu_p)$ as obtained in Theorem 7 and the estimate of Theorem 12 differ only in the constants involved.

The last part of this paper is devoted to the search for a new probability measure on Δ_p^n that would "promote sparsity" in the sense that the mean value of $\sigma_m(x)_q$ decays rapidly with m. One possible candidate is presented in Definition 13 by introducing a new class of measures $\theta_{p,\beta}$, which are given by their density with respect to the cone measure μ_p:

$$\frac{d\theta_{p,\beta}}{d\mu_p}(x) = c_{p,\beta}^{-1} \cdot \prod_{i=1}^n x_i^\beta, \quad x \in \Delta_p^n,$$

where $c_{p,\beta}$ is a normalizing constant. We refer also to Remark 4 for an equivalent characterization.

We show that for an appropriate choice of β, namely $\beta = p/n - 1$, the estimated value of the m-th largest coefficient of elements of the ℓ_p^n-unit sphere decays exponentially with m. Namely, Theorem 16 provides estimates of $\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})$ that in the end imply that

$$\frac{C_p^1}{(\frac{1}{p}+1)^m} \le \liminf_{n\to\infty} \sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1}) \le \limsup_{n\to\infty} \sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1}) \le \frac{C_p^2}{(\frac{1}{p}+1)^m} \tag{8}$$

for two positive real numbers C_p^1 and C_p^2 that depend only on p.

This result (which is also simulated numerically in the very last section of this paper) is in a certain way independent of n. This offers hope that one could apply this approach also to the infinite-dimensional spaces ℓ_p or, using a suitable discretization technique (like wavelet decomposition), to some function spaces. This remains a subject of our further research.

Of course, the class $\theta_{p,\beta}$ provides only one example of measures with rapid decay of their average best m-term widths. We also leave the detailed study of other measures with such properties open to future work.

Note Added in the Proof Let us comment on the relation of our work with recent papers of Cevher [12] and Gribonval, Cevher, and Davis [29]. Cevher uses in [12] the concept of *Order Statistics* [16] to identify the probability distributions whose independent and identically distributed (i.i.d.) realizations result typically in p-compressible signals, i.e.,

$$x_i^* \le C\,R \cdot i^{-1/p}.$$

Our approach here is a bit different and more connected to the geometry of ℓ_p^n spaces. In accordance with [43], this leads to the study of ℓ_p^n*-normalized* vectors with i.i.d. components. This again allows us to better distinguish between the norm of such a vector (i.e., its *size* or *energy*) and its direction (i.e., its *structure*).

The approach of the recent preprint [29] (which was submitted during the review process of this work) comes much closer to ours. Their Definition 1 of "Compressible priors" introduces the quantity called *relative best m-term approximation error* as

$$\bar{\sigma}_m(x)_q = \frac{\sigma_m(x)_q}{\|x\|_q}, \quad x \in \mathbb{R}^n_+.$$

The asymptotic behavior of this quantity for $x = (x_1, \dots, x_n)$ being a vector with i.i.d. components and $\liminf_{n\to\infty} \frac{m_n}{n} \ge \kappa \in (0,1)$ is then used to define q-compressible probability distribution functions. In contrast to [29], we consider ℓ_q approximation of ℓ_p normalized vectors, and therefore our widths depend on two integrability parameters p and q. Furthermore, we do not pose any restrictions on the ratio m/n to any specific regime and consider the average best m-term widths $\sigma_m^{p,q}(\mu)$ for all $0 \le m \le n-1$. In the only case when we speak about asymptotics (i.e., (32) of Theorem 16), we suppose m to be constant and n growing to infinity. Furthermore, Theorem 1 of [29] shows that all distributions with bounded fourth moment do not fit into their scheme and do not "promote sparsity." Because we are interested in distributions that are connected to the geometry of ℓ_p^n-balls (i.e., generalized Gaussian distribution and generalized Gamma distribution), we change the parameters of the distribution $\theta_{p,\beta}$ in dependence on n. Although quite inconvenient from the mathematical point of view, it is not really clear if this presents a serious obstacle to the application of our approach. But the investigation of this goes beyond the scope of this work.

1.5.3 Structure of the Paper

The paper is structured as follows. The rest of Sect. 1 gives some notation used throughout the paper. Sections 2 and 3 provide estimates of the average best m-term widths with respect to the cone and surface measure, respectively. In Sect. 4, we study a new type of measures on the unit ball of ℓ_p^n. We show that the typical element with respect to those measures behaves in a completely different way compared to the situations discussed before. Those results are illustrated by the numerical experiments described in Sect. 5.

1.6 Notation

We denote by $\mathbb{R}$ the set of real numbers, by $\mathbb{R}_+ := [0,\infty)$ the set of nonnegative real numbers, and by $\mathbb{R}^n$ and $\mathbb{R}^n_+$ their n-fold tensor products. The components of $x \in \mathbb{R}^n$ are denoted by $x_1, \dots, x_n$. The symbol λ stands for the Lebesgue measure on $\mathbb{R}^n$ and $\mathcal{H}$ for the $n-1$ dimensional Hausdorff measure in $\mathbb{R}^n$. If $A \subset \mathbb{R}^n$ and $I \subset \mathbb{R}$ is an interval, we write $I \cdot A := \{tx : t \in I, x \in A\}$.

We shall use very often the *Gamma function*, defined by

$$\Gamma(s) := \int_0^\infty t^{s-1} e^{-t}\, dt, \quad s > 0. \tag{9}$$

In one case, we shall use also the *Beta function*

$$B(p,q) := \int_0^1 t^{p-1}(1-t)^{q-1}\, dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}, \quad p,q > 0 \tag{10}$$

and the *digamma function*

$$\Psi(s) := \frac{d}{ds}\log\Gamma(s) = \frac{\Gamma'(s)}{\Gamma(s)}, \quad s > 0.$$

We recommend [1, Chap. 6] as a standard reference for both basic and more advanced properties of these functions. We shall need Stirling's approximation formula (which was implicitly used already in (7)) in its most simple form

$$\Gamma(x) = \sqrt{\frac{2\pi}{x}}\left(\frac{x}{e}\right)^x \left(1 + \mathcal{O}\left(\frac{1}{x}\right)\right), \quad x > 0. \tag{11}$$

If $a = \{a_j\}_{j=1}^\infty$ and $b = \{b_j\}_{j=1}^\infty$ are real sequences, then $a_j \lesssim b_j$ denotes that there is an absolute constant $C > 0$ such that $a_j \le C\,b_j$ for all $j = 1, 2, \dots$. Similar convention is used for $a_j \gtrsim b_j$ and $a_j \approx b_j$. The capital letter C with indices (i.e., C_p) denotes a positive real number depending only on the highlighted parameters, and their meaning can change from one occurrence to another. If, for any reason, we shall need to distinguish between several numbers of this type, we shall write for example C_p^1 and C_p^2, as was already done in (8).

2 Normalized Cone Measure

In this section, we study the average best m-term widths as introduced in Definition 1 for the most important measure (the so-called cone measure) on Δ_p^n, which is well studied in the literature within the geometry of ℓ_p^n spaces, cf. [4, 5, 37, 38]. Essentially, we recover in Theorem 7 an analog of the estimate (7) for all $0 < p < \infty$.

Definition 2 Let $0 < p \le \infty$ and $n \ge 2$. Then

$$\mu_p(\mathcal{A}) = \frac{\lambda([0,1]\cdot\mathcal{A})}{\lambda([0,1]\cdot\Delta_p^n)}, \quad \mathcal{A} \subset \Delta_p^n$$

is the normalized *cone measure* on Δ_p^n.

If ν_p denotes the p-normalized Lebesgue measure, i.e.,

$$\nu_p(A) = \frac{\lambda(A)}{\lambda([0,1]\cdot\Delta_p^n)}, \quad A \subset \mathbb{R}_+^n,$$

then the connection between ν_p and μ_p is given by

$$\nu_p(A) = n\int_0^\infty r^{n-1}\mu_p\left(\frac{\{x \in A : \|x\|_p = r\}}{r}\right)dr. \tag{12}$$

The proof of (12) follows directly for sets of the type $[a,b]\cdot\mathcal{A}$ with $0 < a < b < \infty$ and $\mathcal{A} \subset \Delta_p^n$ and is then finished by standard approximation arguments. The formula

(12) may be generalized to the so-called *polar decomposition identity*, cf. [4],

$$\frac{\int_{\mathbb{R}^n_+} f(x)\,d\lambda(x)}{\lambda([0,1]\cdot\Delta^n_p)} = n\int_0^\infty r^{n-1}\int_{\Delta^n_p} f(rx)\,d\mu_p(x)\,dr, \tag{13}$$

which holds for every $f \in L_1(\mathbb{R}^n_+)$.

Formula (13) allows for the transfer immediately of the results for the average surface best m-term approximation with respect to μ_p to the average volume approximation with respect to ν_p.

Proposition 3 *The identity*

$$\sigma_m^{p,q}(\nu_p) = \sigma_m^{p,q}(\mu_p)\cdot\frac{n}{n+1}$$

holds for all $0 < p \le q \le \infty$, *all* $n \ge 2$, *and all* $0 \le m \le n-1$.

Proof We plug the function

$$f(x) = \sigma_m(x)_q \cdot \chi_{[0,1]\cdot\Delta^n_p}(x)$$

into (13) and obtain

$$\begin{aligned}&\frac{\int_{[0,1]\cdot\Delta^n_p}\sigma_m(x)_q\,d\lambda(x)}{\lambda([0,1]\cdot\Delta^n_p)}\\ &\quad=\int_{[0,1]\cdot\Delta^n_p}\sigma_m(x)_q\,d\nu_p(x)\\ &\quad= n\int_0^1 r^{n-1}\int_{\Delta^n_p}\sigma_m(rx)_q\,d\mu_p(x)\,dr = n\int_0^1 r^n\,dr\cdot\sigma_m^{p,q}(\mu_p),\end{aligned}$$

which gives the result. □

Proposition 3 shows that the ratio between approximation with respect to μ_p and ν_p is equal to $1+1/n$. This justifies our interest in measures on Δ^n_p. Furthermore, it shows that the quantities $\sigma_m^{p,q}(\nu_p)$ and $\sigma_m^{p,q}(\mu_p)$ behave asymptotically (i.e., for $n\to\infty$) very similarly.

Let $p = 2$, and let $\omega_1,\dots,\omega_n$ be independent normally distributed Gaussian random variables. Then

$$\varrho_2(\mathcal{A}) = \mu_2(\mathcal{A}) = \mathbb{P}\biggl(\frac{(|\omega_1|,\dots,|\omega_n|)}{(\sum_{j=1}^n \omega_j^2)^{1/2}} \in \mathcal{A}\biggr), \quad \mathcal{A}\subset\Delta^n_2.$$

As noted in [43], this relation may be generalized to all values of p with $0 < p < \infty$. Let $\omega_1,\dots,\omega_n$ be independent random variables on $\mathbb{R}_+$ each with density

$$c_p e^{-t^p}, \quad t \ge 0,$$

with respect to the Lebesgue measure, where $c_p = \frac{p}{\Gamma(1/p)} = \frac{1}{\Gamma(1/p+1)}$.

Then, cf. [43, Lemma 1],

$$\mu_p(\mathcal{A}) = \mathbb{P}\biggl(\frac{(\omega_1, \dots, \omega_n)}{(\sum_{j=1}^n \omega_j^p)^{1/p}} \in \mathcal{A}\biggr), \quad \mathcal{A} \subset \Delta_p^n. \tag{14}$$

We now fix $\omega_1, \dots, \omega_n$ until the end of this paper. Also the symbols $\mathbb{E}$ and $\mathbb{P}$ are always taken with respect to these variables.

2.1 The Case $q = \infty$

In this section, we deal with uniform approximation, i.e., with the case $q = \infty$. To be able to imitate the calculation (6), we shall need several tools, which are the subject of Lemmas 4, 5, and 6. Our main result of this section (Theorem 7) then provides the estimate of $\sigma_m^{p,\infty}(\mu_p)$ from above for all m with $0 \le m \le n-1$. Furthermore, it is shown that in the range $0 \le m \le \varepsilon_p n$, this estimate is also optimal.

Lemma 4 *Let* $0 < p < \infty$, *and let* $n \ge 2$ *and* $1 \le m \le n$ *be natural numbers. Then*

$$\int_{\Delta_p^n} x_m^* \, d\mu_p(x) = \frac{\Gamma(n/p)}{\Gamma(n/p + 1/p)} \cdot \mathbb{E} x_m^*.$$

Furthermore, there are two positive real numbers C_p^1 *and* C_p^2 *depending only on* p *such that*

$$C_p^1 \cdot \frac{\mathbb{E} x_m^*}{n^{1/p}} \le \int_{\Delta_p^n} x_m^* \, d\mu_p(x) \le C_p^2 \cdot \frac{\mathbb{E} x_m^*}{n^{1/p}}.$$

Proof We set $f(x) = x_m^* e^{-x_1^p - \dots - x_n^p}$ and use the polar decomposition identity (13):

$$\begin{aligned}\frac{\int_{\mathbb{R}_+^n} x_m^* e^{-x_1^p - \dots - x_n^p} \, d\lambda(x)}{\lambda([0,1] \cdot \Delta_p^n)} &= n \int_0^\infty r^{n-1} \int_{\Delta_p^n} (r x_m^*) \cdot e^{-(rx_1)^p - \dots - (rx_n)^p} \, d\mu_p(x) \, dr \\ &= n \int_0^\infty r^{n-1} \cdot r e^{-r^p} \, dr \int_{\Delta_p^n} x_m^* \, d\mu_p(x),\end{aligned}$$

or, equivalently,

$$\int_{\Delta_p^n} x_m^* \, d\mu_p(x) = \frac{\int_{\mathbb{R}_+^n} x_m^* e^{-x_1^p - \dots - x_n^p} \, d\lambda(x)}{\lambda([0,1] \cdot \Delta_p^n) \cdot n \int_0^\infty r^n e^{-r^p} \, dr}. \tag{15}$$

The identity

$$\int_0^\infty r^n e^{-r^p} dr = \frac{\Gamma(n/p + 1/p)}{p}$$

follows by a simple substitution. Furthermore, we shall need the classical formula of Dirichlet for the volume of the unit ball $B_{\ell_p^n}$ of ℓ_p^n, cf. [21, p. 157],

$$\lambda\big([0,1]\cdot\Delta_p^n\big)=\frac{\lambda(B_{\ell_p^n})}{2^n}=\frac{\Gamma(1/p+1)^n}{\Gamma(n/p+1)}.$$

This allows us to reformulate (15) as

$$\int_{\Delta_p^n} x_m^*\,d\mu_p(x)=\frac{\Gamma(n/p+1)\,\mathbb{E}x_m^*}{c_p^n\cdot n/p\cdot\Gamma(n/p+1/p)\Gamma(1/p+1)^n}=\frac{\Gamma(n/p)\,\mathbb{E}x_m^*}{\Gamma(n/p+1/p)}.$$

Finally, we use Stirling's formula (11) to estimate

$$\frac{n^{1/p}\cdot\Gamma(n/p)}{\Gamma(n/p+1/p)}\le C_p^1\frac{n^{1/p}(n/p)^{n/p-1/2}}{(n/p+1/p)^{n/p+1/p-1/2}}\le C_p^2\bigg(\frac{n}{n+1}\bigg)^{n/p+1/p-1/2}\le C_p^3$$

and similarly for the estimate from below. □

Lemma 5 *Let $\alpha\in\mathbb{R}$ and $\delta>0$. Then*

$$\int_\delta^\infty u^\alpha e^{-u}\,du\le\delta^\alpha e^{-\delta}\cdot\begin{cases}1 & \textit{if } \alpha\le 0,\\ \frac{1}{1-\alpha/\delta} & \textit{if } \alpha>0 \textit{ and } \frac{\alpha}{\delta}<1,\\ (\frac{\alpha}{\delta})^\alpha\cdot\frac{\alpha/\delta}{1-\delta/\alpha} & \textit{if } \alpha>0 \textit{ and } \frac{\alpha}{\delta}>1.\end{cases}$$

Proof If $\alpha\le 0$, we may estimate

$$\int_\delta^\infty u^\alpha e^{-u}\,du\le\delta^\alpha\int_\delta^\infty e^{-u}\,du=\delta^\alpha e^{-\delta}.$$

If $0<\alpha\le 1$, we use partial integration and obtain

$$\int_\delta^\infty u^\alpha e^{-u}\,du=\delta^\alpha e^{-\delta}+\alpha\int_\delta^\infty u^{\alpha-1}e^{-u}\,du\le\delta^\alpha e^{-\delta}\big(1+\alpha\delta^{-1}\big).$$

This is smaller than

$$\delta^\alpha e^{-\delta}\bigg(1+\frac{\alpha}{\delta}+\frac{\alpha^2}{\delta^2}+\cdots\bigg)=\delta^\alpha e^{-\delta}\cdot\frac{1}{1-\alpha/\delta}$$

if $\alpha/\delta<1$ and smaller than

$$\delta^\alpha e^{-\delta}\frac{\alpha}{\delta}\bigg(1+\frac{\delta}{\alpha}+\frac{\delta^2}{\alpha^2}+\cdots\bigg)=\delta^\alpha e^{-\delta}\frac{\alpha}{\delta}\cdot\frac{1}{1-\delta/\alpha}$$

if $\alpha/\delta>1$.

If $k-1<\alpha\le k$ for some $k\in\mathbb{N}$, we iterate the partial integration and arrive at

$$\begin{aligned}
&\int_\delta^\infty u^\alpha e^{-u}\,du\\
&\quad\le \delta^\alpha e^{-\delta}\bigl(1+\alpha\delta^{-1}+\alpha(\alpha-1)\delta^{-2}+\cdots+\alpha(\alpha-1)\cdots(\alpha-k+1)\delta^{-k}\bigr)\\
&\quad\le \delta^\alpha e^{-\delta}\left(1+\frac{\alpha}{\delta}+\frac{\alpha^2}{\delta^2}+\cdots+\frac{\alpha^k}{\delta^k}\right)\\
&\quad\le \delta^\alpha e^{-\delta}\begin{cases}\frac{1}{1-\alpha/\delta} & \text{if } \alpha/\delta<1,\\ (\frac{\alpha}{\delta})^{\alpha+1}\frac{1}{1-\delta/\alpha} & \text{if } \alpha/\delta>1.\end{cases}
\end{aligned}$$

□

Lemma 6 *Let* $0<p<\infty$. *Then there is a positive real number* C_p *such that*

$$\mathbb{E}\,x_m^*\le C_p\log^{1/p}\left(\frac{en}{m}\right)$$

for all $1\le m\le n$.

Proof We estimate

$$\begin{aligned}
\mathbb{E}\,x_m^* &= \int_0^\infty \mathbb{P}\bigl(\omega_m^*>t\bigr)\,dt=\delta+\int_\delta^\infty \mathbb{P}\bigl(\omega_m^*>t\bigr)\,dt\\
&\le \delta+\binom{n}{m}\int_\delta^\infty \mathbb{P}(\omega_1>t,\omega_2>t,\dots,\omega_m>t)\,dt\\
&= \delta+\binom{n}{m}\int_\delta^\infty \mathbb{P}(\omega_1>t)^m\,dt. \qquad (16)
\end{aligned}$$

The parameter $\delta>\max(1,3(1/p-1))^{1/p}$ is to be chosen later on. We substitute $v=u^p$ and obtain

$$\mathbb{P}(\omega_1>t)=c_p\int_t^\infty e^{-u^p}\,du=\frac{c_p}{p}\int_{t^p}^\infty v^{1/p-1}e^{-v}\,dv.$$

Using the first two estimates of Lemma 5 (recall that $t^p\ge\delta^p>\max(1,3(1/p-1))$), we arrive at

$$\mathbb{P}(\omega_1>t)\le C_p t^{1-p}e^{-t^p},$$

where C_p depends only on p. We plug this estimate into (16) and obtain

$$\mathbb{E}\,x_m^*\le\delta+\binom{n}{m}(C_p)^m\int_\delta^\infty t^{m(1-p)}e^{-mt^p}\,dt. \qquad (17)$$

If $p\ge1$, then

$$\begin{aligned}
\int_\delta^\infty t^{m(1-p)}e^{-mt^p}\,dt&\le\delta^{m(1-p)}\int_\delta^\infty e^{-mt^p}\,dt\le\delta^{m(1-p)}\int_{m\delta^p}^\infty e^{-u}u^{1/p-1}\,du\\
&\le e^{-m\delta^p}.
\end{aligned}$$

Altogether, we obtain

$$\mathbb{E}x_m^* \le \delta + \binom{n}{m}(C_p)^m e^{-m\delta^p}.$$

Using $\binom{n}{m} \le (\frac{en}{m})^m$ and choosing $\delta = C_p' \ln(\frac{en}{m})^{1/p}$ finishes the proof.

If $p < 1$, we use again the second estimate of Lemma 5,

$$\begin{aligned}\int_\delta^\infty t^{m(1-p)} e^{-mt^p}\,dt &= \frac{1}{mp}\cdot m^{(1/p-1)(m+1)} \int_{m\delta^p}^\infty u^{(1/p-1)(m+1)} e^{-u}\,du \\ &\le \frac{1}{mp}\cdot \delta^{(1-p)(m+1)} e^{-m\delta^p}\cdot \frac{1}{1-\frac{2(1/p-1)}{\delta^p}} \\ &\le C_p' \delta^{(1-p)(m+1)} e^{-m\delta^p}.\end{aligned}$$

Using (17) and $\binom{n}{m} \le (\frac{en}{m})^m$ again, we get

$$\begin{aligned}\mathbb{E}x_1^* &\le \delta + \exp\bigl(-m\delta^p + m\ln(en/m) + (1-p)(m+1)\ln\delta + m\ln C_p + \ln C_p'\bigr) \\ &\le \delta + \exp\bigl[-m\bigl(\delta^p + C_p\ln(en/m) + 2(1-p)\ln\delta\bigr)\bigr].\end{aligned}$$

The choice $\delta = C_p' \ln(\frac{en}{m})^{1/p}$ with C_p' large enough ensures that

$$\frac{\delta^p}{2} \ge C_p \ln(en/m) \quad \text{and} \quad \frac{\delta^p}{2} \ge 2(1-p)\ln\delta$$

and finishes the proof. □

The following theorem gives the basic estimates of $\sigma_m^{p,\infty}(\mu_p)$.

Theorem 7 *Let* $0 < p \le \infty$, *and let* $n \ge 2$.

(i) *Let* $0 \le m \le n-1$. *Then*

$$\sigma_m^{p,\infty}(\mu_p) \le C_p \left[\frac{\log(\frac{en}{m+1})}{n}\right]^{1/p}.$$

(ii) *There is a number* $0 < \varepsilon_p < 1$ *such that for* $0 \le m \le \varepsilon_p n$, *the following estimate holds*:

$$\sigma_m^{p,\infty}(\mu_p) \ge C_p \left[\frac{\log(\frac{en}{m+1})}{n}\right]^{1/p}.$$

Proof Lemmas 4 and 6 imply immediately the first part of the theorem if $p < \infty$. If $p = \infty$, the proof is trivial.

The proof of the second part is divided into two steps.

Step 1. We start first with the case $m = 0$.

If $p = \infty$, then $x_1^* = 1$ for all $x \in \Delta_p^n$ and the proof is trivial. Let us therefore assume that $p < \infty$. According to Lemma 4, we have to estimate $\mathbb{E}x_1^*$ from below.

This was done in [43, Lemma 2]. We include a slightly different proof for the reader's convenience. For every $t_0 > 0$, we have

$$\mathbb{E}\, x_1^* \geq t_0\, \mathbb{P}\bigl(x_1^* > t_0\bigr) = t_0\, \mathbb{P}\Bigl(\max_{1\leq j\leq n} x_j > t_0\Bigr) \geq t_0 \left[n\mathbb{P}(x_1 > t_0) - \binom{n}{2}\mathbb{P}(x_1 > t_0)^2 \right].$$

We define t_0 by $\mathbb{P}(x_1 > t_0) = \frac{1}{n}$ and obtain $\mathbb{E}\, x_1^* \geq t_0/2$.

From the simple estimate

$$\frac{c_p}{p}\int_{T^p}^{\infty} u^{1/p-1} e^{-u} du \geq C_p e^{-2T^p}, \quad T > 1,$$

it follows that there is a positive real number $\gamma_p > 0$ such that

$$\mathbb{P}\bigl(x_1 > \gamma_p \bigl(\log(en)\bigr)^{1/p}\bigr) \geq 1/n.$$

This gives $t_0 \geq \gamma_p (\log(en))^{1/p}$ and $\mathbb{E}\, x_1^* \geq C_p (\log(en))^{1/p}$.

Step 2. Let $0 \leq m \leq \varepsilon_p n$, where $\varepsilon_p > 0$ will be chosen later on.

We shall use the inequality

$$\frac{1}{m}\sum_{j=1}^{m} \log^{1/p}\left(\frac{en}{j}\right) \leq C_p \log^{1/p}\left(\frac{en}{m}\right), \quad 1 \leq m \leq n, \tag{18}$$

which follows by direct calculation for $p = 1$, by Hölder's inequality for $1 < p < \infty$, and by replacing the sum by the corresponding integral and integration by parts if $0 < p < 1$.

We write

$$\|x\|_{(m)} = \frac{1}{m}\sum_{j=1}^{m} x_j^*.$$

By Lemma 6 and (18),

$$\mathbb{E}\,\|x\|_{(m)} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}\, x_j^* \leq \frac{C_p}{m}\sum_{j=1}^{m} \log^{1/p}\left(\frac{en}{j}\right) \leq C_p^1 \log^{1/p}\left(\frac{en}{m}\right). \tag{19}$$

To estimate $\mathbb{E}\,\|x\|_{(m)}$ from below, we assume that $1 \leq m \leq n$ and that n/m is an integer (otherwise one has to slightly modify the argument at the cost of the constants involved). We partition the set $\{1, \dots, n\} = A_1 \cup \cdots \cup A_m$, where each one of the disjoint sets A_j has n/m elements. Then we have

$$\|x\|_{(m)} \geq \frac{1}{m}\sum_{j=1}^{m} \max_{l\in A_j} x_l,$$

and by the first step we obtain

$$\mathbb{E}\,\|x\|_{(m)} \geq \frac{1}{m}\sum_{j=1}^{m} \mathbb{E} \max_{l\in A_j} x_l \geq C_p^2 \log^{1/p}\left(\frac{en}{m}\right). \tag{20}$$

Let $N_p < 1/\varepsilon_p$ be a natural number to be chosen later on. Combining (19) with (20) gives finally

$$\begin{aligned}\mathbb{E}\, x_m^* &\geq \frac{1}{N_p m}\sum_{k=m}^{N_p m}\mathbb{E}\, x_k^* \geq \mathbb{E}\,\|x\|_{(N_p m)} - \frac{1}{N_p}\mathbb{E}\,\|x\|_{(m)}\\ &\geq C_p^2 \log^{1/p}\left(\frac{en}{N_p m}\right) - \frac{C_p^1}{N_p}\log^{1/p}\left(\frac{en}{m}\right)\\ &= \log^{1/p}\left(\frac{en}{m}\right)\left\{C_p^2\left[1-\frac{\log(N_p)}{\log(\frac{en}{m})}\right]^{1/p} - \frac{C_p^1}{N_p}\right\}.\end{aligned}$$

An appropriate choice of N_p and ε_p (i.e., $N_p > 2^{1/p}C_p^1/C_p^2$ and $\varepsilon_p < \min(1/N_p, e/N_p^2)$) with

$$C_p^2\left[1-\frac{\log(N_p)}{\log(\frac{e}{\varepsilon_p})}\right]^{1/p} - \frac{C_p^1}{N_p} > 0$$

gives the result. □

Remark 1

(i) Theorem 7 provides basic estimates of average best m-term widths $\sigma_m^{p,\infty}(\mu_p)$. In the case $m = 0$, a stronger result on concentration of μ_p was obtained already in [43, Theorem 3 and Remark 2]. It would certainly be of interest to obtain a similar statement also for other values of $m > 0$, but this would go beyond the scope of this paper, and we leave this direction open for further study.

(ii) Theorem 7 may be interpreted in the sense of the discussion after formula (7). Namely, the average coordinate of $x \in \Delta_p^n$ is $n^{-1/p}$. Theorem 7 shows that the average value of the largest coordinate is only slightly larger (namely $c[\ln(en)]^{1/p}$ times larger). In this sense, the average point of Δ_p^n is only slightly modified (and properly normalized) white noise.

(iii) Using the interpolation formula (4), one may immediately extend this result to all $0 < p \leq q < \infty$. But we shall see later on that in the case $q < \infty$, one may prove slightly better estimates.

(iv) The behavior of $\sigma_m^{p,\infty}(\mu_p)$ was studied in detail in [28, Example 10] for $p = 2$. It was shown that if x_i are independent $N(0,1)$ Gaussian random variables and $m \leq n/2 + 1$, then

$$c\sqrt{\ln\frac{2n}{m}} \leq \mathbb{E}\, x_m^* \leq C\sqrt{\ln\frac{2n}{m}},$$

where c and C are absolute positive constants. Furthermore, if $m \geq n/2 + 1$, then

$$\sqrt{\frac{\pi}{2}}\,\frac{n-m+1}{n+1} \leq \mathbb{E}\, x_m^* \leq \sqrt{2\pi}\,\frac{n-m+1}{n}.$$

(v) The method used in the proof of the second part of Theorem 7 may be found, for example, in [27].

2.2 The Case $q < \infty$

We discuss briefly also the case when $q < \infty$. It turns out that in this case the logarithmic term disappears. We do not go much into details and restrict ourselves to the case $m = 0$.

Proposition 8 *Let $n \geq 2$ and $0 < p \leq q < \infty$. Then*

(i) $C^1_{p,q} n^{1/q} \leq \mathbb{E}\,\|x\|_q \leq C^2_{p,q} n^{1/q}$,

(ii)

$$C^1_{p,q} \cdot \frac{\mathbb{E}\,\|x\|_q}{n^{1/p}} \leq \sigma_0^{p,q}(\mu_p) = \int_{\Delta_p^n} \|x\|_q \, d\mu_p(x) \leq C^2_{p,q} \cdot \frac{\mathbb{E}\,\|x\|_q}{n^{1/p}},$$

and

(iii) $C^1_{p,q} n^{1/q-1/p} \leq \sigma_0^{p,q}(\mu_p) \leq C^2_{p,q} n^{1/q-1/p}$,

where in all these estimates C^1_p and C^2_p are positive real numbers depending only on p.

Proof

(i) The following two inequalities may be easily proved by Hölder's and Minkowski's inequalities:

$$\left(\sum_{j=1}^n (\mathbb{E}x_j)^q\right)^{1/q} \leq \mathbb{E}\left(\sum_{j=1}^n x_j^q\right)^{1/q} \leq \left(\sum_{j=1}^n \mathbb{E}x_j^q\right)^{1/q}, \quad q \geq 1,$$

$$\left(\sum_{j=1}^n \mathbb{E}x_j^q\right)^{1/q} \leq \mathbb{E}\left(\sum_{j=1}^n x_j^q\right)^{1/q} \leq \left(\sum_{j=1}^n (\mathbb{E}x_j)^q\right)^{1/q}, \quad q \leq 1.$$

This gives for $q \geq 1$,

$$\mathbb{E}\|x\|_q \leq n^{1/q}\big(\mathbb{E}x_j^q\big)^{1/q} \quad \text{and} \quad \mathbb{E}\|x\|_q \geq n^{1/q}\mathbb{E}x_j,$$

and for $q \leq 1$,

$$\mathbb{E}\|x\|_q \leq n^{1/q}\mathbb{E}x_j \quad \text{and} \quad \mathbb{E}\|x\|_q \geq n^{1/q}\big(\mathbb{E}x_j^q\big)^{1/q}.$$

Let us note that the value of $\mathbb{E}x_j$ and $(\mathbb{E}x_j^q)^{1/q}$ does not depend on n, only on p and q.

(ii) The proof of the second part resembles very much the proof of Lemma 4 and is left to the reader.

(iii) The last point follows immediately from (i) and (ii). □

Remark 2 A statement similar to Proposition 8 is included in [43, Lemma 2, point 4].

3 Normalized Surface Measure

In this section, we study the average best m-term widths for another classical measure on Δ_p^n, namely the normalized Hausdorff measure, cf. Definition 9. Intuitively, it would seem that this measure gives more weight to those areas where one or more components of $x \in \Delta_p^n$ are close to zero. It turns out that this really is the case, and the mathematical formulation is given in Lemma 10 below. This relation is then used together with Lemma 11 in Theorem 12 to provide estimates of $\sigma_0^{p,\infty}(\varrho_p)$ from above.

Definition 9 Let $n \geq 2$ be a natural number. We denote by

$$\varrho_p(\mathcal{A}) = \frac{\mathcal{H}(\mathcal{A})}{\mathcal{H}(\Delta_p^n)}, \quad \mathcal{A} \subset \Delta_p^n,$$

the normalized $n-1$ dimensional Hausdorff measure on Δ_p^n.

Let us mention that for $p \in \{1, 2, \infty\}$, the measure ϱ_p coincides with μ_p. The following lemma provides a relationship between the normalized surface measure ϱ_p and the cone measure μ_p. For $p \geq 1$, it was given by [38]. We follow closely their approach, and it turns out that it may be generalized also to the nonconvex case of $0 < p < 1$.

Lemma 10 *Let $0 < p < \infty$ and $n \geq 2$. Then ϱ_p is an absolutely continuous measure with respect to μ_p and for μ_p almost every $x \in \Delta_p^n$, we have*

$$\frac{d\varrho_p}{d\mu_p}(x) = \frac{n\lambda([0,1]\cdot\Delta_p^n)}{\mathcal{H}(\Delta_p^n)} \big\| \nabla\big(\|\cdot\|_p\big)(x)\big\|_2 = c_{p,n}^{-1}\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2},$$

where

$$c_{p,n} = \int_{\Delta_p^n} \left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2} d\mu_p(x)$$

is the normalizing constant.

Proof The proof follows the proof of [38, Lemmas 1 and 2], where the statement was proved for $1 \leq p < \infty$. Hence, we may assume that $0 < p < 1$. First, we introduce some notation.

We fix $x = (x_1, \dots, x_n) \in \Delta_p^n$ such that:

- the mapping $y \to \|y\|_p$ is differentiable at x,
- x is a density point of $\mathcal{H}$, i.e.,

$$\lim_{\varepsilon \to 0+} \frac{\mathcal{H}(B(x,\varepsilon) \cap \Delta_p^n)}{\varepsilon^{n-1} V_{n-1}} = 1, \tag{21}$$

  where V_{n-1} denotes the Lebesgue volume of the $n-1$ dimensional Euclidean unit ball,
- $x_i > 0$ for all $i = 1, \dots, n$.

Obviously, ϱ_p-almost every $x \in \Delta_p^n$ satisfies all the three properties (we refer, for example, to [34, Theorem 16.2] for the second one).

Furthermore, we set $z := \nabla(\|\cdot\|_p)(x)$. This means that

$$\|x+y\|_p = 1 + \langle z, y\rangle + r(y), \tag{22}$$

where

$$\theta(\delta) := \sup\left\{\frac{|r(y)|}{\|y\|_2} : 0 < \|y\|_2 \le \delta\right\}, \quad \delta > 0$$

tends to zero if δ tends to zero. Using (22) for $y = \delta x$, one observes that $\langle z, x\rangle = 1$. We denote by $H = x + z^{\perp}$ the tangent hyperplane to Δ_p^n at x. Let us note that for $0 < p < 1$, the set $\mathbb{R}_+^n \setminus [0,1)\cdot\Delta_p^n = [1,\infty)\cdot\Delta_p^n$ is convex. Next, we show that $\langle z, y\rangle \ge 1$ for every $y \in [1,\infty)\cdot\Delta_p^n$. Indeed,

$$\begin{aligned}1 &\le \big\|x + \lambda(y-x)\big\|_p = 1 + \big\langle z, \lambda(y-x)\big\rangle + r\big(\lambda(y-x)\big)\\ &= 1 - \lambda + \lambda\langle z, y\rangle + r\big(\lambda(y-x)\big).\end{aligned}$$

Dividing by $\lambda > 0$ and letting $\lambda \to 0$ gives the statement.

The proof of the lemma is based on the following two inclusions, namely

$$[0,1]\cdot\big(B\big(x, \varepsilon\big(1-\theta(\varepsilon)\big)\big) \cap H\big) \subset [0,1]\cdot\big(B(x,\varepsilon)\cap\Delta_p^n\big) \tag{23}$$

and

$$[0,1]\cdot\big(B(x,\varepsilon)\cap\Delta_p^n\big) \subset \big[0, 1+\varepsilon\theta(\varepsilon)\big]\cdot\big(B\big(x,\varepsilon\big(1+\theta(\varepsilon)\|x\|_2\big)\big)\cap H\big), \tag{24}$$

which hold for all $\varepsilon > 0$ small enough.

First, we prove (23). Given $0 \le s \le 1$ and $v \in B(x, \varepsilon(1-\theta(\varepsilon)) \cap H$, we need to find $0 \le t \le 1$ and $w \in B(x,\varepsilon)\cap\Delta_p^n$ such that $sv = tw$. To do this, we set

$$w := \frac{v}{\|v\|_p} \in \Delta_p^n \quad \text{and} \quad t := s\|v\|_p.$$

We need to show that $t \le 1$ and $\|x - w\|_2 \le \varepsilon$.

We choose $0 < \varepsilon \le \min_i x_i$. Then

$$x_i \le |x_i - v_i| + v_i \le \|x - v\|_2 + v_i \le \varepsilon + v_i$$

for every $i = 1, \dots, n$, which implies that $v_i \ge 0$ and $v \in \mathbb{R}_+^n$. From $v \in H$ and $v \in \mathbb{R}_+^n$, we deduce that $\|v\|_p \le 1$. Hence $t = s\|v\|_p \le \|v\|_p \le 1$.

Next, we write

$$\begin{aligned}\|x - w\|_2 &= \left\|x - \frac{v}{\|v\|_p}\right\|_2 \le \|x - v\|_2 + \left\|v - \frac{v}{\|v\|_p}\right\|_2\\ &\le \varepsilon\big(1-\theta(\varepsilon)\big) + \|v\|_2 \cdot \frac{1-\|v\|_p}{\|v\|_p} \le \varepsilon\big(1-\theta(\varepsilon)\big) + 1 - \|v\|_p\end{aligned}$$

$$\begin{aligned}&=\varepsilon\bigl(1-\theta(\varepsilon)\bigr)+1-\bigl\{1+\langle v-x,z\rangle+r(v-x)\bigr\}\\&=\varepsilon\bigl(1-\theta(\varepsilon)\bigr)+r(v-x)\le\varepsilon.\end{aligned}$$

Next, we prove (24). We need to find for given $0\le t\le 1$ and $w\in B(x,\varepsilon)\cap\Delta_p^n$ some $0\le s\le 1+\varepsilon\theta(\varepsilon)$ and $v\in B(x,\varepsilon(1+\theta(\varepsilon)\|x\|_2))\cap H$ such that $tw=sv$. We put

$$s:=t\langle w,z\rangle\quad\text{and}\quad v:=\frac{w}{\langle w,z\rangle}.$$

Let us recall that we have shown above that $w\in\Delta_p^n$ implies that $\langle w,z\rangle\ge 1$.

Of course, $tw=sv$ and $v\in H$ (as $\langle v,z\rangle=1$). Hence, it remains to show that $s\le 1+\varepsilon\theta(\varepsilon)$ and $\|v-x\|_2\le\varepsilon(1+\theta(\varepsilon)\|x\|_2)$.

The application of (22) gives

$$1=\|w\|_p=\bigl\|x+(w-x)\bigr\|_p=1+\langle w-x,z\rangle+r(w-x),$$

which again forces $\langle w,z\rangle\le 1+\varepsilon\theta(\varepsilon)$. Then $s=t\langle w,z\rangle\le\langle w,z\rangle\le 1+\varepsilon\theta(\varepsilon)$.

Finally, we write

$$\begin{aligned}\|v-x\|_2&=\left\|\frac{w}{\langle w,z\rangle}-x\right\|_2\le\left\|\frac{w}{\langle w,z\rangle}-\frac{x}{\langle w,z\rangle}\right\|_2+\left\|\frac{x}{\langle w,z\rangle}-x\right\|_2\\&\le\frac{\|w-x\|_2}{\langle w,z\rangle}+\|x\|_2\frac{\langle w,z\rangle-1}{\langle w,z\rangle}\le\varepsilon+\varepsilon\theta(\varepsilon)\|x\|_2.\end{aligned}$$

Equipped with (23) and (24), we may finish the proof of the lemma. We write

$$\begin{aligned}&\lim_{\varepsilon\to 0}\frac{\varrho_p(B(x,\varepsilon)\cap\Delta_p^n)}{\mu_p(B(x,\varepsilon)\cap\Delta_p^n)}\\&\quad=\lim_{\varepsilon\to 0}\frac{\mathcal{H}(B(x,\varepsilon)\cap\Delta_p^n)}{\mathcal{H}(\Delta_p^n)}\cdot\frac{\varepsilon^{n-1}V_{n-1}}{\varepsilon^{n-1}V_{n-1}}\cdot\frac{\lambda([0,1]\cdot\Delta_p^n)}{\lambda([0,1]\cdot[B(x,\varepsilon)\cap\Delta_p^n])}\\&\quad=\frac{\lambda([0,1]\cdot\Delta_p^n)}{\mathcal{H}(\Delta_p^n)}\cdot\lim_{\varepsilon\to 0}\frac{\varepsilon^{n-1}V_{n-1}}{\lambda([0,1]\cdot[B(x,\varepsilon)\cap\Delta_p^n])},\end{aligned}\tag{25}$$

where we have used (21). As the perpendicular distance between zero and H is equal to $1/\|z\|_2$, we observe that

$$\operatorname{vol}\bigl(B(x,a)\cap H\bigr)=\frac{a^{n-1}V_{n-1}}{n\|z\|_2}$$

holds for every $a>0$. Using this, we get from (23) and (24),

$$\begin{aligned}&\lambda\bigl([0,1]\cdot\bigl(B\bigl(x,\varepsilon\bigl(1-\theta(\varepsilon)\bigr)\bigr)\cap H\bigr)\bigr)\\&\quad=\frac{[\varepsilon(1-\theta(\varepsilon))]^{n-1}V_{n-1}}{n\|z\|_2}\end{aligned}$$

$$\begin{aligned}
&\le \lambda\big([0,1]\cdot\big(B(x,\varepsilon)\cap\Delta_p^n\big)\big)\\
&\le \lambda\big(\big[0,1+\varepsilon\theta(\varepsilon)\big]\cdot\big(B\big(x,\varepsilon\big(1+\theta(\varepsilon)\|x\|_2\big)\big)\cap H\big)\big)\\
&=\big[1+\varepsilon\theta(\varepsilon)\big]^n\cdot\frac{[\varepsilon(1+\theta(\varepsilon)\|x\|_2)]^{n-1}V_{n-1}}{n\|z\|_2}.
\end{aligned}$$

Combining these estimates with (25) gives the result. □

The following lemma is analogous to Lemma 4 and reduces the calculation of $\sigma_0^{p,\infty}(\varrho_p)$ to inequalities for the estimated values of functions of the random variables $x_1,\dots,x_n$.

Lemma 11 *Let $0<p<\infty$. There exist two positive real numbers C_p^1 and C_p^2 such that*

$$\begin{aligned}
C_p^1\cdot\frac{\mathbb{E}\,x_1^*(\sum_{i=1}^n x_i^{2p-2})^{1/2}}{\mathbb{E}(\sum_{i=1}^n x_i^{2p-2})^{1/2}}\cdot n^{-1/p}&\le\sigma_0^{p,\infty}(\varrho_p)=\int_{\Delta_p^n}x_1^*\,d\varrho_p\\
=\frac{\int_{\Delta_p^n}x_1^*(\sum_{i=1}^n x_i^{2p-2})^{1/2}d\mu_p(x)}{\int_{\Delta_p^n}(\sum_{i=1}^n x_i^{2p-2})^{1/2}d\mu_p(x)}&\le C_p^2\frac{\mathbb{E}\,x_1^*(\sum_{i=1}^n x_i^{2p-2})^{1/2}}{\mathbb{E}(\sum_{i=1}^n x_i^{2p-2})^{1/2}}\cdot n^{-1/p}
\end{aligned}$$

for all $n\ge 2$.

Proof Only the inequalities need a proof. It resembles the proof of Lemma 4 and is again based on the polar decomposition formula (13).

We plug the functions

$$f_1(x)=x_1^*\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2}e^{-x_1^p-\cdots-x_n^p}\quad\text{and}\quad f_2(x)=\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2}e^{-x_1^p-\cdots-x_n^p}$$

into (13) and obtain

$$\begin{aligned}
\sigma_0^{p,\infty}(\varrho_p)&=\frac{\int_{\mathbb{R}_+^n}f_1(x)\,dx\cdot\int_0^\infty r^{n+p-2}e^{-r^p}\,dr}{\int_{\mathbb{R}_+^n}f_2(x)\,dx\cdot\int_0^\infty r^{n+p-1}e^{-r^p}\,dr}\\
&=\frac{\mathbb{E}\,x_1^*(\sum_{i=1}^n x_i^{2p-2})^{1/2}}{\mathbb{E}(\sum_{i=1}^n x_i^{2p-2})^{1/2}}\cdot\frac{\Gamma(n/p+1-1/p)}{\Gamma(n/p+1)}.
\end{aligned}$$

By Stirling's formula, the last expression is equivalent to $n^{-1/p}$ with constants of equivalence depending only on p. □

Theorem 12 *Let $0<p<\infty$. Then there is a positive real number C_p such that*

$$\sigma_0^{p,\infty}(\varrho_p)\le C_p\left[\frac{\log(n+1)}{n}\right]^{1/p}$$

for all $n\ge 2$.

Proof We define a probability measure $\alpha_{p,n}$ on $\mathbb{R}_n^+$ by the density

$$\tilde{c}_{p,n}^{-1}\cdot\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2} e^{-x_1^p-\cdots-x_n^p}, \qquad \tilde{c}_{p,n} := \int_{\mathbb{R}_+^n}\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2} e^{-x_1^p-\cdots-x_n^p}\,dx$$

with respect to the Lebesgue measure. Let us note that due to the inequality

$$\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2} \le \sum_{i=1}^n x_i^{p-1},$$

the integral in the definition of $\tilde{c}_{p,n}$ really converges, and $\alpha_{p,n}$ is well defined.

According to Lemma 11, we need to estimate

$$\int_{\mathbb{R}_+^n} x_1^*\,d\alpha_{p,n}(x).$$

We calculate for $\delta > 1$, which is to be chosen later on,

$$\begin{aligned}\int_{\mathbb{R}_+^n} x_1^*\,d\alpha_{p,n}(x) &= \int_0^\infty \alpha_{p,n}\big(x_1^* > t\big)\,dt \le \delta + \int_\delta^\infty \alpha_{p,n}\big(x_1^* > t\big)\,dt \\ &\le \delta + n\int_\delta^\infty \alpha_{p,n}(x_1 > t)\,dt.\end{aligned}$$

We write $x' = (x_2, \dots, x_n) \in \mathbb{R}_+^{n-1}$. Then

$$\begin{aligned}\alpha_{p,n}(x_1 > t) &= \tilde{c}_{p,n}^{-1}\int_t^\infty e^{-x_1^p}\int_{\mathbb{R}_+^{n-1}}\left(\sum_{i=1}^n x_i^{2p-2}\right)^{1/2} e^{-x_2^p-\cdots-x_n^p}\,dx'\,dx_1 \\ &\le \tilde{c}_{p,n}^{-1}\int_t^\infty e^{-x_1^p}\int_{\mathbb{R}_+^{n-1}}\left[x_1^{p-1} + \left(\sum_{i=2}^n x_i^{2p-2}\right)^{1/2}\right] e^{-x_2^p-\cdots-x_n^p}\,dx'\,dx_1 \\ &= \tilde{c}_{p,n}^{-1}\int_t^\infty e^{-x_1^p} x_1^{p-1}\,dx_1 \cdot \int_{\mathbb{R}_+^{n-1}} e^{-x_2^p-\cdots-x_n^p}\,dx' \\ &\quad + \tilde{c}_{p,n}^{-1}\int_t^\infty e^{-x_1^p}\,dx_1 \cdot \int_{\mathbb{R}_+^{n-1}}\left(\sum_{i=2}^n x_i^{2p-2}\right)^{1/2} e^{-x_2^p-\cdots-x_n^p}\,dx' \\ &:= I_1 + I_2.\end{aligned}$$

The inequality

$$\begin{aligned} c_p^n \tilde{c}_{p,n} &= c_p^n \int_{\mathbb{R}_+^n} \left(\sum_{i=1}^n x_i^{2p-2} \right)^{1/2} e^{-x_1^p - \cdots - x_n^p} \, dx \\ &\geq c_p^n \int_{\mathbb{R}_+^n} \left(\sum_{i=2}^n x_i^{2p-2} \right)^{1/2} e^{-x_1^p - \cdots - x_n^p} \, dx \\ &= c_p^n \int_0^\infty e^{-x_1^p} \, dx_1 \int_{\mathbb{R}_+^{n-1}} \left(\sum_{i=2}^n x_i^{2p-2} \right)^{1/2} e^{-x_2^p - \cdots - x_n^p} \, dx' \\ &= c_p^{n-1} \tilde{c}_{p,n-1} \end{aligned} \tag{26}$$

shows that

$$I_1 = \frac{c_p \int_t^\infty x_1^{p-1} e^{-x_1^p} \, dx_1}{c_p^n \tilde{c}_{p,n}} \leq \frac{c_p \int_t^\infty x_1^{p-1} e^{-x_1^p} \, dx_1}{c_p \tilde{c}_{p,1}} = \tilde{c}_{p,1}^{-1} \cdot \frac{e^{-t^p}}{p}.$$

Using (26) again, we also get

$$I_2 = \tilde{c}_{p,n}^{-1} \cdot \tilde{c}_{p,n-1} \int_t^\infty e^{-x_1^p} \, dx_1 \leq c_p \int_t^\infty e^{-x_1^p} \, dx_1 = \frac{c_p}{p} \cdot \int_{t^p}^\infty s^{1/p-1} e^{-s} \, ds.$$

If $p \geq 1$, we get

$$I_1 + I_2 \leq C_p e^{-t^p}, \quad t > 1, \tag{27}$$

and

$$\int_{\mathbb{R}_+^n} x_1^* \, d\alpha_{p,n}(x) \leq \delta + C_p n \int_\delta^\infty e^{-t^p} \, dt \leq \delta + C_p' n e^{-\delta^p}.$$

By choosing $\delta = C_p \log(n+1)^{1/p}$, we get the result.

If $p < 1$, we use the second estimate of Lemma 5 and replace (27) with

$$I_1 + I_2 \leq C_p t^{1-p} e^{-t^p}, \quad t > t_0,$$

for $t_0 > 1$ large enough, and the result again follows by the choice of δ.

□

Remark 3

(i) Theorem 12 shows that the average size of the largest coordinate of $x \in \Delta_p^n$ taken with respect to the normalized Hausdorff measure is again only slightly larger than $n^{-1/p}$. Hence, also in this case, the typical element of Δ_p^n seems to be far from being sparse and resembles rather properly normalized white noise in the sense described in introduction.

(ii) Using interpolation inequality (4), one may again obtain a similar estimate also for $0 < p \leq q < \infty$, namely

$$\sigma_0^{p,q}(\varrho_p) \leq C_{p,q} \left[\frac{\log(n+1)}{n} \right]^{1/p-1/q}.$$

It would probably be possible to avoid the logarithmic terms and provide improved estimates also for $m > 0$, but we shall not go in this direction. Our main aim of this section was to show that normalized Hausdorff measure does not prefer sparse (or nearly sparse) vectors, and this was clearly demonstrated by Theorem 12.

4 Tensor Product Measures

As discussed already in the introduction and proved in Theorems 7 and 12, the average vectors of Δ_p^n with respect to the cone measure μ_p and with respect to surface measure ϱ_p behave "badly," meaning that (roughly speaking) many of their coordinates are approximately of the same size. As promised before, we shall now introduce a new class of measures for which the random vector behaves in a completely different way. These measures are defined through their density with respect to the cone measure μ_p. This density has a strong singularity near the points with vanishing coordinates.

Definition 13 Let $0 < p < \infty$, $\beta > -1$, and $n \geq 2$. Then we define the probability measure $\theta_{p,\beta}$ on Δ_p^n by

$$\frac{d\theta_{p,\beta}}{d\mu_p}(x) = c_{p,\beta}^{-1} \cdot \prod_{i=1}^n x_i^\beta, \quad x \in \Delta_p^n, \tag{28}$$

where

$$c_{p,\beta} = \int_{\Delta_p^n} \prod_{i=1}^n x_i^\beta \, d\mu_p(x). \tag{29}$$

Remark 4

(i) If $0 > \beta > -1$, then (28) defines the density of $\theta_{p,\beta}$ with respect to μ_p only for points where $x_i \neq 0$ for all $i = 1, \dots, n$. That means that this density is defined μ_p-almost everywhere. The definition is then complemented by the statement that $\theta_{p,\beta}$ is absolutely continuous with respect to μ_p.

(ii) We shall see later on that the condition $\beta > -1$ ensures that (29) is finite.

(iii) It was observed already in [4] that the measures $\theta_{p,\beta}$ allow a formula similar to (14). We plug the function $f(x) = \chi_{[0,\infty)\cdot\mathcal{A}} \prod_{i=1}^n x_i^\beta e^{-\|x\|_p^p}$ into (13), where $\mathcal{A}$ is any μ_p-measurable subset of Δ_p^n, and obtain

$$\int_{[0,\infty)\cdot\mathcal{A}} \prod_{i=1}^n x_i^\beta e^{-\|x\|_p^p} \, d\lambda(x) = \lambda\big([0,1] \cdot \Delta_p^n\big) \cdot n \cdot \int_0^\infty r^{n-1+n\beta} e^{-r^p} \, dr \cdot \int_{\mathcal{A}} \prod_{i=1}^n x_i^\beta \, d\mu_p(x).$$

We use a similar formula also for $\mathcal{A}=\Delta_p^n$, which leads to

$$\int_{\mathcal{A}} 1 d\theta_{p,\beta} = \frac{\int_{\mathcal{A}} \prod_{i=1}^n x_i^\beta d\mu_p(x)}{\int_{\Delta_p^n} \prod_{i=1}^n x_i^\beta d\mu_p(x)} = \frac{\int_{[0,\infty)\cdot\mathcal{A}} \prod_{i=1}^n x_i^\beta e^{-\|x\|_p^p} dx}{\int_{\mathbb{R}_+^n} \prod_{i=1}^n x_i^\beta e^{-\|x\|_p^p} dx}.$$

Let $\omega'=(\omega'_1,\dots,\omega'_n)$ be a vector with independent identically distributed components with respect to the density $c_{p,\beta}t^\beta e^{-t^p}$, $t>0$, where $c_{p,\beta}^{-1}=\int_0^\infty t^\beta e^{-t^p}dt$ is a normalizing constant. Up to a simple substitution, this is the well-known *gamma distribution*. We observe that the distribution of random points with respect to $\theta_{p,\beta}$ equals to the distribution of ℓ_p^n normalized vectors ω', i.e.,

$$\theta_{p,\beta}(\mathcal{A}) = \mathbb{P}\left(\frac{(\omega'_1,\dots,\omega'_n)}{(\sum_{j=1}^n (\omega'_j)^p)^{1/p}} \in \mathcal{A}\right), \quad \mathcal{A}\subset\Delta_p^n.$$

(iv) Of course, the same procedure might be considered also for other distributions. We leave this to future work. We also refer to the discussion on the recent work of Gribonval, Cevher, and Davies [29] in the introduction.

Lemma 14 *Let* $0<p<\infty$, $\beta>-1$, *and* $n\ge 2$.

(i) *Let* $1\le m\le n$. *Then*

$$\sigma_{m-1}^{p,\infty}(\theta_{p,\beta}) = \int_{\Delta_p^n} x_m^* \, d\theta_{p,\beta} = \frac{\mathbb{E}x_m^* \prod_{i=1}^n x_i^\beta}{\mathbb{E}\prod_{i=1}^n x_i^\beta} \cdot \frac{\Gamma(n(\beta+1)/p)}{\Gamma(n(\beta+1)/p+1/p)}.$$

(ii)

$$\mathbb{E}\prod_{i=1}^n x_i^\beta = \left[\frac{c_p}{p}\cdot\Gamma\big((\beta+1)/p\big)\right]^n.$$

Proof The proof of the first part follows again by (13), this time used for the functions

$$f_1(x) = x_m^*\left(\prod_{i=1}^n x_i^\beta\right)e^{-x_1^p-\cdots-x_n^p} \quad\text{and}\quad f_2(x) = \left(\prod_{i=1}^n x_i^\beta\right)e^{-x_1^p-\cdots-x_n^p}.$$

The proof of the second part is straightforward. □

It follows directly from (9) that $\Gamma(s)$ tends to infinity when s tends to zero. The following lemma quantifies this phenomenon. Although the statement seems to be well known, we were not able to find a reference, and we therefore provide at least a sketch of the proof.

Lemma 15 *Let* $C\simeq 0.577\ldots$ *denote the Euler constant. Then*

$$\lim_{n\to\infty}\left(\frac{\Gamma(1/n)}{n}\right)^n = e^{-C}.$$

Proof It is enough to show that

$$\lim_{n\to\infty} n\cdot\log\bigl(\Gamma(1+1/n)\bigr)=-C,$$

which (by using the l'Hospital rule) follows from

$$\lim_{n\to\infty}\frac{\int_0^\infty s^{1/n}e^{-s}\log s\,ds}{\int_0^\infty s^{1/n}e^{-s}\,ds}=-C.$$

But the numerator of this fraction is equal to $\Gamma'(1+1/n)$ and its denominator to $\Gamma(1+1/n)$. The whole fraction is therefore equal to $\Psi(1+1/n)$ and $\Psi(1+1/n)\to\Psi(1)=-C$ as n tends to infinity, cf. [1, Sect. 6.3.2, p. 258]. □

The next theorem shows that if $\beta=p/n-1$, then the measure $\theta_{p,\beta}$ promotes sparsity, and one may even consider limiting behavior of n growing to infinity.

Theorem 16 *Let* $0<p<\infty$, *and let* $n\ge 2$ *and* $1\le m\le n$ *be integers. Then*

$$\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})\ge C_p^1\cdot\frac{\Gamma(n+1)}{\Gamma(n-m+1)}\cdot\frac{\Gamma(n/p+n-m+1)}{\Gamma(n/p+n+1)}\tag{30}$$

and

$$\begin{aligned}&\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})\\&\quad\le C_p^2\cdot\frac{\Gamma(n+1)}{\Gamma(n-m+1)}\left\{\frac{\Gamma(n/p+n-m+1)}{\Gamma(n/p+n+1)}+\frac{1}{m!}\cdot\left(\frac{e^{-1}}{\Gamma(1/n)}\right)^m\right\},\end{aligned}\tag{31}$$

where C_p^1 *and* C_p^2 *are positive real numbers depending only on* p.

Furthermore, for every fixed $m\in\mathbb{N}$,

$$\frac{C_p^1}{(\frac{1}{p}+1)^m}\le\liminf_{n\to\infty}\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})\le\limsup_{n\to\infty}\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})\le\frac{C_p^2}{(\frac{1}{p}+1)^m},\tag{32}$$

where C_p^1 *and* C_p^2 *are positive real numbers depending only on* p.

Proof First observe that $n(\beta+1)/p=1$ for $\beta=p/n-1$, and therefore

$$\frac{\Gamma(n(\beta+1)/p)}{\Gamma(n(\beta+1)/p+1/p)}=\frac{1}{\Gamma(1+1/p)}$$

depends only on p. Due to Lemma 14, we have to estimate

$$\mathbb{E}x_m^*\left(\prod_{i=1}^n x_i^{p/n-1}\right)=c_p^n\int_{\mathbb{R}_+^d}x_m^*\prod_{i=1}^n x_i^{p/n-1}e^{-x_1^p-\cdots-x_n^p}\,dx.\tag{33}$$

Let $t=x_m^*$, and let us assume that there is only one coordinate $j=1,\dots,n$ such that $x_j=t$. Obviously, this assumption holds almost everywhere. Of course, we have n

possibilities for j. Furthermore, $m-1$ from the remaining $n-1$ components of x are bigger than t and the remaining $n-m$ components are smaller. This allows us to rewrite (33) as

$$\begin{aligned} &c_p^n n\binom{n-1}{m-1}\int_0^\infty t^{p/n}e^{-t^p}\left(\int_0^t u^{p/n-1}e^{-u^p}\,du\right)^{n-m} \\ &\quad\times\left(\int_t^\infty u^{p/n-1}e^{-u^p}\,du\right)^{m-1}dt \\ &=\frac{c_p^n n}{p^n}\binom{n-1}{m-1}\int_0^\infty \omega^{1/p+1/n-1}e^{-\omega}\left(\int_0^\omega s^{1/n-1}e^{-s}\,ds\right)^{n-m} \\ &\quad\times\left(\int_\omega^\infty s^{1/n-1}e^{-s}\,ds\right)^{m-1}d\omega. \end{aligned}$$

Let us write

$$\gamma=\Gamma(1/n)=\int_0^\infty s^{1/n-1}e^{-s}\,ds \quad\text{and}\quad y(\omega)=\gamma^{-1}\cdot\int_0^\omega s^{1/n-1}e^{-s}\,ds.$$

Then $y(\omega)$ is a nondecreasing function of ω, $y(0)=0$ and $\lim_{\omega\to\infty}y(\omega)=1$. We denote by $\omega(y)$ its inverse function, i.e.,

$$y=\gamma^{-1}\cdot\int_0^{\omega(y)} s^{1/n-1}e^{-s}\,ds,\quad 0\le y\le 1. \tag{34}$$

Using this notation, we obtain

$$\mathbb{E}x_m^*\left(\prod_{i=1}^n x_i^{p/n-1}\right)=\frac{c_p^n\gamma^n}{p^n}n\binom{n-1}{m-1}\int_0^1\omega(y)^{1/p}y^{n-m}(1-y)^{m-1}\,dy$$

and

$$\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})=\frac{\Gamma(n+1)}{\Gamma(m)\Gamma(n-m+1)}\int_0^1\omega(y)^{1/p}y^{n-m}(1-y)^{m-1}\,dy, \tag{35}$$

where $\omega(y)$ is given by (34). □

Step 1. Estimate from below The estimate

$$\gamma y=\int_0^{\omega(y)}s^{1/n-1}e^{-s}\,ds\le\int_0^{\omega(y)}s^{1/n-1}\,ds=n\omega(y)^{1/n}$$

implies, together with Lemma 15,

$$\omega(y)\ge\left(\frac{\gamma y}{n}\right)^n\ge cy^n,$$

with c independent of n. This gives finally

$$\begin{aligned}\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1}) &\geq c^{1/p}\cdot\frac{\Gamma(n+1)}{\Gamma(m)\Gamma(n-m+1)}\cdot\int_0^1 y^{n/p+n-m}(1-y)^{m-1}\,dy\\ &= c^{1/p}\cdot\frac{\Gamma(n+1)}{\Gamma(m)\Gamma(n-m+1)}\cdot B(n/p+n-m+1,m)\\ &= c^{1/p}\cdot\frac{\Gamma(n+1)}{\Gamma(n-m+1)}\cdot\frac{\Gamma(n/p+n-m+1)}{\Gamma(n/p+n+1)},\end{aligned}$$

where we used the Beta function (10), and the proof of (30) is complete.

Step 2. Estimate from above Let us first take y such that $1-e^{-1}/\gamma\leq y\leq 1$. Then $-\ln(\gamma(1-y))\geq 1$, and

$$\int_{-\ln(\gamma(1-y))}^{\infty} s^{1/n-1}e^{-s}\,ds\leq\int_{-\ln(\gamma(1-y))}^{\infty} e^{-s}\,ds=\gamma(1-y).$$

Hence,

$$\omega(y)\leq -\ln\bigl(\gamma(1-y)\bigr),\quad 1-e^{-1}/\gamma\leq y\leq 1. \tag{36}$$

Finally, we observe that

$$f:y\to\int_{Cy^n}^{\infty} s^{1/n-1}e^{-s}\,ds$$

is a convex function on $\mathbb{R}_+$, $f(0)=\gamma$ and

$$\begin{aligned}f\bigl(1-e^{-1}/\gamma\bigr)&=\int_{C(1-e^{-1}/\gamma)^n}^{\infty} s^{1/n-1}e^{-s}\,ds\\ &\leq\int_1^{\infty} s^{1/n-1}e^{-s}\,ds\leq e^{-1}\end{aligned}$$

if we choose C so large that $C(1-e^{-1}/\gamma)^n\geq 1$ for all $n\in\mathbb{N}$. This is indeed possible, while a byproduct of Lemma 15 is also a relation $\lim_{n\to\infty}\gamma/n=1$. Using the convexity of f, we obtain

$$f(y)\leq\gamma(1-y),\quad 0\leq y\leq 1-e^{-1}/\gamma,$$

which further leads to

$$\omega(y)\leq Cy^n,\quad 0\leq y\leq 1-e^{-1}/\gamma. \tag{37}$$

We insert (36) and (37) into (35) and obtain

$$\sigma_{m-1}^{p,\infty}(\theta_{p,p/n-1})\leq\frac{\Gamma(n+1)}{\Gamma(m)\Gamma(n-m+1)}\bigl\{C^{1/p}I_1+I_2\bigr\}, \tag{38}$$

where

$$I_1 := \int_0^{1-e^{-1}/\gamma} y^{n/p+n-m}(1-y)^{m-1}\,dy$$

and

$$I_2 := \int_{1-e^{-1}/\gamma}^{1} \left|\ln\bigl(\gamma(1-y)\bigr)\right|^{1/p} y^{n-m}(1-y)^{m-1}\,dy.$$

The first integral may be estimated again using the Beta function, which gives

$$I_1 \le B(n/p+n-m+1, m). \tag{39}$$

We denote by k the uniquely defined integer such that $1/p \le k < 1/p+1$ holds, and estimate

$$I_2 \le \int_{1-e^{-1}/\gamma}^{1} \left|\ln\bigl(\gamma(1-y)\bigr)\right|^{1/p}(1-y)^{m-1}\,dy \le I_{k,m} := \int_0^{e^{-1}/\gamma} \left|\ln(\gamma y)\right|^{k} y^{m-1}\,dy.$$

Next, we use partial integration to estimate $I_{k,m}$. We obtain

$$I_{k,m} = \frac{1}{m}\left(\frac{e^{-1}}{\gamma}\right)^m + \frac{k}{m}\cdot I_{k-1,m}.$$

Together with $I_{0,m} = 1/m \cdot (e^{-1}/\gamma)^m$, this leads finally to

$$I_{k,m} \le \frac{(k+1)!}{m}\left(\frac{e^{-1}}{\gamma}\right)^m.$$

This, together with (38) and (39), finishes the proof of (31).

The proof of (32) then follows directly by Stirling's formula (11).

Remark 5

(i) Let us take $m = 0$. Then the formula (32) describes an essentially different behavior compared to the normalized cone and surface measure. Namely, the expected value of the largest coordinate of $x \in \Delta_p^n$ with respect to $\theta_{p,p/n-1}$ does not decay to zero with n growing to infinity. We shall demonstrate this effect also numerically in the next section.

(ii) If $m > 0$, then (32) shows that $\sigma_m^{p,\infty}(\theta_{p,p/n-1})$ decays exponentially fast with m as soon as n is large enough. That means that for n large enough, the average vector of Δ_p^n exhibits a strong sparsity-like structure. Namely, its m-th largest component decays exponentially with m.

(iii) We have chosen in (28) a different β for each n; namely, $\beta_n = p/n - 1 > -1$. This was of course a crucial ingredient in the proof of Theorem 16. It is not difficult to adapt the analysis of the proof of Theorem 16 to the situation when $\beta > -1$ is fixed for all $n \in \mathbb{N}$. In this case, we obtain again that (up to logarithmic factors) $\sigma_0^{p,\infty}(\theta_{p,\beta})$ is equivalent to $n^{-1/p}$ with constants of equivalence depending on $p > 0$ and $\beta > -1$.

(iv) Last, but not least, we observe that one may choose $p = 1$ or even $p = 2$ in Theorem 16 and still obtain the exponential decay of coordinates as described by (32). It seems that there is no significant connection between sparsity of an average vector of $x \in \Delta_p^n$ and the size of $p > 0$.

5 Numerical Experiments

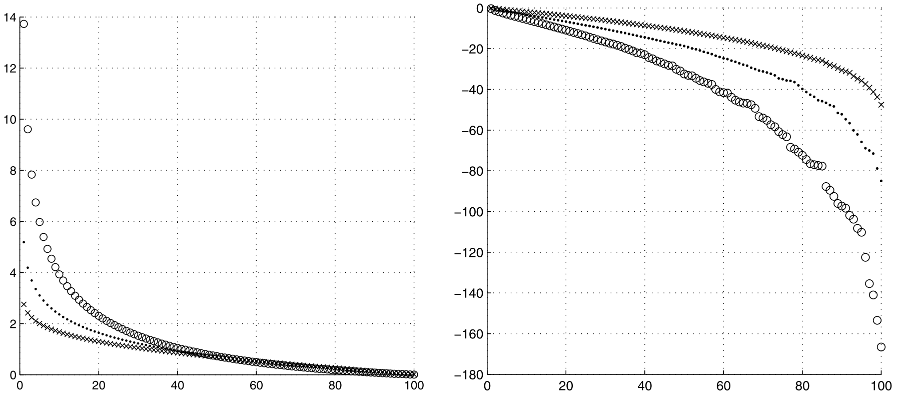
5.1 Cone Measure

We would like to demonstrate the most significant effects of the theory also by numerical experiments. We start with the case of the cone measure. The key role is played by (14). It may be interpreted in the following way. To generate a random point on Δ_p^n with respect to the normalized cone measure, it is enough to generate $\omega_1, \dots, \omega_n$ with respect to the density $c_p e^{-t^p}$, $t > 0$, and then calculate

$$\frac{(\omega_1, \dots, \omega_n)}{(\sum_{j=1}^n \omega_j^p)^{1/p}} \in \Delta_p^n.$$

This method is very practical, as the running time of this algorithm depends only linearly on n.

Let us note that the values of ω_i may be generated very easily. For example, the package *GNU Scientific Library* [26] implements a random number generator with respect to the gamma distribution using the method described in the classical work of Knuth [31]. Using this package, we generated 10^8 random points $x \in \Delta_p^n$ for $n = 100$ and $p \in \{1/2, 1, 2\}$ to approximate numerically the value of $n^{1/p} \cdot \int_{\Delta_p^n} x_m^* d\mu_p(x)$. The result may be found in Fig. 1.



(a) $n^{1/p} \cdot \int_{\Delta_p^n} x_m^* d\mu_p(x)$

(b) $\log_{10}(\int_{\Delta_p^n} x_m^* d\theta_{p,p/n-1})$

Fig. 1 Approximations of $n^{1/p} \cdot \int_{\Delta_p^n} x_m^* \, d\mu_p(x)$ (*left*) and $\log_{10}(\int_{\Delta_p^n} x_m^* \, d\theta_{p,p/n-1})$ (*right*) for $n = 100$, $p = 1/2(\circ)$, $p = 1(\bullet)$ and $p = 2(\times)$ based on sampling of 10^8 random points

5.2 Tensor Measures

As pointed out in Remark 4, point (iii), a random point on Δ_p^n with respect to $\theta_{p,\beta}$ may be generated in the following way. We generate $\omega_1',\dots,\omega_n'$ with respect to the density $c_{p,\beta}t^\beta e^{-t^p}$, $t>0$, where $c_{p,\beta}^{-1}=\int_0^\infty t^\beta e^{-t^p}\,dt$ is a normalizing constant, and we consider the vector

$$\frac{(\omega_1',\dots,\omega_n')}{(\sum_{j=1}^n(\omega_j')^p)^{1/p}}\in\Delta_p^n.$$

Also this may be easily done with the help of [26]. We generated again 10^8 random points $x\in\Delta_p^n$ with respect to $\theta_{p,p/n-1}$ for $n=100$ and $p\in\{1/2,1,2\}$. Then we used those points to numerically approximate the expression $\log_{10}(\int_{\Delta_p^n}x_m^*d\theta_{p,p/n-1})$.

Acknowledgements I would like to thank to Stephan Dahlke, Massimo Fornasier, Aicke Hinrichs, Erich Novak and Henryk Woźniakowski for their interest in this topic and the anonymous referees for their valuable comments and remarks, which helped to greatly improve the quality of the paper. In particular, the proof of the second part of Theorem 7 was suggested by one of the referees. I acknowledge the financial support provided by the START-award "Sparse Approximation and Optimization in High Dimensions" of the Fonds zur Förderung der wissenschaftlichen Forschung (FWF, Austrian Science Foundation).

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. U.S. Government Printing Office, Washington (1964)
2. Anttila, M., Ball, K., Perissinaki, I.: The central limit problem for convex bodies. Trans. Am. Math. Soc. **355**(12), 4723–4735 (2003)
3. Ball, K., Perissinaki, I.: The subindependence of coordinate slabs in ℓ_p^n balls. Isr. J. Math. **107**, 289–299 (1998)
4. Barthe, F., Csörnyei, M., Naor, A.: A note on simultaneous polar and Cartesian decomposition. In: Geometric Aspects of Functional Analysis. Lecture Notes in Mathematics. Springer, Berlin (2003)
5. Barthe, F., Guédon, O., Mendelson, S., Naor, A.: A probabilistic approach to the geometry of the l_p^n-ball. Ann. Probab. **33**(2), 480–513 (2005)
6. Bennett, C., Sharpley, R.: Interpolation of Operators. Pure and Applied Mathematics, vol. 129. Academic Press, Boston (1988)
7. Bobin, J., Starck, J.-L., Fadili, J.M., Moudden, Y., Donoho, D.L.: Morphological component analysis: an adaptive thresholding strategy. IEEE Trans. Image Process. **16**(11), 2675–2681 (2007)
8. Candés, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid, Spain (2006)
9. Candés, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
10. Candés, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
11. Candés, E.J., Romberg, J.K., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
12. Cevher, V.: Learning with compressible priors. In: Neural Information Processing Systems (NIPS) (2009)
13. Champagnat, F., Goussard, Y., Idier, J.: Unsupervised deconvolution of sparse spike trains using stochastic approximation. IEEE Trans. Signal Process. **44**(12), 2988–2998 (1996)
14. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**(1), 211–231 (2009)
15. Dahlke, S., Novak, E., Sickel, W.: Optimal approximation of elliptic problems by linear and nonlinear mappings I. J. Complex. **22**(1), 29–49 (2006)

16. David, H.A., Nagaraja, H.N.: Order Statistics. Wiley-Interscience, New York (2004)
17. DeVore, R.A.: Nonlinear approximation. Acta Numer. 51–150 (1998)
18. DeVore, R.A., Jawerth, B., Popov, V.: Compression of wavelet decompositions. Am. J. Math. **114**(4), 737–785 (1992)
19. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
20. Dvoretzky, A.: Some results on convex bodies and Banach spaces. In: Proc. Internat. Sympos. Linear Spaces, Jerusalem, pp. 123–160 (1960)
21. Edwards, J.: A Treatise on the Integral Calculus, vol. II. Chelsea, New York (1922)
22. Figiel, T.: A short proof of Dvoretzky's theorem on almost spherical sections of convex bodies. Compos. Math. **33**(3), 297–301 (1976)
23. Figiel, T., Lindenstrauss, J., Milman, V.D.: The dimension of almost spherical sections of convex bodies. Acta Math. **139**(1–2), 53–94 (1977)
24. Fornasier, M.: Numerical methods for sparse recovery. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Series on Computational and Applied Mathematics, vol. 9. Springer, Berlin (2010)
25. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Appl. Numer. Harmon. Anal., Birkhäuser, Boston (in preparation)
26. GNU Scientific Library: http://www.gnu.org/software/gsl/
27. Gluskin, E.D.: An octahedron is poorly approximated by random subspaces. Funkc. Anal. Prilozh. **20**(1), 14–20 (1986). 96
28. Gordon, Y., Litvak, A.E., Schütt, C., Werner, E.: On the minimum of several random variables. Proc. Am. Math. Soc. **134**(12), 3665–3675 (2006)
29. Gribonval, R., Cevher, V., Davies, M.: Compressible priors for high-dimensional statistics. Preprint (2011)
30. Gribonval, R., Schnass, K.: Dictionary identification—sparse matrix factorisation via ℓ_1 minimisation. IEEE Trans. Inf. Theory **56**(7), 3523–3539 (2010)
31. Knuth, D.E.: Seminumerical Algorithms, 3rd edn. The Art of Computer Programming, vol. 2. Addison-Wesley, Reading (1998)
32. Ledoux, M.: The Concentration of Measure Phenomenon. AMS, Providence (2001)
33. Ledoux, M., Talagrand, M.: Probability in Banach Spaces. Springer, Berlin (1991)
34. Mattila, P.: Geometry of Sets and Measures in Euclidean Spaces. Cambridge University Press, Cambridge (1995)
35. Milman, V.D.: A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. Funkc. Anal. Prilozh. **5**(4), 28–37 (1971)
36. Milman, V.D., Schechtman, G.: Asymptotic Theory of Finite-Dimensional Normed Spaces. Lecture Notes in Mathematics, vol. 1200. Springer, Berlin (1986)
37. Naor, A.: The surface measure and cone measure on the sphere of l_p^n. Trans. Am. Math. Soc. **359**(3), 1045–1079 (2007)
38. Naor, A., Romik, D.: Projecting the surface measure of the sphere of l_p^n. Ann. Inst. Henri Poincaré Probab. Stat. **39**(2), 241–261 (2003)
39. Oskolkov, K.: Polygonal approximation of functions of two variables. Math. USSR Sb. **35**, 851–861 (1979)
40. Pesquet, J.C., Krim, H., Leporini, D., Hamman, E.: Bayesian approach to best basis selection. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., pp. 2634–2637 (1996)
41. Rachev, S.T., Rüschendorf, L.: Approximate independence of distributions on spheres and their stability properties. Ann. Probab. **19**(3), 1311–1337 (1991)
42. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery, vol. 9 (2010)
43. Schechtman, G., Zinn, J.: On the volume of the intersection of two L_p^n balls. Proc. Am. Math. Soc. **110**(1), 217–224 (1990)
44. Schmidt, E.: Zur Theorie der linearen und nichtlinearen Integralgleichungen I. Math. Anal. **63**, 433–476 (1907)
45. Temlyakov, V.N.: Nonlinear methods of approximation. Found. Comput. Math. **3**(1), 33–107 (2003)

PARTICLE SYSTEMS AND KINETIC EQUATIONS MODELING INTERACTING AGENTS IN HIGH DIMENSION

M. FORNASIER[†‡], J. HAŠKOVEC[‡], AND J. VYBÍRAL[‡]

Abstract. In this paper we explore how concepts of high-dimensional data compression via random projections onto lower-dimensional spaces can be applied for tractable simulation of certain dynamical systems modeling complex interactions. In such systems, one has to deal with a large number of agents (typically millions) in spaces of parameters describing each agent of high dimension (thousands or more). Even with today's powerful computers, numerical simulations of such systems are prohibitively expensive. We propose an approach for the simulation of dynamical systems governed by functions of *adjacency matrices* in high dimension, by random projections via Johnson-Lindenstrauss embeddings, and recovery by compressed sensing techniques. We show how these concepts can be generalized to work for associated kinetic equations, by addressing the phenomenon of the delayed curse of dimension, known in information-based complexity for optimal numerical integration problems in high dimensions.

Key words. Dimensionality reduction, dynamical systems, flocking and swarming, Johnson-Lindenstrauss embedding, compressed sensing, high-dimensional kinetic equations, delayed curse of dimension, optimal integration of measures in high dimension.

AMS subject classifications. 34C29, 35B35, 35Q91, 35Q94, 60B20, 65Y20.

1. Introduction.

The dimensionality scale of problems arising in our modern information society has become very large and finding appropriate methods for dealing with them is one of the great challenges of today's numerical simulation. The most notable recent advances in data analysis are based on the observation that in many situations, even for very complex phenomena, the intrinsic dimensionality of the data is significantly lower than the ambient dimension. Remarkable progresses have been made in data compression, processing, and acquisition. We mention, for instance, the use of *diffusion maps* for data clouds and graphs in high dimension [5, 6, 17, 18, 19] in order to define low-dimensional local representations of data with small distance distortion, and meaningful automatic clustering properties. In this setting the embedding of data is performed by a *highly nonlinear* procedure, obtained by computing the eigenfunctions of suitable normalized diffusion kernels, measuring the probability of transition from one data point to another over the graph.

Quasi-isometrical *linear* embeddings of high-dimensional point clouds into low-dimensional spaces of parameters are provided by the well-known Johnson-Lindenstrauss Lemma [1, 22, 35]: any cloud of $\mathcal{N}$ points in $\mathbb{R}^d$ can be embedded by a random linear projection M nearly isometrically into $\mathbb{R}^k$ with $k = \mathcal{O}(\varepsilon^{-2}\log(\mathcal{N}))$ (a precise statement will be given below). This embedding strategy is simpler than the use of diffusion maps, as it is linear, however it is "blind" to the specific geometry and local dimensionality of the data, as the embedding dimension k depends exclusively on the number of points in the cloud. In many applications, this is sufficient, as the number of points $\mathcal{N}$ is supposed to be a power of the dimension d, and the embedding produces an effective reduction to $k = \mathcal{O}(\varepsilon^{-2}\log(\mathcal{N})) = \mathcal{O}(\varepsilon^{-2}\log(d))$ dimensions. As clarified in [3, 37], the Johnson-Lindenstrauss Lemma is also at the basis of the possibility of performing optimal compressed and nonadaptive acquisition of high-dimensional

[†] Faculty of Mathematics, Technical University of Munich, Boltzmannstrasse 3, D-85748 Garching, Germany

[‡] Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria

data. In *compressed sensing* [12, 24, 28] a vector $x \in \mathbb{R}^d$ is encoded in a vector $y \in \mathbb{R}^k$ by applying a random projection M, which is modeling a linear acquisition device with random sensors, i.e., $y = Mx$. From y it is possible to decode x approximatively (see Theorem 3.7 below) by solving the convex optimization problem

$$x^{\#} = \arg \min_{Mz=y} \left(\|z\|_{\ell_1^d} := \sum_{i=1}^{d} |z_i| \right),$$

with the error distortion

$$\|x^{\#} - x\|_{\ell_1^d} \leq C\sigma_K(x)_{\ell_1^d},$$

where $\sigma_K(x)_{\ell_1^d} = \inf_{z:\#\mathrm{supp}\,(z)\leq K} \|z - x\|_{\ell_1^d}$ and $K = \mathcal{O}(k/(\log(d/k) + 1))$. We denote $\Sigma_K = \{z \in \mathbb{R}^d : \#\mathrm{supp}\,(z) \leq K\}$ the set of K-sparse vectors, i.e., the union of K-dimensional coordinate subspaces in $\mathbb{R}^d$. In particular, if $x \in \Sigma_K$, then $x^{\#} = x$. Hence, not only is M a Johnson-Lindenstrauss embedding, quasi-isometrical on point clouds and K-dimensional coordinate subspaces, but also allows for the recovery of the most relevant components of high-dimensional vectors, from low-dimensional encoded information. A recent work [4, 48] extends the quasi-isometrical properties of the Johnson-Lindenstrauss embedding from point clouds and K-dimensional coordinate subspaces to smooth compact Riemannian manifolds with bounded curvature. Inspired by this work, in [34] the authors extend the principles of compressed sensing in terms of point recovery on smooth compact Riemannian manifolds.

Besides these relevant results in compressing and coding-decoding high-dimensional "stationary" data, dimensionality reduction of complex dynamical systems and high-dimensional partial differential equations is a subject of recent intensive research. Several tools have been employed, for instance, the use of diffusion maps for dynamical systems [39], tensor product bases and sparse grids for the numerical solution of linear high-dimensional PDEs [23, 10, 30, 31], the reduced basis method for solving high-dimensional parametric PDEs [7, 9, 38, 43, 44, 46].

In this paper we shall further explore the connection between data compression and *tractable* numerical simulation of dynamical systems, and solutions of associated high-dimensional kinetic equations. We are specially interested in dynamical systems of the type

$$\dot{x}_i(t) = f_i(\mathcal{D}x(t)) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}x(t))x_j(t), \tag{1.1}$$

where we use the following notation:

- $N \in \mathbb{N}$ - number of agents,
- $x(t) = (x_1(t), \ldots, x_N(t)) \in \mathbb{R}^{d\times N}$, where $x_i : [0, T] \to \mathbb{R}^d$, $i = 1, \ldots, N$,
- $f_i : \mathbb{R}^{N\times N} \to \mathbb{R}^d, \quad i = 1, \ldots, N$,
- $f_{ij} : \mathbb{R}^{N\times N} \to \mathbb{R}, \quad i, j = 1, \ldots, N$,
- $\mathcal{D} : \mathbb{R}^{d\times N} \to \mathbb{R}^{N\times N}$, $\mathcal{D}x := (\|x_i - x_j\|_{\ell_2^d})_{i,j=1}^N$ is the *adjacency matrix* of the point cloud x.

We shall assume that the governing functions f_i and f_{ij} are Lipschitz, but we shall specify the details later on. The system (1.1) describes the dynamics of multiple complex agents $x(t) = (x_1(t), \ldots, x_N(t)) \in \mathbb{R}^{d\times N}$, interacting on the basis of their mutual "social" distance $\mathcal{D}x(t)$, and its general form includes several models for swarming and collective motion of animals and micro-organisms, aggregation of cells, etc. Several relevant effects can be included in the model by means of the functions f_i and f_{ij}, in particular, fundamental binary mechanisms of *attraction, repulsion, aggregation* and *alignment* [13, 14, 20, 21, 41, 36]. Moreover, possibly adding stochastic terms of random noise may also allow to consider *diffusion* effects [8, 14]. However, these models and motion mechanisms are mostly derived borrowing a leaf from physics, by assuming the agents (animals, micro-organisms, cells etc.) as pointlike and exclusively determined by their spatial position and velocity in $\mathbb{R}^d$ for $d = 3+3$. In case we wished to extend such models of social interaction to more "sophisticated" agents, described by many parameters ($d \gg 3 + 3$), the simulation may become computationally prohibitive. Our motivation for considering high-dimensional situations stems from the modern development of communication technology and Internet, for which we witness the development of larger and larger communities accessing information (interactive databases), services (financial market), social interactions (social networks) etc. For instance, we might be interested to simulate the behavior of certain subsets of the financial market where the agents are many investors, who are characterized by their portfolios of several hundreds of investments. The behavior of each individual investor depends on the dynamics of others according to a suitable social distance determined by similar investments. Being able to produce meaningful simulations and learning processes of such complex dynamics is an issue, which might be challenged by using suitable compression/dimensionality reduction techniques.

The idea we develop in this paper is to project randomly the system and its initial condition by Johnson-Lindenstrauss embeddings to a lower-dimensional space where an independent simulation can be performed with significantly reduced complexity. We shall show that the use of multiple projections and parallel computations allows for an approximate reconstruction of the high-dimensional dynamics, by means of compressed sensing techniques. After we explore the tractable simulation of the dynamical systems (1.1) when the dimension d of the parameter space is large, we also address the issue of whether we can perform tractable simulations when also the number N of agents is getting very large. Unlike the control of a finite number of agents, the numerical simulation of a rather large population of interacting agents ($N \gg 0$) can constitute a serious difficulty which stems from the accurate solution of a possibly very large system of ODEs. Borrowing the strategy from the kinetic theory of gases [16], we want instead to consider a density distribution of agents, depending on their d-parameters, which interact with stochastic influence (corresponding to classical collisional rules in kinetic theory of gases) – in this case the influence is "smeared" since two individuals may interact also when they are far apart in terms of their "social distance" $\mathcal{D}x$. Hence, instead of simulating the behavior of each individual agent, we shall describe the collective behavior encoded by a density distribution μ, whose evolution is governed by one sole mesoscopic partial differential equation. We shall show that, under realistic assumptions on the concentration of the measure μ on sets of lower dimension, we can also acquire information on the properties of the high-dimensional measure solution μ of the corresponding kinetic equation, by considering random projections to lower dimension. Such approximation properties are determined by means of the combination of optimal numerical integration principles for the high-dimensional measure μ [29, 32] and the results previously achieved for particle dynamical systems.

1.1. Fundamental assumptions. We introduce the following notation for ℓ_p-norms of vectors $v \in \mathbb{R}^d$,

$$\|v\|_{\ell_p^d} := \left(\sum_{i=1}^{d} |v_i|^p \right)^{1/p} \qquad \text{for } 1 \le p < \infty,$$

and

$$\|v\|_{\ell_\infty^d} := \max_{i=1,\dots,d} |v_i|.$$

For matrices $x \in \mathbb{R}^{n \times m}$ we consider the mixed norm

$$\|x\|_{\ell_p^m(\ell_q^n)} := \|(\|x_i\|_{\ell_p^n})_{i=1}^m\|_{\ell_q^m},$$

where $x_i \in \mathbb{R}^n$ is the i^{th}-column of the matrix x.

For the rest of the paper we impose three fundamental assumptions about Lipschitz and boundedness properties of f_i and f_{ij},

$$|f_i(a) - f_i(b)| \le L\|a-b\|_{\ell_\infty^N(\ell_\infty^N)}, \quad i = 1, \dots, N \tag{1.2}$$

$$\max_{i=1,\dots,N} \sum_{j=1}^{N} |f_{ij}(a)| \le L', \tag{1.3}$$

$$\max_{i=1,\dots,N} \sum_{j=1}^{N} |f_{ij}(a) - f_{ij}(b)| \le L''\|a-b\|_{\ell_\infty^N(\ell_\infty^N)}, \tag{1.4}$$

for every $a, b \in \mathbb{R}^{N \times N}$. Unfortunately, models of real-life phenomena would not always satisfy these conditions, for instance models of financial markets or socio-economic interactions can be expected to exhibit severely discontinuous behavior. However, these assumptions are reasonable in certain regimes and allow us to prove the concept we are going to convey in this paper, i.e., the possibility of simulating high-dimensional dynamics by multiple independent simulations in low dimension.

1.2. Euler scheme, a classical result of stability and convergence, and its complexity. We shall consider the system of ordinary differential equations of the form (1.1) with the initial condition

$$x_i(0) = x_i^0\,, \qquad i = 1, \dots, N\,. \tag{1.5}$$

The Euler method for this system is given by (1.5) and

$$x_i^{n+1} := x_i^n + h \left[f_i(\mathcal{D}x^n) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}x^n) x_j^n \right], \quad n = 0, \dots, n_0 - 1. \tag{1.6}$$

where $h > 0$ is the time step and $n_0 := T/h$ is the number of iterations. We consider here the *explicit* Euler scheme exclusively for the sake of simplicity, for more sophisticated integration methods might be used. We start with a classical result, which we report in detail for the sake of the reader, and for simplicity we assume $f_{ij} = 0$ for all $i, j = 1, \dots N$.

THEOREM 1.1 (Stability and convergence of the Euler scheme). *Fix $x^0 \in \mathbb{R}^{d\times N}$ and let $x(t)$ be the unique solution of the ODE*

$$\dot{x}(t) = f(\mathcal{D}x(t))\,, \qquad x(0) = x^0\,, \tag{1.7}$$

on the interval $[0,T]$, $T > 0$, for $f = (f_i)_{i=1}^N$ satisfying (1.2). Moreover, fix $h > 0$ and let $t^n := nh$ and $\tilde{x}^n$ be the approximate solution obtained by the explicit Euler method, i.e.,

$$\tilde{x}^{n+1} = \tilde{x}^n + hf(\mathcal{D}\tilde{x}^n)\,, \qquad \tilde{x}_0 = \tilde{x}^0\,,$$

for $n = 0,\dots,n_0-1$. Note that we allow different initial conditions x^0 and $\tilde{x}^0$ for the continuous and, resp., discrete solutions. Then, we have the error estimate

$$\mathcal{E}^n \le \exp(2Lt^n)\left(\mathcal{E}^0 + ht^n\frac{\|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}}{2}\right),$$

where $\mathcal{E}^n = \|x(t^n) - \tilde{x}^n\|_{\ell_\infty^N(\ell_2^d)}$.

Proof. For the sake of the proof, we extend $\tilde{x}$ to the full interval $[0,T]$ by linear interpolation between the grid points t^n, i.e.,

$$\tilde{x}(t^n+s) = \tilde{x}(t^n) + sf(\mathcal{D}\tilde{x}(t^n)) \qquad \text{for } s\in[0,h]\,,$$

such that $\tilde{x}$ is a continuous, piecewise linear function on $[0,T]$.

For a fixed n and $t := t^n$, let us consider the exact and approximate solutions in the interval $[t, t+\tau]$ with $\tau\in[0,h]$:

$$x(t+\tau) = x(t) + \int_0^\tau f(\mathcal{D}x(t+s))\,\mathrm{d}s\,, \tag{1.8}$$

$$\tilde{x}(t+\tau) = \tilde{x}(t) + \int_0^\tau f(\mathcal{D}\tilde{x}(t))\,\mathrm{d}s\,. \tag{1.9}$$

Subtracting (1.9) from (1.8) and using (1.2), we obtain

$$\begin{aligned}\|x(t+\tau)-\tilde{x}(t+\tau)\|_{\ell_\infty^N(\ell_2^d)} &\le \|x(t)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)} + \int_0^\tau \|f(\mathcal{D}x(t+s)) - f(\mathcal{D}\tilde{x}(t))\|_{\ell_\infty^N(\ell_2^d)}\,\mathrm{d}s\\ &\le \|x(t)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)} + L\int_0^\tau \|\mathcal{D}x(t+s)-\mathcal{D}\tilde{x}(t)\|_{\ell_\infty^N(\ell_\infty^N)}\,\mathrm{d}s\\ &\le \|x(t)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)} + 2L\int_0^\tau \|x(t+s)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)}\,\mathrm{d}s\,.\end{aligned}$$

Moreover, for $s\in[0,h]$,

$$\begin{aligned}\|x(t+s)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)} &\le \|x(t+s)-\tilde{x}(t+s)\|_{\ell_\infty^N(\ell_2^d)} + \|\tilde{x}(t+s)-\tilde{x}(t)\|_{\ell_\infty^N(\ell_2^d)}\\ &= \|x(t+s)-\tilde{x}(t+s)\|_{\ell_\infty^N(\ell_2^d)} + s\|f(\mathcal{D}\tilde{x}(t))\|_{\ell_\infty^N(\ell_2^d)}\,.\end{aligned}$$

The term $\|f(\mathcal{D}\tilde{x}(t))\|_{\ell_\infty^N(\ell_2^d)} = \|f(\mathcal{D}\tilde{x}^n)\|_{\ell_\infty^N(\ell_2^d)}$ is bounded by $(1+2Lh)^n\|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}$, which can be seen from the simple induction

$$\begin{aligned}\|f(\mathcal{D}\tilde{x}^n)\|_{\ell_\infty^N(\ell_2^d)} &\le \|f(\mathcal{D}\tilde{x}^n) - f(\mathcal{D}\tilde{x}_{n-1})\|_{\ell_\infty^N(\ell_2^d)} + \|f(\mathcal{D}\tilde{x}_{n-1})\|_{\ell_\infty^N(\ell_2^d)}\\ &\le L\|\mathcal{D}\tilde{x}_n - \mathcal{D}\tilde{x}_{n-1}\|_{\ell_\infty^N(\ell_\infty^N)} + \|f(\mathcal{D}\tilde{x}_{n-1})\|_{\ell_\infty^N(\ell_2^d)}\\ &\le 2L\|\tilde{x}_n - \tilde{x}_{n-1}\|_{\ell_\infty^N(\ell_2^d)} + \|f(\mathcal{D}\tilde{x}_{n-1})\|_{\ell_\infty^N(\ell_2^d)}\\ &= (1+2Lh)\|f(\mathcal{D}\tilde{x}_{n-1})\|_{\ell_\infty^N(\ell_2^d)} \le (1+2Lh)^n\|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}\,.\end{aligned}$$

Consequently, defining $\mathcal{E}(t+\tau) := \|x(t+\tau) - \tilde{x}(t+\tau)\|_{\ell_\infty^N(\ell_2^d)}$, we obtain

$$\begin{aligned}\mathcal{E}(t+\tau) &\leq \mathcal{E}(t) + 2L\int_0^\tau \left(\mathcal{E}(t+s) + s(1+2Lh)^n \|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}\right) \mathrm{d}s \\ &\leq \mathcal{E}(t) + 2L\int_0^\tau \mathcal{E}(t+s)\,\mathrm{d}s + \frac{h^2}{2}(1+2Lh)^n \|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)} \,.\end{aligned}$$

An application of the Gronwall lemma yields

$$\mathcal{E}(t+h) \leq \left(\mathcal{E}(t) + \frac{h^2}{2}(1+2Lh)^n \|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}\right) \exp(2Lh) \,.$$

By another simple induction we obtain

$$\mathcal{E}^n \leq \exp(2Lnh)\mathcal{E}^0 + \left(\sum_{k=1}^n \exp(2Lkh)(1+2Lh)^{n-k}\right) \frac{h^2}{2} \|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)} \,,$$

where we turned back to the notation $\mathcal{E}^n = \mathcal{E}(t^n)$. Using $(1+2Lh)^{n-k} \leq \exp(2Lh(n-k))$, we have

$$\mathcal{E}^n \leq \exp(2Lnh)\mathcal{E}^0 + \exp(2Lnh) n \frac{h^2}{2} \|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)} \,,$$

and, finally, writing t^n for nh, we conclude

$$\mathcal{E}^n \leq \exp(2Lt^n) \left(\mathcal{E}^0 + ht^n \frac{\|f(\mathcal{D}\tilde{x}^0)\|_{\ell_\infty^N(\ell_2^d)}}{2}\right) \,.$$

□

The simulation of the dynamical system (1.7) has a complexity which is at least the one of computing the *adjacency matrix* $\mathcal{D}\tilde{x}^n$ at each discrete time t^n, i.e., $\mathcal{O}(d \times N^2)$. The scope of the next sections is to show that, up to an ε-distortion, we can approximate the dynamics of (1.1) by projecting the system into lower dimension and by executing in parallel computations with reduced complexity. Computation of the adjacency matrix in the new dimension requires only $\mathcal{O}(\varepsilon^{-2}\log(N) \times N^2)$ operations. Especially if the distortion parameter $\varepsilon > 0$ is not too small and the number of agents is of a polynomial order in d, we reduce the complexity of computing the adjacency matrix to $\mathcal{O}(\log(d) \times N^2)$.

2. Projecting the Euler method: dimensionality reduction of discrete dynamical systems.

2.1. Johnson-Lindenstrauss embedding. We wish to project the dynamics of (1.1) into a lower-dimensional space by employing a well-known result of Johnson and Lindenstrauss [35], which we informally rephrase for our purposes as follows.

Lemma 2.1 (Johnson and Lindenstrauss). *Let $\mathcal{P}$ be an arbitrary set of $\mathcal{N}$ points in $\mathbb{R}^d$. Given a distortion parameter $\varepsilon > 0$, there exists a constant*

$$k_0 = \mathcal{O}(\varepsilon^{-2}\log(\mathcal{N})),$$

such that for all integers $k \geq k_0$, there exists a $k \times d$ matrix M for which

$$(1-\varepsilon)\|x-\tilde{x}\|_{\ell_2^d}^2 \leq \|Mx - M\tilde{x}\|_{\ell_2^k}^2 \leq (1+\varepsilon)\|x-\tilde{x}\|_{\ell_2^d}^2, \tag{2.1}$$

for all $x, \tilde{x} \in \mathcal{P}$. It is easy to see that the condition

$$(1-\varepsilon)\|p\|_{\ell_2^d}^2 \le \|Mp\|_{\ell_2^k}^2 \le (1+\varepsilon)\|p\|_{\ell_2^d}^2, \quad p \in \mathbb{R}^d, \tag{2.2}$$

implies

$$(1-\varepsilon)\|p\|_{\ell_2^d} \le \|Mp\|_{\ell_2^k} \le (1+\varepsilon)\|p\|_{\ell_2^d}, \quad p \in \mathbb{R}^d, \tag{2.3}$$

for $0 < \varepsilon < 1$, which will be used in the following sections. On the other hand, (2.3) implies (2.2) with 3ε instead of ε.

Our aim is to apply this lemma to dynamical systems. As the mapping M from Lemma 2.1 is linear and almost preserves distances between the points (up to the $\varepsilon > 0$ distortion as described above), we restrict ourselves to dynamical systems which are linear or whose non-linearity depends only on the mutual distances of the points involved, as in (1.1).

Let us define the additional notation, which is going to be fixed throughout the paper:

- $d \in \mathbb{N}$ - dimension (large),
- $\varepsilon > 0$ - the distortion parameter from Lemma 2.1,
- $k \in \mathbb{N}$ - new dimension (small),
- $M \in \mathbb{R}^{k\times d}$ - randomly generated matrix as described below.

The only constructions of a matrix M as in Lemma 2.1 known up to now are stochastic, i.e., the matrix is randomly generated and has the quasi-isometry property (2.1) with high probability. We refer the reader to [22] and [1, Theorem 1.1] for two typical versions of the Johnson-Lindenstrauss Lemma.

We briefly collect below some well-known instances of random matrices, which satisfy the statement of Lemma 2.1 with high probability:

- $k \times d$ matrices M whose entries $m_{i,j}$ are independent realizations of Gaussian random variables
$$m_{i,j} \sim \mathcal{N}\left(0, \frac{1}{k}\right);$$
- $k \times d$ matrices M whose entries are independent realizations of ± Bernoulli random variables
$$m_{i,j} := \begin{cases} +\frac{1}{\sqrt{k}}, & \text{with probability } \frac{1}{2} \\ -\frac{1}{\sqrt{k}}, & \text{with probability } \frac{1}{2} \end{cases}$$

Several other random projections suitable for Johnson-Lindenstrauss embeddings can be constructed following Theorem 3.6 recalled below, and we refer the reader to [37] for more details.

2.2. Uniform estimate for a general model. If $M \in \mathbb{R}^{k\times d}$ is a matrix, we consider the projected Euler method in $\mathbb{R}^k$ associated to the high-dimensional system (1.5)-(1.6), namely

$$y_i^0 := Mx_i^0, \tag{2.4}$$

$$y_i^{n+1} := y_i^n + h\left[Mf_i(\mathcal{D}'y^n) + \sum_{j=1}^N f_{ij}(\mathcal{D}'y^n)y_j^n\right], \quad n = 0,\dots,n_0-1. \tag{2.5}$$

We denote here $\mathcal{D}' : \mathbb{R}^{k\times N} \to \mathbb{R}^{N\times N}$, $\mathcal{D}'y := (\|y_i - y_j\|_{\ell_2^k})_{i,j=1}^N$, the *adjacency matrix* of the agents $y = (y_1, \ldots, y_N)$ in $\mathbb{R}^{k\times N}$. The first result of this paper reads as follows.

THEOREM 2.2. *Let the sequences*

$$\{x_i^n, i = 1,\ldots,N \text{ and } n = 0,\ldots,n_0\} \quad \text{and} \quad \{y_i^n, i = 1,\ldots,N \text{ and } n = 0,\ldots,n_0\}$$

be defined by (1.5)-(1.6) *and* (2.4)-(2.5) *with f_i and f_{ij} satisfying* (1.2)–(1.4) *and a matrix $M \in \mathbb{R}^{k\times d}$ with*

$$\|Mf_i(\mathcal{D}'y^n) - Mf_i(\mathcal{D}x^n)\|_{\ell_2^k} \le (1+\varepsilon)\,\|f_i(\mathcal{D}'y^n) - f_i(\mathcal{D}x^n)\|_{\ell_2^d}, \tag{2.6}$$

$$\|Mx_j^n\|_{\ell_2^k} \le (1+\varepsilon)\|x_j^n\|_{\ell_2^d}, \tag{2.7}$$

$$(1-\varepsilon)\|x_i^n - x_j^n\|_{\ell_2^d} \le \|Mx_i^n - Mx_j^n\|_{\ell_2^k} \le (1+\varepsilon)\|x_i^n - x_j^n\|_{\ell_2^d} \tag{2.8}$$

for all $i, j = 1,\ldots,N$ and all $n = 0,\ldots,n_0$. Moreover, let us assume that

$$\alpha \ge \max_j \|x_j^n\|_{\ell_2^d} \quad \text{for all} \quad n = 0,\ldots,n_0, \quad j = 1,\ldots,N.$$

Let

$$e_i^n := \|y_i^n - Mx_i^n\|_{\ell_2^k}, \; i = 1,\ldots,N \text{ and } n = 0,\ldots,n_0 \tag{2.9}$$

and set $\mathcal{E}^n := \max_i e_i^n$. Then

$$\mathcal{E}^n \le \varepsilon h n B \exp(hnA), \tag{2.10}$$

where $A := L' + 2(1+\varepsilon)(L+\alpha L'')$ and $B := 2\alpha(1+\varepsilon)(L+\alpha L'')$.

We remark that conditions (2.6)-(2.8) are in fact satisfied as soon as M is a suitable Johnson-Lindenstrauss embedding as in Lemma 2.1.

Proof. Using (2.9) and (1.5)-(1.6) and (2.4)-(2.5) combined with (2.6) and (2.7), we obtain

$$\begin{aligned}
e_i^{n+1} &\le e_i^n + h\,\|Mf_i(\mathcal{D}'y^n) - Mf_i(\mathcal{D}x^n)\|_{\ell_2^k} + h\left\|\sum_{j=1}^N f_{ij}(\mathcal{D}'y^n)y_j^n - f_{ij}(\mathcal{D}x^n)Mx_j^n\right\|_{\ell_2^k}\\
&\le e_i^n + h(1+\varepsilon)\,\|f_i(\mathcal{D}'y^n) - f_i(\mathcal{D}x^n)\|_{\ell_2^d}\\
&\quad + h\sum_{j=1}^N\Big(\|f_{ij}(\mathcal{D}'y^n)y_j^n - f_{ij}(\mathcal{D}'y^n)Mx_j^n\|_{\ell_2^k} + \|f_{ij}(\mathcal{D}'y^n)Mx_j^n - f_{ij}(\mathcal{D}x^n)Mx_j^n\|_{\ell_2^k}\Big)\\
&\le e_i^n + h(1+\varepsilon)\,\|f_i(\mathcal{D}'y^n) - f_i(\mathcal{D}x^n)\|_{\ell_2^d}\\
&\quad + h\sum_{j=1}^N\Big(|f_{ij}(\mathcal{D}'y^n)|e_j^n + (1+\varepsilon)\|x_j^n\|_{\ell_2^d}\cdot|f_{ij}(\mathcal{D}'y^n) - f_{ij}(\mathcal{D}x^n)|\Big).
\end{aligned}$$

Taking the maximum on both sides, this becomes

$$\begin{aligned}
\mathcal{E}^{n+1} &\le \mathcal{E}^n + h(1+\varepsilon)\max_i\|f_i(\mathcal{D}'y^n) - f_i(\mathcal{D}x^n)\|_{\ell_2^d}\\
&\quad + h\mathcal{E}^n\max_i\sum_{j=1}^N|f_{ij}(\mathcal{D}'y^n)| + h(1+\varepsilon)\alpha\cdot\max_i\sum_{j=1}^N|f_{ij}(\mathcal{D}'y^n) - f_{ij}(\mathcal{D}x^n)|.
\end{aligned}$$

We use (1.2)–(1.4) for $a = \mathcal{D}'y^n$ and $b = \mathcal{D}x^n$ to estimate all the terms on the right-hand side. This gives

$$\begin{aligned}\mathcal{E}^{n+1} &\leq \mathcal{E}^n + h(1+\varepsilon)L\|\mathcal{D}'y^n - \mathcal{D}x^n\|_{\ell_\infty^N(\ell_\infty^N)} + h\mathcal{E}^n L' + h(1+\varepsilon)\alpha L''\|\mathcal{D}'y^n - \mathcal{D}x^n\|_{\ell_\infty^N(\ell_\infty^N)}\\ &\leq \mathcal{E}^n(1+hL') + h(1+\varepsilon)(L+\alpha L'')\left[\|\mathcal{D}'y^n - \mathcal{D}'Mx^n\|_{\ell_\infty^N(\ell_\infty^N)} + \|\mathcal{D}'Mx^n - \mathcal{D}x^n\|_{\ell_\infty^N(\ell_\infty^N)}\right]\\ &\leq \mathcal{E}^n(1+hL') + 2h(1+\varepsilon)(L+\alpha L'')(\mathcal{E}^n + \alpha\varepsilon),\end{aligned}$$

where we used (2.8) in the last line. This, together with $\mathcal{E}^0 = 0$, leads to

$$\mathcal{E}^n \leq \varepsilon hnB\exp(hnA),$$

where $A := L' + 2(1+\varepsilon)(L+\alpha L'')$ and $B := 2\alpha(1+\varepsilon)(L+\alpha L'')$. □

2.3. Uniform estimate for the Cucker-Smale model. As a relevant example, let us now show that Theorem 2.2 can be applied to the well-known Cucker-Smale model, introduced and analyzed in [20, 21], which is described by

$$\dot{x}_i = v_i \in \mathbb{R}^d, \tag{2.11}$$

$$\dot{v}_i = \frac{1}{N}\sum_{j=1}^N g(\|x_i - x_j\|_{\ell_2^d})(v_j - v_i), \quad i = 1,\ldots,N. \tag{2.12}$$

The function $g : [0,\infty) \to \mathbb{R}$ is given by $g(s) = \frac{G}{(1+s^2)^\beta}$, for $\beta > 0$, and bounded by $g(0) = G > 0$. This model describes the *emerging of consensus* in a group of interacting agents, trying to *align* (also in terms of abstract consensus) with their neighbors. One of the motivations of the model from Cucker and Smale was to describe the formation and evolution of languages [21, Section 6], although, due to its simplicity, it has been eventually related mainly to the description of the *emergence of flocking* in groups of birds [20]. In the latter case, in fact, spatial and velocity coordinates are sufficient to describe a pointlike agent ($d = 3+3$), while for the evolution of languages, one would have to take into account a much broader dictionary of parameters, hence a higher dimension $d \gg 3+3$ of parameters, which is in fact the case of our interest in the present paper.

Let us show that the model is indeed of the type (1.1). We interprete the system as a group of $2N$ agents in $\mathbb{R}^d$, whose dynamics is given by the following equations

$$\begin{aligned}\dot{x}_i &= \sum_{j=1}^N f_{ij}^x v_j \in \mathbb{R}^d,\\ \dot{v}_i &= \sum_{j=1}^N f_{ij}^v(\mathcal{D}x)v_j, \quad i = 1,\ldots,N\end{aligned}$$

with $f_{ij}^x := \delta_{ij}$, $f_{ii}^v(\mathcal{D}x) := -\frac{1}{N}\sum_{k=1}^N g(\|x_i - x_k\|_{\ell_2^d})$, and $f_{ij}^v(\mathcal{D}x) := \frac{1}{N}g(\|x_i - x_j\|_{\ell_2^d})$, for $i \neq j$. The condition (1.2) is empty, (1.3) reads

$$L' \geq \max(1, 2G) \geq \max_i \left\{1, \frac{2}{N}\sum_{k=1}^N g(\|x_i^n - x_k^n\|_{\ell_2^d})\right\}.$$

Finally,

$$\begin{aligned}
\max_i \frac{2}{N} \sum_{j=1}^{N} &\Big| g(\|x_i^n - x_j^n\|_{\ell_2^d}) - g(\|y_i^n - y_j^n\|_{\ell_2^k}) \Big| \\
&\le \max_i \frac{2\|g\|_{\mathrm{Lip}}}{N} \cdot \sum_{j=1}^{N} \Big| \|x_i^n - x_j^n\|_{\ell_2^d} - \|y_i^n - y_j^n\|_{\ell_2^k} \Big| \\
&\le 2\|g\|_{\mathrm{Lip}} \cdot \|\mathcal{D}' y^n - \mathcal{D} x^n\|_{\ell_\infty^N(\ell_\infty^N)}
\end{aligned}$$

shows that $L'' \le 2\|g\|_{\mathrm{Lip}}$.

2.4. Least-squares estimate of the error for the Cucker-Smale model. The formula (2.10) provides the estimate of the maximum of the individual errors, i.e., $\mathcal{E}^n := \|(y_i^n - Mx_i^n)_{i=1}^N\|_{\ell_\infty^N(\ell_2^k)}$. In this section we address the stronger $\ell_2^N(\ell_2^k)$-estimate for the error. For generic dynamical systems (1.1) such estimate is not available in general, and one has to perform a case-by-case analysis. As a typical example of how to proceed, we restrict ourselves to the Cucker-Smale model, just recalled in the previous section. The forward Euler discretization of (2.11)–(2.12) is given by

$$\begin{aligned}
x_i^{n+1} &= x_i^n + hv_i^n, \\
v_i^{n+1} &= v_i^n + \frac{h}{N} \sum_{j=1}^{N} g(\|x_i^n - x_j^n\|_{\ell_2^d})(v_j^n - v_i^n)
\end{aligned} \tag{2.13}$$

with initial data x_i^0 and v_i^0 given. Let M be again a suitable random matrix in the sense of Lemma 2.1. The Euler method of the projected system is given by the initial conditions $y_i^0 = Mx_i^0$ and $w_i^0 = Mv_i^0$ and the formulas

$$\begin{aligned}
y_i^{n+1} &= y_i^n + hw_i, \\
w_i^{n+1} &= w_i^n + \frac{h}{N} \sum_{j=1}^{N} g(\|y_i^n - y_j^n\|_{\ell_2^k})(w_j^n - w_i^n).
\end{aligned} \tag{2.14}$$

We are interested in the estimates of the following quantities

$$\begin{aligned}
e_{x,i}^n &:= \|y_i^n - Mx_i^n\|_{\ell_2^k}, \quad \mathcal{E}_x^n := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_{x,i}^n)^2} = \frac{\|(y_i^n - Mx_i^n)_{i=1}^N\|_{\ell_2^N(\ell_2^k)}}{\sqrt{N}}, \\
e_{v,i}^n &:= \|w_i^n - Mv_i^n\|_{\ell_2^k}, \quad \mathcal{E}_v^n := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_{v,i}^n)^2} = \frac{\|(w_i^n - Mv_i^n)_{i=1}^N\|_{\ell_2^N(\ell_2^k)}}{\sqrt{N}}.
\end{aligned}$$

Using (2.13) and (2.14), we obtain

$$e_{x,i}^{n+1} \le e_{x,i}^n + he_{v,i}^n \quad \text{and} \quad \mathcal{E}_x^{n+1} \le \mathcal{E}_x^n + h\mathcal{E}_v^n.$$

To bound the quantity $\mathcal{E}_v^n$ we have to work more. Another application of (2.13) and

(2.14) leads to

$$\begin{aligned}
e_{v,i}^{n+1} &\leq e_{v,i}^n + \frac{h}{N}\sum_{j=1}^N \Big(\|g(\|y_i^n - y_j^n\|_{\ell_2^k})(w_j^n - w_i^n) \pm g(\|y_i^n - y_j^n\|_{\ell_2^k})(Mv_j^n - Mv_i^n) \\
&\qquad - g(\|x_i^n - x_j^n\|_{\ell_2^d})(Mv_j^n - Mv_i^n)\|_{\ell_2^k}\Big) \\
&\leq e_{v,i}^n + \frac{h}{N}\sum_{j=1}^N g(\|y_i^n - y_j^n\|_{\ell_2^k})(e_{v,j}^n + e_{v,i}^n) \qquad (2.15)\\
&\qquad + \frac{(1+\varepsilon)h\|g\|_{\mathrm{Lip}}}{N}\cdot\sum_{j=1}^N \|v_j^n - v_i^n\|_{\ell_2^d}\cdot\big|\|x_i^n - x_j^n\|_{\ell_2^d} - \|y_i^n - y_j^n\|_{\ell_2^k}\big|.
\end{aligned}$$

We estimate the first summand in (2.15)

$$\frac{h}{N}\sum_{j=1}^N g(\|y_i^n - y_j^n\|_{\ell_2^k})(e_{v,j}^n + e_{v,i}^n) \leq \frac{hG}{N}\big[Ne_{v,i}^n + \sum_{j=1}^N e_{v,j}^n\big] = hGe_{v,i}^n + \frac{hG}{N}\sum_{j=1}^N e_{v,j}^n$$

and its ℓ_2-norm with respect to i by Hölder's inequality

$$h\sqrt{N}G\mathcal{E}_v^n + \frac{hG}{N}\left(\sum_{i=1}^N\Big(\sum_{j=1}^N e_{v,j}^n\Big)^2\right)^{1/2} \leq 2h\sqrt{N}G\mathcal{E}_v^n. \qquad (2.16)$$

To estimate the second summand in (2.15), let us set $V := \max_{i,j,n}\|v_i^n - v_j^n\|_{\ell_2^d}$ and make use of

$$\begin{aligned}
&\big|\|x_i^n - x_j^n\|_{\ell_2^d} - \|y_i^n - y_j^n\|_{\ell_2^k}\big| \\
&\quad\leq \big|\|x_i^n - x_j^n\|_{\ell_2^d} - \|Mx_i^n - Mx_j^n\|_{\ell_2^k}\big| + \big|\|Mx_i^n - Mx_j^n\|_{\ell_2^k} - \|y_i^n - y_j^n\|_{\ell_2^k}\big| \\
&\quad\leq \varepsilon\|x_i^n - x_j^n\|_{\ell_2^d} + e_{x,i}^n + e_{x,j}^n.
\end{aligned}$$

We arrive at

$$\begin{aligned}
&\frac{(1+\varepsilon)h\|g\|_{\mathrm{Lip}}}{N}\sum_{j=1}^N \|v_j^n - v_i^n\|_{\ell_2^d}(\varepsilon\|x_i^n - x_j^n\|_{\ell_2^d} + e_{x,i}^n + e_{x,j}^n) \\
&\qquad \leq \frac{(1+\varepsilon)h\|g\|_{\mathrm{Lip}}V}{N}\left\{\varepsilon\sum_{j=1}^N\|x_i^n - x_j^n\|_{\ell_2^d} + Ne_{x,i}^n + \sum_{j=1}^N e_{x,j}^n\right\}.
\end{aligned}$$

The ℓ_2-norm of this expression with respect to i is bounded by

$$\begin{aligned}
&\frac{(1+\varepsilon)h\|g\|_{\mathrm{Lip}}V}{N}\left\{\varepsilon\Big(\sum_{i=1}^N\Big(\sum_{j=1}^N\|x_i^n - x_j^n\|_{\ell_2^d}\Big)^2\Big)^{1/2} + N\Big(\sum_{i=1}^N (e_{x,i}^n)^2\Big)^{1/2} + \sqrt{N}\sum_{j=1}^N e_{x,j}^n\right\} \\
&\leq (1+\varepsilon)h\|g\|_{\mathrm{Lip}}V\sqrt{N}(\varepsilon X + 2\mathcal{E}_x^n), \qquad (2.17)
\end{aligned}$$

where $X := \max_{i,j,n}\|x_i^n - x_j^n\|_{\ell_2^d}$. Combining (2.15) with (2.16) and (2.17) leads to the recursive estimate

$$\begin{aligned}
\mathcal{E}_x^{n+1} &\leq \mathcal{E}_x^n + h\mathcal{E}_v^n, \qquad (2.18)\\
\mathcal{E}_v^{n+1} &\leq \mathcal{E}_v^n + 2hG\mathcal{E}_v^n + h(1+\varepsilon)\|g\|_{\mathrm{Lip}}V\left\{\varepsilon X + 2\mathcal{E}_x^n\right\},
\end{aligned}$$

which we put into the matrix form

$$\begin{pmatrix} \mathcal{E}_x^{n+1} \\ \mathcal{E}_v^{n+1} \end{pmatrix} = \mathcal{A} \begin{pmatrix} \mathcal{E}_x^{n} \\ \mathcal{E}_v^{n} \end{pmatrix} + \begin{pmatrix} 0 \\ (1+\varepsilon)\varepsilon h \|g\|_{\mathrm{Lip}} V X \end{pmatrix}, \tag{2.19}$$

where $\mathcal{A}$ is a 2×2 matrix given by

$$\mathcal{A} = \mathcal{A}_1 + h\mathcal{A}_2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + h \begin{pmatrix} 0 & 1 \\ 2(1+\varepsilon)\|g\|_{\mathrm{Lip}} V & 2G \end{pmatrix}.$$

Taking the norms on both sides of (2.19) leads to

$$\sqrt{(\mathcal{E}_x^{n+1})^2 + (\mathcal{E}_v^{n+1})^2} \leq (1 + h\|\mathcal{A}_2\|)\sqrt{(\mathcal{E}_x^{n})^2 + (\mathcal{E}_v^{n})^2} + \varepsilon(1+\varepsilon)h\|g\|_{\mathrm{Lip}} V X$$

and the least-squares error estimate finally reads as follows.

$$\sqrt{(\mathcal{E}_x^{n})^2 + (\mathcal{E}_v^{n})^2} \leq \varepsilon(1+\varepsilon)hn\|g\|_{\mathrm{Lip}} V X \exp(hn\|\mathcal{A}_2\|).$$

3. Dimensionality reduction for continuous dynamical systems.

3.1. Uniform estimates for continuous dynamical systems. In this section we shall establish the analogue of the above results for the continuous time setting of dynamical systems of the type (1.1),

$$\dot{x}_i = f_i(\mathcal{D}x) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}x)x_j\,, \qquad i = 1,\dots,N\,, \tag{3.1}$$

$$x_i(0) = x_i^0\,, \qquad i = 1,\dots,N\,. \tag{3.2}$$

We adopt again the assumptions about Lipschitz continuity and boundedness of the right-hand side made in Section 2, namely (1.2), (1.3) and (1.4).

Theorem 3.1. *Let $x(t) \in \mathbb{R}^{d\times N}$, $t \in [0,T]$, be the solution of the system* (3.1)–(3.2) *with f_i's and f_{ij}'s satisfying* (1.2)–(1.4), *such that*

$$\max_{t\in[0,T]} \max_{i,j} \|x_i(t) - x_j(t)\|_{\ell_2^d} \leq \alpha\,. \tag{3.3}$$

Let us fix $k \in \mathbb{N}$, $k \leq d$, and a matrix $M \in \mathbb{R}^{k\times d}$ such that

$$(1-\varepsilon)\|x_i(t) - x_j(t)\|_{\ell_2^d} \leq \|Mx_i(t) - Mx_j(t)\|_{\ell_2^k} \leq (1+\varepsilon)\|x_i(t) - x_j(t)\|_{\ell_2^d}\,, \tag{3.4}$$

for all $t \in [0,T]$ and i, $j = 1,\dots,N$. Let $y(t) \in \mathbb{R}^{k\times N}$, $t \in [0,T]$ be the solution of the projected system

$$\dot{y}_i = Mf_i(\mathcal{D}'y) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}'y)y_j\,, \qquad i = 1,\dots,N\,,$$

$$y_i(0) = Mx_i^0\,, \qquad i = 1,\dots,N\,, \tag{3.5}$$

such that for a suitable $\beta > 0$,

$$\max_{t\in[0,T]} \|y(t)\|_{\ell_\infty^N(\ell_2^d)} \leq \beta\,. \tag{3.6}$$

Let us define the column-wise ℓ_2-error $e_i(t) := \|y_i - Mx_i\|_{\ell_2^k}$ for $i = 1, \dots, N$ and

$$\mathcal{E}(t) := \max_{i=1,\dots,N} e_i(t) = \|y - Mx\|_{\ell_\infty^N(\ell_2^k)} \,.$$

Then we have the estimate

$$\mathcal{E}(t) \leq \varepsilon\alpha t(L\|M\| + L''\beta)\exp\left[(2L\|M\| + 2\beta L'' + L')t\right] \,. \tag{3.7}$$

Proof. Due to (1.2)–(1.4), we have for every $i = 1, \dots, N$ the estimate

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}e_i &= \frac{\langle y_i - Mx_i, \frac{\mathrm{d}}{\mathrm{d}t}(y_i - Mx_i)\rangle}{\|y_i - Mx_i\|_{\ell_2^k}} \leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}(y_i - Mx_i)\right\|_{\ell_2^k} \\ &\leq \|Mf_i(\mathcal{D}'y) - Mf_i(\mathcal{D}x)\|_{\ell_2^k} + \sum_{j=1}^N \|f_{ij}(\mathcal{D}'y)y_j - f_{ij}(\mathcal{D}x)Mx_j\|_{\ell_2^k} \\ &\leq L\|M\|\,\|\mathcal{D}'y - \mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)} + \sum_{j=1}^N \Big(\|f_{ij}(\mathcal{D}x)(Mx_j - y_j)\|_{\ell_2^k} + \|(f_{ij}(\mathcal{D}x) - f_{ij}(\mathcal{D}'y))y_j\|_{\ell_2^k}\Big) \\ &\leq L\|M\|\,\|\mathcal{D}'y - \mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)} + L'\|Mx - y\|_{\ell_\infty^N(\ell_2^k)} + L''\|\mathcal{D}x - \mathcal{D}'y\|_{\ell_\infty^N(\ell_\infty^N)}\|y\|_{\ell_\infty^N(\ell_2^k)} \,.\end{aligned}$$

The term $\|\mathcal{D}'y - \mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)} \leq \|\mathcal{D}'y - \mathcal{D}'Mx\|_{\ell_\infty^N(\ell_\infty^N)} + \|\mathcal{D}'Mx - \mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)}$ is estimated by

$$\begin{aligned}\|\mathcal{D}'y - \mathcal{D}Mx\|_{\ell_\infty^N(\ell_\infty^N)} &= \max_{i,j}\Big|\|y_i - y_j\|_{\ell_2^k} - \|Mx_i - Mx_j\|_{\ell_2^k}\Big| \\ &\leq \max_{i,j}\|y_i - Mx_i\|_{\ell_2^k} + \|y_j - Mx_j\|_{\ell_2^k} \leq 2\mathcal{E}(t)\,,\end{aligned}$$

and, using the assumption (3.4),

$$\|\mathcal{D}'Mx - \mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)} = \max_{i,j}\Big|\|Mx_i - Mx_j\|_{\ell_2^k} - \|x_i - x_j\|_{\ell_2^d}\Big| \leq \varepsilon\max_{i,j}\|x_i - x_j\|_{\ell_2^k} = \varepsilon\|\mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)} \,.$$

Finally, by the a priori estimate (3.3) for $\|\mathcal{D}x\|_{\ell_\infty^N(\ell_\infty^N)}$ and (3.6) for $\|y\|_{\ell_\infty^N(\ell_2^d)}$, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}e_i &\leq L\|M\|(2\mathcal{E}(t) + \varepsilon\alpha) + L'\mathcal{E}(t) + L''\beta(2\mathcal{E}(t) + \varepsilon\alpha) \\ &= (2L\|M\| + 2\beta L'' + L')\mathcal{E}(t) + \varepsilon\alpha(L\|M\| + L''\beta)\,.\end{aligned}$$

Now, let us split the interval $[0, T)$ into a union of finite disjoint intervals $I_j = [t_{j-1}, t_j)$, $j = 1, \dots, K$ for a suitable $K \in \mathbb{N}$, such that $\mathcal{E}(t) = e_{i(j)}(t)$ for $t \in I_j$. Consequently, on every I_j we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(t) = \frac{\mathrm{d}}{\mathrm{d}t}e_{i(j)}(t) \leq (2L\|M\| + 2\beta L'' + L')\mathcal{E}(t) + \varepsilon\alpha(L\|M\| + L''\beta)\,,$$

and the Gronwall lemma yields

$$\mathcal{E}(t) \leq [\varepsilon\alpha(L\|M\| + L''\beta)(t - t_{j-1}) + \mathcal{E}(t_{j-1})]\exp\left((2L\|M\| + 2\beta L'' + L')(t - t_{j-1})\right)$$

for $t \in [t_{j-1}, t_j)$. A concatenation of these estimates over the intervals I_j leads finally to the expected error estimate

$$\mathcal{E}(t) \leq \varepsilon\alpha t(L\|M\| + L''\beta)\exp\left[(2L\|M\| + 2\beta L'' + L')t\right] \,.$$

□

3.2. A continuous Johnson-Lindenstrauss Lemma. Let us now go through the assumptions we made in the formulation of Theorem 3.1 and discuss how they restrict the validity and applicability of the result. First of all, let us mention that (3.3) and (3.6) can be easily proven to hold for locally Lipschitz right-hand sides f_i and f_{ij} on finite time intervals. Obviously, the critical point for the applicability of Theorem 3.1 is the question how to find a matrix M satisfying the condition (3.4), i.e., being a quasi-isometry along the trajectory solution $x(t)$ for *every* $t \in [0,T]$. The answer is provided by the following generalization of the Johnson-Lindenstrauss Lemma (Lemma 2.1) for rectifiable $\mathcal{C}^1$-curves, by a suitable continuity argument. Let us stress that our approach resembles the "sampling and ϵ-net" argument in [3, 4, 48] for the extension of the quasi-isometry property of Johnson-Lindenstrauss embeddings to smooth Riemmanian manifolds. From this point of view the following result can be viewed as a specification of the work [4, 48].

We first prove an auxiliary technical result:

LEMMA 3.2. *Let* $0 < \varepsilon < \varepsilon' < 1$*,* $a \in \mathbb{R}^d$ *and let* $M : \mathbb{R}^d \to \mathbb{R}^k$ *be a linear mapping such that*

$$(1-\varepsilon)\|a\|_{\ell_2^d} \le \|Ma\|_{\ell_2^k} \le (1+\varepsilon)\|a\|_{\ell_2^d}.$$

Let $x \in \mathbb{R}^d$ *satisfy*

$$\|a-x\| \le \frac{(\varepsilon'-\varepsilon)\|a\|_{\ell_2^d}}{\|M\|+1+\varepsilon'}. \tag{3.8}$$

Then

$$(1-\varepsilon')\|x\|_{\ell_2^d} \le \|Mx\|_{\ell_2^k} \le (1+\varepsilon')\|x\|_{\ell_2^d}. \tag{3.9}$$

Proof. If $a = 0$, the statement is trivial. If $a \neq 0$, we denote the right-hand side of (3.8) by $\tau > 0$ and estimate by the triangle inequality

$$\begin{aligned}\frac{\|Mx\|_{\ell_2^k}}{\|x\|_{\ell_2^d}} &= \frac{\|M(x-a)+Ma\|_{\ell_2^k}}{\|x-a+a\|_{\ell_2^d}} \le \frac{\|M\|\cdot\|x-a\|_{\ell_2^d}+(1+\varepsilon)\|a\|_{\ell_2^d}}{\|a\|_{\ell_2^d}-\|x-a\|_{\ell_2^d}}\\ &\le \frac{\|M\|\cdot\tau+(1+\varepsilon)\|a\|_{\ell_2^d}}{\|a\|_{\ell_2^d}-\tau} \le 1+\varepsilon'.\end{aligned}$$

A similar chain of inequalities holds for the estimate from below. □

Now we are ready to establish a continuous version of Lemma 2.1.

THEOREM 3.3. *Let* $\varphi : [0,1] \to \mathbb{R}^d$ *be a* $\mathcal{C}^1$ *curve. Let* $0 < \varepsilon < \varepsilon' < 1$*,*

$$\gamma := \max_{\xi\in[0,1]} \frac{\|\varphi'(\xi)\|_{\ell_2^d}}{\|\varphi(\xi)\|_{\ell_2^d}} < \infty \quad \textit{and} \quad \mathcal{N} \ge (\sqrt{d}+2)\cdot\frac{\gamma}{\varepsilon'-\varepsilon}.$$

Let k be such that a randomly chosen (and properly normalized) projector M satisfies the statement of the Johnson-Lindenstrauss Lemma 2.1 with ε, d, k and $\mathcal{N}$ arbitrary points with high probability. Without loss of generality we assume that $\|M\| \le \sqrt{d/k}$ within the same probability (this is in fact the case, e.g., for the examples of Gaussian and Bernoulli random matrices reported in Section 2).

Then

$$(1-\varepsilon')\|\varphi(t)\|_{\ell_2^d} \le \|M\varphi(t)\|_{\ell_2^k} \le (1+\varepsilon')\|\varphi(t)\|_{\ell_2^d}, \ \textit{for all}\ t \in [0,1] \tag{3.10}$$

holds with the same probability.

Proof. Let $t_i = i/\mathcal{N}$, $i = 0, \dots, \mathcal{N}$ and put

$$T_i := \arg\max_{\xi \in [t_i, t_{i+1}]} \|\varphi'(\xi)\|_{\ell_2^d}, \quad i = 0, \dots, \mathcal{N} - 1.$$

Let $M : \mathbb{R}^d \to \mathbb{R}^k$ be the randomly chosen and normalized projector (see Lemma 2.1). Hence $\|M\| \leq \sqrt{d/k}$ and

$$(1 - \varepsilon')\|\varphi(T_i)\|_{\ell_2^d} \leq \|M(\varphi(T_i))\|_{\ell_2^k} \leq (1 + \varepsilon')\|\varphi(T_i)\|_{\ell_2^d}, \qquad i = 1, \dots, \mathcal{N} \tag{3.11}$$

with high probability. We show that (3.10) holds with (at least) the same probability.

This follows easily from (3.11) and the following estimate, which holds for every $t \in [t_i, t_{i+1}]$,

$$\begin{aligned}\|\varphi(t) - \varphi(T_i)\|_{\ell_2^d} &\leq \int_t^{T_i} \|\varphi'(s)\|_{\ell_2^d} ds \leq \frac{\|\varphi'(T_i)\|_{\ell_2^d}}{\mathcal{N}} \leq \frac{\|\varphi'(T_i)\|_{\ell_2^d}(\varepsilon' - \varepsilon)}{\gamma(\sqrt{d} + 2)} \\ &\leq \frac{\|\varphi(T_i)\|_{\ell_2^d}(\varepsilon' - \varepsilon)}{\sqrt{d} + 2} \leq \frac{\|\varphi(T_i)\|_{\ell_2^d}(\varepsilon' - \varepsilon)}{\|M\| + 1 + \varepsilon'}.\end{aligned}$$

The proof is then finished by a straightforward application of Lemma 3.2. □

Remark 1. *We show now that the condition*

$$\gamma := \max_{\xi \in [0,1]} \frac{\|\varphi'(\xi)\|_{\ell_2^d}}{\|\varphi(\xi)\|_{\ell_2^d}} < \infty$$

is necessary, hence it is a restriction to the type of curves one can quasi-isometrically project. Let $d \geq 3$. It is known that there is a continuous curve $\varphi : [0,1] \to [0,1]^{d-1}$, such that $\varphi([0,1]) = [0,1]^{d-1}$, i.e., φ goes onto *$[0,1]^{d-1}$. The construction of such a* space-filling *curve goes back to Peano and Hilbert. After a composition with suitable dilations and d-dimensional spherical coordinates we observe that there is also a* surjective *continuous curve $\varphi : [0,1] \to \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ denotes the ℓ_2^d unit sphere in* $\mathbb{R}^d$.

As M was supposed to be a projection, (3.10) *cannot hold for all t's with $\varphi(t) \in$* ker $M \neq \emptyset$.

Obviously, the key condition for applicability of Theorem 3.3 for finding a projection matrix M satisfying (3.4) is that

$$\sup_{t \in [0,T]} \max_{i,j} \frac{\|\dot{x}_i - \dot{x}_j\|_{\ell_2^d}}{\|x_i - x_j\|_{\ell_2^d}} \leq \gamma < \infty\,. \tag{3.12}$$

This condition is, for instance, trivially satisfied when the right-hand sides f_i's and f_{ij}'s have the following Lipschitz continuity:

$$\begin{aligned}\|f_i(\mathcal{D}x) - f_j(\mathcal{D}x)\|_{\ell_2^d} &\leq L'''\|x_i - x_j\|_{\ell_2^d} \qquad \text{for all } i, j = 1, \dots, N\,, \\ |f_{i,k}(\mathcal{D}x) - f_{j,k}(\mathcal{D}x)| &\leq L''''\|x_i - x_j\|_{\ell_2^d} \qquad \text{for all } i, j, k = 1, \dots, N.\end{aligned}$$

We will show in the examples below how condition (3.12) is verified in cases of dynamical systems modeling standard social mechanisms of *attraction, repulsion, aggregation* and *alignment.*

3.3. Applicability to fundamental examples of dynamical systems describing social dynamics. In this section we show the applicability of our dimensionality reduction theory to well-known dynamical systems driven by "social forces" of *alignment, attraction, repulsion* and *aggregation*. Although these models were proposed as descriptions of *group motion in physical space*, the fundamental social effects can be considered as building blocks in the more abstract context of many parameter social dynamics. It has been shown [14, 41] that these models are able to produce meaningful *patterns*, for instance *mills* in two spatial dimensions (see Figure 3.1), reproducing the behavior of certain biological species. However, we should expect that in higher dimension the possible patterns produced by the combination of fundamental effects can be much more complex.

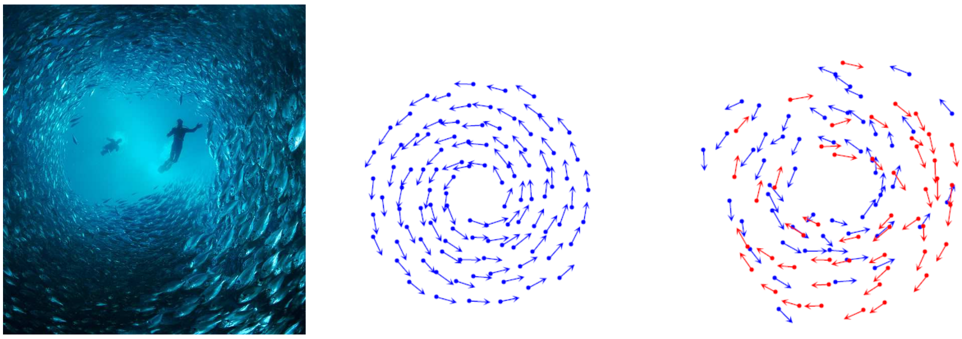


FIG. 3.1. *Mills in nature and in models*

3.3.1. The Cucker-Smale system (alignment effect). As shown in Section 2, the Cucker and Smale flocking model (2.11)–(2.12) is of the type (1.1) and satisfies the Lipschitz continuity assumptions (1.2)–(1.4). Therefore, to meet all the assumptions of Theorem 3.1, we only need to check that it also satisfies the condition (3.12). However, for this we need to consider a slightly different framework than in Section 2.3; instead of considering the $2N$ d-dimensional variables (N position variables and N velocity variables), we need to arrange the model as N variables in $\mathbb{R}^{2d}$, each variable consisting of the position part (first d entries) and of the velocity part (the other d entries). We have then

$$\begin{aligned}\|\dot{x}_i-\dot{x}_j\|_{\ell_2^d}+\|\dot{v}_i-\dot{v}_j\|_{\ell_2^d} &\le \|v_i-v_j\|_{\ell_2^d}+\frac{1}{N}\sum_{k=1}^{N}\left|g(\|x_i-x_k\|_{\ell_2^d})-g(\|x_j-x_k\|_{\ell_2^d})\right|\|v_k\|_{\ell_2^d}\\ &\le \|v_i-v_j\|_{\ell_2^d}+\frac{\|g\|_{Lip}}{N}\sum_{k=1}^{N}\left|\|x_i-x_k\|_{\ell_2^d}-\|x_j-x_k\|_{\ell_2^d}\right|\|v_k\|_{\ell_2^d}\\ &\le \|v_i-v_j\|_{\ell_2^d}+\frac{\|g\|_{Lip}}{N}\left(\sum_{k=1}^{N}\|v_k\|_{\ell_2^d}\right)\|x_i-x_j\|_{\ell_2^d}\\ &\le \|v_i-v_j\|_{\ell_2^d}+c\|x_i-x_j\|_{\ell_2^d}\,,\end{aligned}$$

for a suitable constant c depending on the initial data. We used here the a-priori boundedness of the term $\frac{1}{N}\left(\sum_{k=1}^{N}\|v_k\|_{\ell_2^d}\right)$, see [21] or [33] for details. Consequently, we can satisfy (3.12) with $\gamma = \max(1, c)$.

3.3.2. D'Orsogna model, gravitational and electrostatic interaction (attraction and repulsion effects). Another practically relevant model which fits into the class given by (1.1) is the so-called D'Orsogna model of flocking, [41]:

$$\dot{x}_i = v_i\,, \tag{3.13}$$

$$\dot{v}_i = (a - b\|v_i\|^2_{\ell_2^d})v_i - \frac{1}{N}\sum_{j\neq i}\nabla_{x_i}U(\|x_i - x_j\|_{\ell_2^d})\,, \qquad i = 1, \ldots, N, \tag{3.14}$$

where a and b are positive constants and $U : [0, \infty) \to \mathbb{R}$ is a smooth potential. We denote $u(s) = U'(s)/s$ and assume that u is a bounded, Lipschitz continuous function. We again arrange the model as a system of N variables in $\mathbb{R}^{2d}$, each variable consisting of the position part (first d entries) and of the velocity part (the other d entries). Consequently, the model can be put into a form compliant with (1.1) as follows:

$$\begin{aligned}\dot{x}_i &= \sum_{j=1}^{N} f_{ij}^{xv} v_j\,,\\ \dot{v}_i &= \sum_{j=1}^{N} f_{ij}^{vv}(\mathcal{D}v)v_j + \sum_{j=1}^{N} f_{ij}^{vx}(\mathcal{D}x)x_j\,,\end{aligned}$$

with $f_{ij}^{xv} = \delta_{ij}$, $f_{ii}^{vx}(\mathcal{D}x) = -\frac{1}{N}\sum_{j\neq i} u(\|x_i - x_j\|_{\ell_2^d})$ and $f_{ij}^{vx}(\mathcal{D}x) = \frac{1}{N}u(\|x_i - x_j\|_{\ell_2^d})$ for $i \neq j$. Moreover, we may set $f_{ij}^{vv}(\mathcal{D}v) = \delta_{ij}(a - b\|v_i\|^2_{\ell_2^d})$ by introducing an auxiliary, noninfluential constant zero particle $(x_0, v_0) = (0, 0)$ with null dynamics, i.e., $f_0^{*\star} = 0$ and $f_{0j}^{*\star} = 0$, where $*, \star \in \{x, v\}$. Then, (1.2) is void, while (1.3) is satisfied by

$$\begin{aligned}\max_i \sum_j (|f_{ij}^{xv}(\mathcal{D}x, \mathcal{D}v)| + |f_{ij}^{vx}(\mathcal{D}x, \mathcal{D}v)| + |f_{ij}^{vv}(\mathcal{D}x, \mathcal{D}v)|)&\\ \leq 1 + a + b\max_i \|v_i\|^2_{\ell_2^d} + 2\,\|u\|_{L_\infty} \leq L'\,,&\end{aligned}$$

since the theory provides an apriori bound on $\beta_v := \sup_{t\in[0,T]}\max_i \|v_i\|_{\ell_2^d}$, see [41]. Condition (1.4) for f_{ij}^{xv} is void, while for f_{ij}^{vv} it is satisfied by

$$\begin{aligned}\max_i \sum_j \left|f_{ij}^{vv}(\mathcal{D}v) - f_{ij}^{vv}(\mathcal{D}w)\right| &\leq b\max_i \left|\|v_i\|^2_{\ell_2^d} - \|w_i\|^2_{\ell_2^d}\right|\\ &\leq b\max_i \left(\|v_i\|_{\ell_2^d} + \|w_i\|_{\ell_2^d}\right)\|v_i - w_i\|_{\ell_2^d}\\ &\leq L''\,\|\mathcal{D}v - \mathcal{D}w\|_{\ell_\infty^N(\ell_\infty^N)}\,,\end{aligned}$$

where we again use the apriori boundedness of β_v. For f_{ij}^{vx} is (1.4) satisfied by

$$\begin{aligned}\max_i \sum_j \left|f_{ij}^{vx}(\mathcal{D}x) - f_{ij}^{vx}(\mathcal{D}y)\right| &\leq \max_i \frac{2}{N}\sum_{j\neq i}\left|u(\|x_i - x_j\|_{\ell_2^d}) - u(\|y_i - y_j\|_{\ell_2^d})\right|\\ &\leq \max_i \frac{2}{N}\|u\|_{\mathrm{Lip}}\sum_{j\neq i}\left|\|x_i - x_j\|_{\ell_2^d} - \|y_i - y_j\|_{\ell_2^d}\right|\\ &\leq 2\,\|u\|_{\mathrm{Lip}}\,\|\mathcal{D}x - \mathcal{D}y\|_{\ell_\infty^N(\ell_\infty^N)}\,.\end{aligned}$$

Finally, it can be easily checked that condition (3.12) is satisfied by

$$\|\dot{x}_i - \dot{x}_j\|_{\ell_2^d} + \|\dot{v}_i - \dot{v}_j\|_{\ell_2^d} \leq (1 + a + 3b\beta_v^2)\|v_i - v_j\|_{\ell_2^d} + \Big(\|u\|_{L_\infty} + 2\beta_x \|u\|_{\text{Lip}}\Big) \|x_i - x_j\|_{\ell_2^d} ,$$

where $\beta_x := \sup_{t\in[0,T]} \max_i \|x_i\|_{\ell_2^d}$.

In fact, the D'Orsogna model is a generalization of the classical model of interacting particles through a potential U,

$$\begin{aligned} \dot{x}_i &= v_i , \qquad i = 1, \dots, N , \\ \dot{v}_i &= -\frac{1}{N} \sum_{j\neq i} \nabla U(\|x_i - x_j\|_{\ell_2^d}) , \qquad i = 1, \dots, N , \end{aligned}$$

for instance, gravitational or electrostatic interaction. However, in these cases the function $u(s) = U'(s)/s$ does not meet the assumptions of boundedness and Lipschitz continuity that are needed for the applicability of our method. Consequently, we only can consider models with regular enough potentials.

3.4. Recovery of the dynamics in high dimension from multiple simulations in low dimension. The main message of Theorem 3.1 is that, under suitable assumptions on the governing functions f_i, f_{ij}, the trajectory of the solution $y(t)$ of the *projected* dynamical system (3.5) is at an ε error from the trajectory of the *projection* of the solution $x(t)$ of the dynamical system (3.1)-(3.2), i.e.,

$$y_i(t) \approx Mx_i(t) \text{ or, more precisely, } \|Mx_i(t) - y_i(t)\|_{\ell_2^k} \leq C(t)\varepsilon, \quad t \in [0,T]. \quad (3.15)$$

We wonder whether this approximation property can allow us to "learn" properties of the original trajectory $x(t)$ in high dimension.

3.4.1. Sparse recovery. To address this issue we recall first some relevant and useful concepts from the field of *compressed sensing* [25, 28]. Again a central role here is played by (random) matrices with the so-called *Restricted Isometry Property* RIP, cf. [11].

Definition 3.4 (Restricted Isometry Property). *A $k \times d$ matrix M is said to have the Restricted Isometry Property of order $K \leq d$ and level $\delta \in (0,1)$ if*

$$(1-\delta)\|x\|_{\ell_2^d}^2 \leq \|Mx\|_{\ell_2^k}^2 \leq (1+\delta)\|x\|_{\ell_2^d}^2$$

for all K-sparse $x \in \Sigma_K = \{z \in \mathbb{R}^d : \#\text{supp}\,(z) \leq K\}$.

Both the typical matrices used in Johnson-Lindenstrauss embeddings (cf. Lemma 2.1) and matrices with RIP used in compressed sensing are usually generated at random. It was observed by [3] and [37], that there is an intimate connection between these two notions. A simple reformulation of the arguments of [3] yields the following.

Theorem 3.5 (Baraniuk, Davenport, DeVore, and Wakin). *Let M be a $k \times d$ matrix drawn at random which satisfies*

$$(1-\delta/2)\|x\|_{\ell_2^d}^2 \leq \|Mx\|_{\ell_2^k}^2 \leq (1+\delta/2)\|x\|_{\ell_2^d}^2, \quad x \in \mathcal{P}$$

for every set $\mathcal{P} \subset \mathbb{R}^d$ with $\#\mathcal{P} \leq \left(\frac{12ed}{\delta K}\right)^K$ with probability $0 < \nu < 1$. Then M satisfies the Restricted Isometry Property of order K and level $\delta/3$ with probability at least equal to ν.

Combined with several rather elementary constructions of Johnson-Lindenstrauss embedding matrices available in literature, cf. [1] and [22], this result provides a simple construction of RIP matrices. The converse direction, namely the way from RIP matrices to matrices suitable for Johnson-Lindenstrauss embedding was discovered only recently in [37].

Theorem 3.6 (Krahmer and Ward). *Fix $\eta > 0$ and $\varepsilon > 0$, and consider a finite set $\mathcal{P} \subset \mathbb{R}^d$ of cardinality $|\mathcal{P}| = \mathcal{N}$. Set $K \geq 40 \log \frac{4\mathcal{N}}{\eta}$, and suppose that the $k \times d$ matrix $\tilde{M}$ satisfies the Restricted Isometry Property of order K and level $\delta \leq \varepsilon/4$. Let $\xi \in \mathbb{R}^d$ be a Rademacher sequence, i.e., uniformly distributed on $\{-1,1\}^d$. Then with probability exceeding $1-\eta$,*

$$(1-\varepsilon)\|x\|^2_{\ell_2^d} \leq \|Mx\|^2_{\ell_2^k} \leq (1+\varepsilon)\|x\|^2_{\ell_2^d}.$$

uniformly for all $x \in \mathcal{P}$, where $M := \tilde{M}\operatorname{diag}(\xi)$, where $\operatorname{diag}(\xi)$ is a $d \times d$ diagonal matrix with ξ on the diagonal.

We refer to [42] for additional details.

Remark 2. *Notice that M as constructed in Theorem 3.6 is both a Johnson-Lindenstrauss embedding and a matrix with RIP, because*

$$\begin{aligned}(1-\delta)\|x\|^2_{\ell_2^d} = (1-\delta)\|\operatorname{diag}(\xi)x\|^2_{\ell_2^d} &\leq \|\underbrace{\tilde{M}\operatorname{diag}(\xi)}_{:=M}x\|^2_{\ell_2^k}\\ &\leq (1+\delta)\|\operatorname{diag}(\xi)x\|^2_{\ell_2^d} = (1+\delta)\|x\|^2_{\ell_2^d}.\end{aligned}$$

The matrices considered in Section 2 satisfy with high probability the RIP with

$$K = \mathcal{O}\left(\frac{k}{1+\log(d/k)}\right).$$

Equipped with the notion of RIP matrices we may state the main result of the theory of compressed sensing, as appearing in [25], which we shall use for the recovery of the dynamical system in $\mathbb{R}^d$.

Theorem 3.7. *Assume that the matrix $M \in \mathbb{R}^{k\times d}$ has the RIP of order $2K$ and level*

$$\delta_{2K} < \frac{2}{3+\sqrt{7/4}} \approx 0.4627.$$

Then the following holds for all $x \in \mathbb{R}^d$. Let the low-dimensional approximation $y = Mx + \eta$ be given with $\|\eta\|_{\ell_2^k} \leq C\varepsilon$. Let $x^{\#}$ be the solution of

$$\min_{z\in\mathbb{R}^d} \|z\|_{\ell_1^d} \quad \text{subject to } \|Mz-y\|_{\ell_2^k} \leq \|\eta\|_{\ell_2^k}. \tag{3.16}$$

Then

$$\|x - x^{\#}\|_{\ell_2^d} \leq C_1\varepsilon + C_2 \frac{\sigma_K(x)_{\ell_1^d}}{\sqrt{K}}$$

for some constants $C_1, C_2 > 0$ that depend only on δ_{2K}, and $\sigma_K(x)_{\ell_1^d} = \inf_{z:\#\operatorname{supp}(z)\leq K} \|z-x\|_{\ell_1^d}$ is the best-K-term approximation error in ℓ_1^d.

This result says that provided the stability relationship (3.15), we can approximate the individual trajectories $x_i(t)$, for each $t \in [0,T]$ fixed, by a vector $x_i^{\#}(t)$ solution of an optimization problem of the type (3.16), and the accuracy of the approximation depends on the best-K-term approximation error $\sigma_K(x_i(t))_{\ell_1^d}$. Actually, when $x_i(t)$ is a vector in $\mathbb{R}^d$ with few large entries in absolute value, then $x_i^{\#}(t) \approx x_i(t)$ is a very good approximation, up to the ineliminable ε-distortion. However, if the vector $x_i(t)$ has many relevant entries, then this approximation will be rather poor. One possibility to improve the recovery error is to increase the dimension k (leading to a smaller distortion parameter $\varepsilon > 0$ in the Johnson-Lindenstrauss embedding). But we would like to explore another possibility, namely projecting and simulating *in parallel and independently* the dynamical system L-times in the lower dimension k

$$\dot{y}_i^\ell = M^\ell f_i(\mathcal{D}' y^\ell) + \sum_{j=1}^N f_{ij}(\mathcal{D}' y^\ell) y_j^\ell\,, \qquad y_i^\ell(0) = M^\ell x_i^0\,, \quad \ell = 1, \dots, L. \tag{3.17}$$

Let us give a brief overview of the corresponding error estimates. The number of points needed in every of the cases is $\mathcal{N} \approx N \times n_0$, where N is the number of agents and $n_0 = T/h$ is the number of iterations.

- We perform 1 projection and simulation in $\mathbb{R}^k$: Then $\varepsilon = \mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{k}}\Big)$, $K = \mathcal{O}\left(\frac{k}{1+\log(d/k)}\right)$ and an application of Theorem 3.7 leads to

$$\|x_i(t) - x_i^{\#}(t)\|_{\ell_2^d} \le C'(t) \left(\sqrt{\frac{\log \mathcal{N}}{k}} + \frac{\sigma_K(x_i(t))_{\ell_1^d}}{\sqrt{K}} \right). \tag{3.18}$$

  Here, $C'(t)$ combines both the constants from Theorem 3.7 and the time-dependent $C(t)$ from (3.15). So, to reach the precision of order $C'(t)\epsilon > 0$, we have to choose $k \in \mathbb{N}$ large enough, such that $\sqrt{\frac{\log \mathcal{N}}{k}} \le \epsilon$ and $\frac{\sigma_K(x_i(t))_{\ell_1^d}}{\sqrt{K}} \le \epsilon$. We then need $k \times N^2$ operations to evaluate the adjacency matrix.
- We perform 1 projection and simulation in $\mathbb{R}^{L\times k}$: Then $\varepsilon' = \mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{Lk}}\Big)$ and $K' = \mathcal{O}\left(\frac{Lk}{1+\log(d/Lk)}\right)$ and an application of Theorem 3.7 leads to

$$\|x_i(t) - x_i^{\#}(t)\|_{\ell_2^d} \le C'(t) \left(\sqrt{\frac{\log \mathcal{N}}{Lk}} + \frac{\sigma_{K'}(x_i(t))_{\ell_1^d}}{\sqrt{K'}} \right). \tag{3.19}$$

  The given precision of order $C'(t)\epsilon > 0$, may be then reached by choosing $k, L \in \mathbb{N}$ large enough, such that $\sqrt{\frac{\log \mathcal{N}}{Lk}} \le \epsilon$ and $\frac{\sigma_{K'}(x_i(t))_{\ell_1^d}}{\sqrt{K'}} \le \epsilon$. We then need $Lk \times N^2$ operations to evaluate the adjacency matrix.
- We perform L independent and parallel projections and simulations in $\mathbb{R}^k$: Then we assemble the following system corresponding to (3.17)

$$\mathcal{M}x = \begin{pmatrix} M^1 \\ M^2 \\ \dots \\ \dots \\ M^L \end{pmatrix} x_i = \begin{pmatrix} y_i^1 \\ y_i^2 \\ \dots \\ \dots \\ y_i^L \end{pmatrix} - \begin{pmatrix} \eta_i^1 \\ \eta_i^2 \\ \dots \\ \dots \\ \eta_i^L \end{pmatrix},$$

  where for all $\ell = 1, \dots, L$ the matrices $M^\ell \in \mathbb{R}^{k\times d}$ are (let us say) random matrices with each entry generated independently with respect to the properly normalized Gaussian distribution as described in Section 2. Then

$\mathcal{M}/\sqrt{L}$ is a $Lk \times d$ matrix with Restricted Isometry Property of order $K' = \mathcal{O}\left(\frac{Lk}{1+\log(d/Lk)}\right)$ and level $\delta < 0.4627$. The initial distortion of each of the projections is still $\varepsilon = \mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{k}}\Big)$. Therefore, by applying Theorem 3.7, we can compute $x_i^{\#}(t)$ such that

$$\|x_i(t) - x_i^{\#}(t)\|_{\ell_2^d} \leq C'(t)\left(\sqrt{\frac{\log \mathcal{N}}{k}} + \frac{\sigma_{K'}(x_i(t))_{\ell_1^d}}{\sqrt{K'}}\right). \tag{3.20}$$

Notice that the computation of $x_i^{\#}(t)$ can also be performed in parallel, see, e.g., [26]. The larger is the number L of projections we perform, the larger is K' and the smaller is the second summand in (3.20); actually $\sigma_{K'}(x_i(t))_{\ell_1^d}$ vanishes for $K' \geq d$. Unfortunately, the parallelization can not help to reduce the initial distortion $\varepsilon > 0$. To reach again the precision of order $C'(t)\epsilon > 0$, we have to choose $k \in \mathbb{N}$ large enough, such that $\sqrt{\frac{\log \mathcal{N}}{k}} \leq \epsilon$. Then we chose $L \geq 1$ large enough such that $\frac{\sigma_{K'}(x_i(t))_{\ell_1^d}}{\sqrt{K'}} \leq \epsilon$. We again need $k \times N^2$ operations to evaluate the adjacency matrix.

In all three cases, we obtain the estimate

$$\|x_i(t) - x_i^{\#}(t)\|_{\ell_2^d} \leq C'(t)\left(\varepsilon + \frac{\sigma_K(x_i(t))_{\ell_1^d}}{\sqrt{K}}\right), \tag{3.21}$$

where the corresponding values of $\varepsilon > 0$ and K together with the number of operations needed to evaluate the adjacency matrix may be found in the following table.

	ε	K	number of operations
1 projection into $\mathbb{R}^k$	$\mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{k}}\Big)$	$\mathcal{O}\left(\frac{k}{1+\log(d/k)}\right)$	$k \times N^2$
1 projection into $\mathbb{R}^{L\times k}$	$\mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{Lk}}\Big)$	$\mathcal{O}\left(\frac{Lk}{1+\log(d/Lk)}\right)$	$Lk \times N^2$
L projections into $\mathbb{R}^k$	$\mathcal{O}\Big(\sqrt{\frac{\log \mathcal{N}}{k}}\Big)$	$\mathcal{O}\left(\frac{Lk}{1+\log(d/Lk)}\right)$	$k \times N^2$

3.4.2. Manifold recovery. In recent papers [4, 48, 34], the concepts of compressed sensing and sparse recovery were extended to vectors on smooth manifolds. These methods could become very useful in our context if (for any reason) we would have an apriori knowledge that the trajectories $x_i(t)$ keep staying on or near such a smooth manifold. We leave this direction open for future research.

3.5. Numerical experiments. In this section we illustrate the practical use and performances of our projection method for the Cucker-Smale system (2.11)–(2.12). As already mentioned, this system models the emergence of consensus in a group of interacting agents, trying to align with their neighbors. The qualitative behavior of its solutions is formulated by this well known result [20, 21, 33]:

THEOREM 3.8. *Let $(x_i(t), v_i(t))$ be the solutions of* (2.11)–(2.12)*. Let us define the fluctuation of positions around the center of mass $x_c(t) = \frac{1}{N}\sum_{i=1}^N x_i(t)$, and, resp., the fluctuation of the rate of change around its average $v_c(t) = \frac{1}{N}\sum_{i=1}^N v_i(t)$ as*

$$\Lambda(t) = \frac{1}{N}\sum_{i=1}^N \|x_i(t) - x_c(t)\|_{\ell_2^d}^2, \qquad \Gamma(t) = \frac{1}{N}\sum_{i=1}^N \|v_i(t) - v_c(t)\|_{\ell_2^d}^2.$$

Then if either $\beta \leq 1/2$ or the initial fluctuations $\Lambda(0)$ and $\Gamma(0)$ are small enough (see [20] for details), then $\Gamma(t) \to 0$ as $t \to \infty$.

The phenomenon of $\Gamma(t)$ tending to zero as $t \to \infty$ is called *flocking* or *emergence of consensus*. If $\beta > 1/2$ and the initial fluctuations are not small, it is not known whether a given initial configuration will actually lead to flocking or not, and the only way to find out the possible formation of *consensus patterns* is to perform numerical simulations. However, these can be especially costly if the number of agents N and the dimension d are large; the algorithmic complexity of the calculation is $\mathcal{O}(d \times N^2)$. Therefore, a significant reduction of the dimension d, which can be achieved by our projection method, would lead to a corresponding reduction of the computational cost.

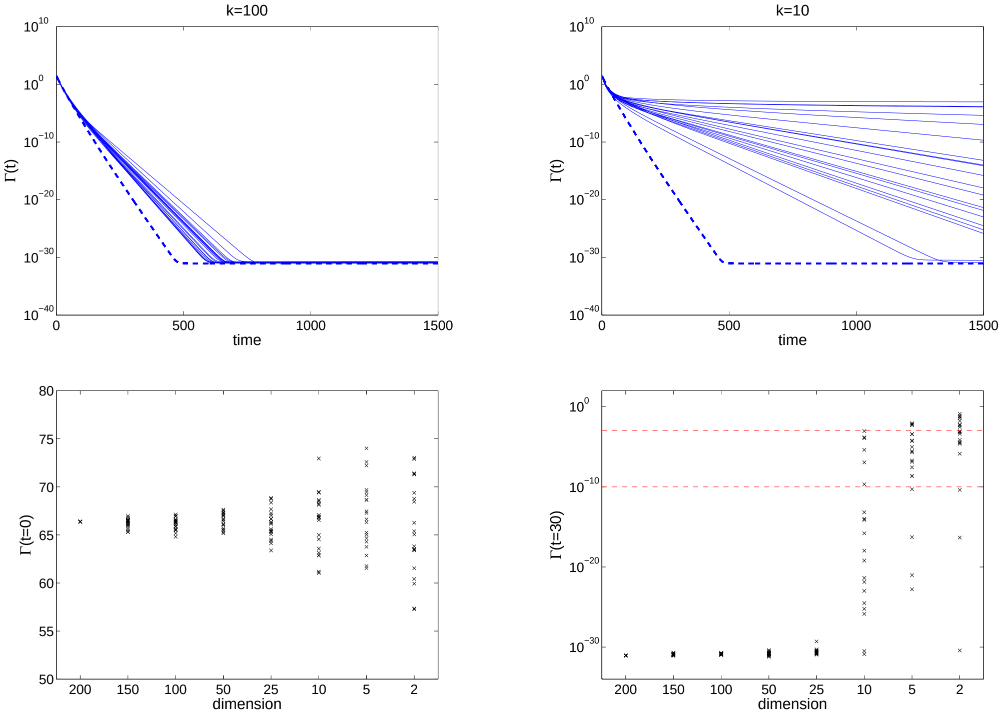


FIG. 3.2. *Numerical results for $\beta = 1.5$: First row shows the evolution of $\Gamma(t)$ of the system projected to dimension $k = 100$ (left) and $k = 10$ (right) in the twenty realizations, compared to the original system (bold dashed line). Second row shows the initial values $\Gamma(t = 0)$ and final values $\Gamma(t = 30)$ in all the performed simulations.*

We illustrate this fact by a numerical experiment, where we choose $N = 1000$ and $d = 200$, i.e., every agent i is determined by a 200-dimensional vector x_i of its state and a 200-dimensional vector v_i giving the rate of change of its state. The initial datum (x^0, v^0) is generated randomly, every component of x^0 being drawn independently from the uniform distribution on $[0, 1]$ and every component of v^0 being drawn independently from the uniform distribution on $[-1, 1]$. We choose $\beta = 1.5$, 1.62 and 1.7, and for each of these values we perform the following set of simulations:

1. Simulation of the original system in 200 dimensions.

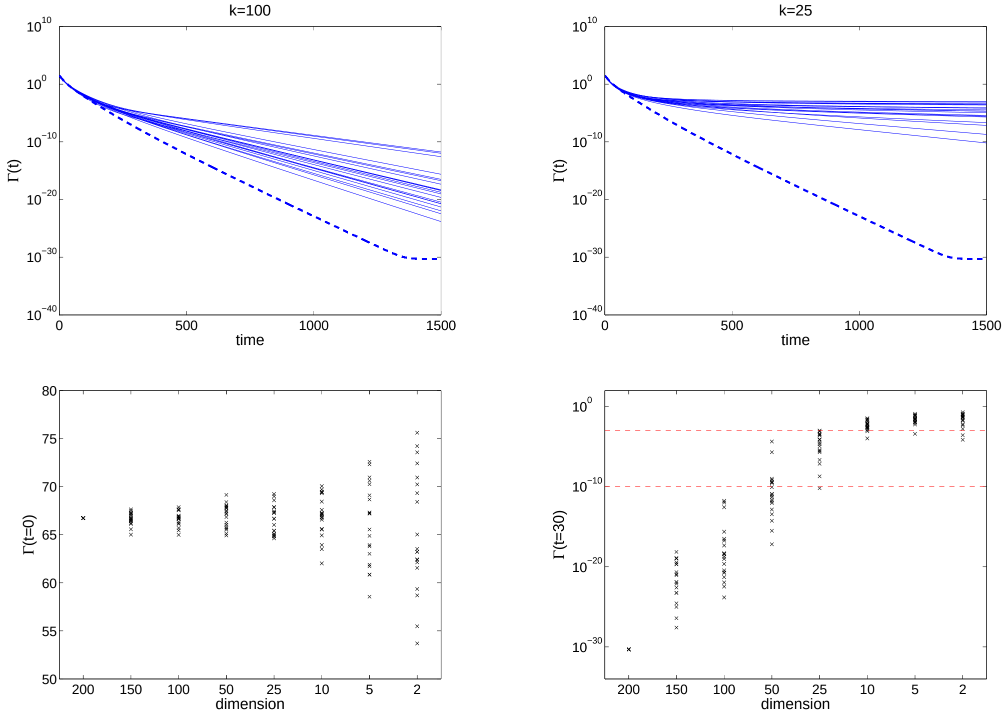


FIG. 3.3. *Numerical results for $\beta = 1.62$: First row shows the evolution of $\Gamma(t)$ of the system projected to dimension $k = 100$ (left) and $k = 25$ (right) in the twenty realizations, compared to the original system (bold dashed line). Second row shows the initial values $\Gamma(t = 0)$ and final values $\Gamma(t = 30)$ in all the performed simulations.*

2. Simulations in lower dimensions k: the initial condition (x^0, v^0) is projected into the k-dimensional space with a random Johnson-Lindenstrauss projection matrix M with Gaussian entries. The dimension k takes the values 150, 100, 50, 25, 10, 5, and 2. For every k, we perform the simulation twenty times, each time with a new random projection matrix M.

All the simulations were implemented in MATLAB, using 1500 steps of the forward Euler method with time step size 0.02. The paths of $\Gamma(t)$ from the twenty experiments with $k = 100$ and $k = 25$ or $k = 10$ are shown in the first rows of Figs. 3.2, 3.3 and, resp., 3.4 for $\beta = 1.5$, 1.62 and, resp., 1.7.

The information we are actually interested in is whether flocking takes place, in other words, whether the fluctuations of velocities $\Gamma(t)$ tend to zero. Typically, after an initial phase, the graph of $\Gamma(t)$ gives a clear indication either about exponentially fast convergence to zero (due to rounding errors, "zero" actually means values of the order 10^{-30} in the simulations) or about convergence to a positive value. However, in certain cases the decay may be very slow and a very long simulation of the system would be needed to see if the limiting value is actually zero or not. Therefore, we propose the following heuristic rules to decide about flocking from numerical simulations:

- If the value of Γ at the final time $t = 30$ is smaller than 10^{-10}, we conclude that flocking took place.

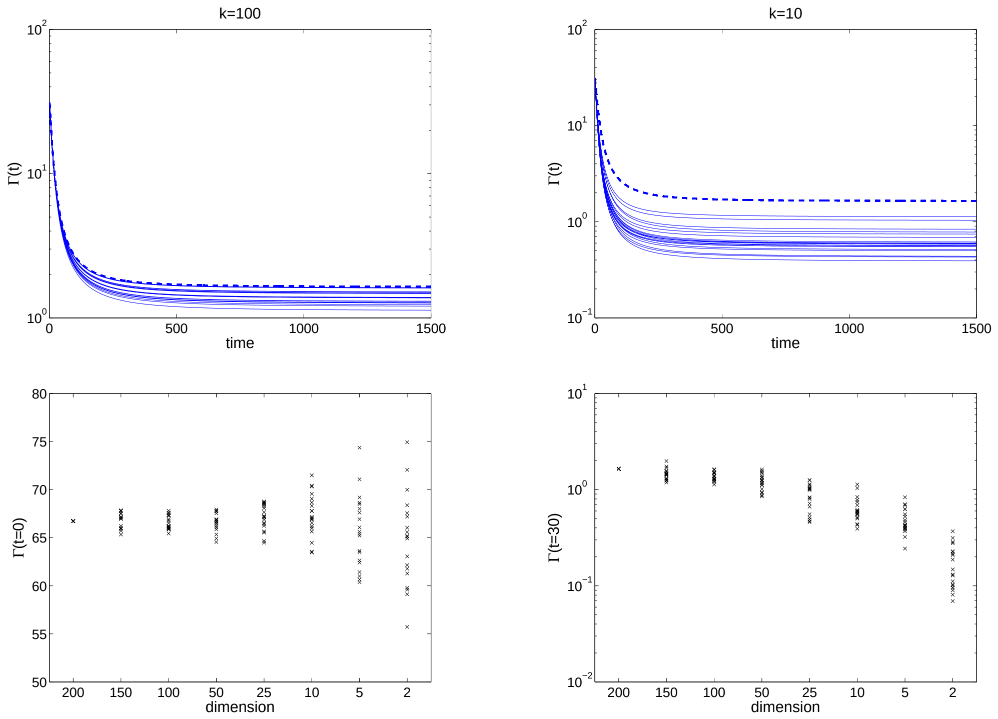


FIG. 3.4. *Numerical results for $\beta = 1.7$: First row shows the evolution of $\Gamma(t)$ of the system projected to dimension $k = 100$ (left) and $k = 10$ (right) in the twenty realizations, compared to the original system (bold dashed line). Second row shows the initial values $\Gamma(t = 0)$ and final values $\Gamma(t = 30)$ in all the performed simulations.*

- If the value of $\Gamma(30)$ is larger than 10^{-3}, we conclude that flocking did not take place.
- Otherwise, we do not make any conclusion.

In the second rows of Figs. 3.2, 3.3 and 3.4 we present the initial and final values of Γ of the twenty simulations for all the dimensions k, together with the original dimension $d = 200$. In accordance with the above rules, flocking takes place if the final value of Γ lies below the lower dashed line, does not take place if it lies above the upper dashed line, otherwise the situation is not conclusive. The results are summarized in Table 3.1.

Experience gained with a large amount of numerical experiments shows the following interesting fact: The flocking behavior of the Cucker-Smale system is very stable with respect to the Johnson-Lindenstrauss projections. Usually, the projected systems show the same flocking behavior as the original one, even if the dimension is reduced dramatically, for instance from $d = 200$ to $k = 10$ (see Figs 3.2 and 3.4). This stability can be roughly explained as follows: Since the flocking behavior depends mainly on the initial values of Γ and Λ, which are statistical properties of the random distributions used for the generation of initial data, and since N is sufficiently large, the concentration of measure phenomenon takes place. Its effect is that the initial values of the fluctuations of the projected data are very close to the original ones, and

	$\beta = 1.5$				$\beta = 1.62$				$\beta = 1.7$		
dim	pos	neg	??	dim	pos	neg	??	dim	pos	neg	??
200	1	0	0	200	1	0	0	200	0	1	0
150	20	0	0	150	20	0	0	150	0	20	0
100	20	0	0	100	20	0	0	100	0	20	0
50	20	0	0	50	13	0	7	50	0	20	0
25	20	0	0	25	1	1	18	25	0	20	0
10	14	0	6	10	0	18	2	10	0	20	0
5	4	4	12	5	0	19	1	5	0	20	0
2	3	8	9	2	0	18	2	2	0	20	0

TABLE 3.1

Statistics of the flocking behaviors of the systems in the original dimension $d = 200$ and in the projected dimensions. With $\beta = 1.5$ and $\beta = 1.62$, the original system ($d = 200$) exhibited flocking behavior. With $\beta = 1.5$, even after random projections into 25 dimensions, the system exhibited flocking in all 20 repetitions of the experiment, and still in 14 cases in dimension 10. With $\beta = 1.62$, the deterioration of the flocking behavior with decreasing dimension was much faster, and already in dimension 25 the situation was not conclusive. This is related to the fact that the value $\beta = 1.62$ was chosen to intentionally bring the system close to the borderline between flocking and non-flocking. Finally, with $\beta = 1.7$, the original system did not flock, and, remarkably, all the projected systems (even to two dimensions) exhibit the same behavior.

thus the flocking behavior is (typically) the same. There is only a narrow interval of values of β (in our case this interval is located around the value $\beta = 1.62$), which is a borderline region between flocking and non-flocking, and the projections to lower dimensions spoil the flocking behavior, see Fig 3.3. Let us note that in our simulations we were only able to detect cases when flocking took place in the original system, but did not take place in some of the projected ones. Interestingly, we never observed the inverse situation, a fact which we are not able to explain satisfactorily. In fact, one can make other interesting observations, deserving further investigation. For instance, Figs. 3.2 and 3.3 show that if the original system exhibits flocking, then the curves of $\Gamma(t)$ of the projected systems tend to lie above the curve of $\Gamma(t)$ of the original one. The situation is reversed if the original system does not flock, see Fig. 3.4.

From a practical point of view, we can make the following conclusion: To obtain an indication about the flocking behavior of a highly dimensional Cucker-Smale system, it is typically satisfactory to perform a limited number of simulations of the system projected into a much lower dimension, and evaluate the statistics of their flocking behavior. If the result is the same for the majority of simulations, one can conclude that the original system very likely has the same flocking behavior as well.

4. Mean-field limit and kinetic equations in high dimension. In the previous sections we were concerned with tractable simulation of the dynamical systems of the type (1.1) when the dimension d of the parameter space is large. Another source of possible intractability in numerical simulations appears in the situation where the number of agents N is very large. Therefore, in the next sections we consider the so-called *mean-field limit* of (1.1) as $N \to \infty$, where the evolution of the system is described by time-dependent probability measures $\mu(t)$ on $\mathbb{R}^d$, representing the density distribution of agents, and satisfying mesoscopic partial differential equations of the type (4.1). This strategy originated from the kinetic theory of gases, see [16] for classical references. We show how our projection method can be applied for dimensionality reduction of the corresponding kinetic equations and explain how the probability measures can be approximated by atomic measures. Using the concepts of *delayed curse of dimension* and *measure quantization* known from optimal integration problems in high dimension, we show that under the assumption that the measure concentrates along low-dimensional subspaces (and in general along low-dimensional sets or manifolds), it can be approximated by atomic measures with sub-exponential (with respect to d) number of atoms. Through such approximation, we shall show that we can approximate suitable random averages of the solution of the original partial differential equation in high dimension by tractable simulations of corresponding solutions of lower-dimensional kinetic equations.

4.1. Formal derivation of mean-field equations. In this section we briefly explain how the mean-field limit description corresponding to (1.1) can be derived. This is given, under suitable assumptions on the family of the governing functions $\mathcal{F}_N = \{f_i, f_{ij} : i, j = 1, \dots N\}$, by the general formula

$$\frac{\partial \mu}{\partial t} + \nabla \cdot (\mathcal{H}_{\mathcal{F}}[\mu]\mu) = 0, \tag{4.1}$$

where $\mathcal{H}_{\mathcal{F}}[\mu]$ is a field in $\mathbb{R}^d$, determined by the sequence $\mathcal{F} = (\mathcal{F}_N)_{N\in\mathbb{N}}$.

In order to provide an explicit example, we show how to formally derive the mean field limit of systems of the type

$$\dot{x}_i = v_i \,, \tag{4.2}$$

$$\dot{v}_i = \sum_{j=1}^{N} f_{ij}^{vv}(\mathcal{D}x, \mathcal{D}v)v_j + \sum_{j=1}^{N} f_{ij}^{vx}(\mathcal{D}x)x_j \,, \tag{4.3}$$

with

$$f_{ij}^{vx}(\mathcal{D}x) = -\frac{\delta_{ij}}{N}\sum_{k\neq i} u(\|x_i - x_k\|_{\ell_2^d}) + \frac{1-\delta_{ij}}{N} u(\|x_i - x_j\|_{\ell_2^d}) \,,$$

$$f_{ij}^{vv}(\mathcal{D}x, \mathcal{D}v) = \delta_{ij}\left(h(\|v_i\|_{\ell_2^d}^2) - \frac{1}{N}\sum_{k=1}^{N} g(\|x_i - x_k\|_{\ell_2^d})\right) + \frac{1-\delta_{ij}}{N} g(\|x_i - x_j\|_{\ell_2^d}) \,.$$

Note that for suitable choices of the functions h, g, u this formalism includes both the Cucker-Smale model (2.11)–(2.12) and D'Orsogna model (3.13)–(3.14). We define the empirical measure associated to the solutions $x_i(t)$, $v_i(t)$ of (4.2)–(4.3) as

$$\mu^N(t) := \mu^N(t, x, v) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i(t)}(x)\delta_{v_i(t)}(v) \,.$$

Taking a smooth, compactly supported test function $\xi \in C_0^\infty(\mathbb{R}^{2d})$ and using (4.2)–(4.3), one easily obtains by a standard formal calculation (see [14])

$$\begin{aligned} \frac{d}{dt}\langle \mu^N(t), \xi\rangle &= \frac{d}{dt}\left(\frac{1}{N}\sum_{i=1}^{N} \xi(x_i(t), v_i(t))\right) \qquad (4.4) \\ &= \int_{\mathbb{R}^{2d}} \nabla_x \xi(x, v) \cdot v \, \mathrm{d}\mu^N(t, x, v) + \int_{\mathbb{R}^{2d}} \nabla_v \xi(x, v) \cdot \mathcal{H}[\mu^N(t)](x, v) \, \mathrm{d}\mu^N(t, x, v) \,, \end{aligned}$$

with

$$\mathcal{H}[\mu](x, v) = h(\|v\|_{\ell_2^d})v + \int_{\mathbb{R}^{2d}} g(\|x - y\|_{\ell_2^d})(w - v) \, \mathrm{d}\mu(y, w) + \int_{\mathbb{R}^{2d}} u(\|x - y\|_{\ell_2^d})(y - x) \, \mathrm{d}\mu(y, w) \,.$$

We now assume weak convergence of a subsequence of $(\mu^N(t))_{N\in\mathbb{N}}$ to a time-dependent measure $\mu(t) = \mu(t,x,v)$ and boundedness of its first order moment, which indeed can be established rigorously for the Cucker-Smale and D'Orsogna systems (see [33], [41]). Then, passing to the limit $N \to \infty$ in (4.4), one obtains in the strong formulation that μ is governed by

$$\frac{\partial \mu}{\partial t}(t,x,v) + v \cdot \nabla_x \mu(t,x,v) + \nabla_v \cdot (\mathcal{H}[\mu(t)](x,v)\mu(t,x,v)) = 0\,,$$

which is an instance of the general prototype (4.1).

Using the same formal arguments as described above, one can easily derive mean field limit equations corresponding to (1.1) with different choices of the family $\mathcal{F}$.

4.2. Monge-Kantorovich-Rubinstein distance and stability. In several relevant cases, including the Cucker-Smale and D'Orsogna systems [13], solutions of equations of the type (4.1) are stable with respect to suitable distances. We consider the space $\mathcal{P}_1(\mathbb{R}^d)$, consisting of all probability measures on $\mathbb{R}^d$ with finite first moment. In $\mathcal{P}_1(\mathbb{R}^d)$ and for solutions of (4.1), a natural metric to work with is the so-called *Monge-Kantorovich-Rubinstein distance* [47],

$$W_1(\mu,\nu) := \sup\{|\langle \mu - \nu, \xi\rangle| = \left|\int_{\mathbb{R}^d} \xi(x) d(\mu-\nu)(x)\right|, \xi \in \mathrm{Lip}(\mathbb{R}^d), \mathrm{Lip}(\xi) \le 1\}. \tag{4.5}$$

We further denote $\mathcal{P}_c(\mathbb{R}^d)$ the space of compactly supported probability measures on $\mathbb{R}^d$. In particular, throughout the rest of this paper, we will assume that for any compactly supported measure valued weak solutions $\mu(t), \nu(t) \in C([0,T], \mathcal{P}_c(\mathbb{R}^d))$ of (4.1) we have the following stability inequality

$$W_1(\mu(t),\nu(t)) \le C(t) W_1(\mu(0),\nu(0)), \quad t \in [0,T], \tag{4.6}$$

where $C(t)$ is a positive increasing function of t with $C(0) > 0$, independent of the dimension d. We address the interested reader to [13, Section 4] for a sample of general conditions on the vector field $\mathcal{H}[\mathcal{F}](\mu)$ which guarantee stability (4.6) for solutions of equations (4.1).

4.3. Dimensionality reduction of kinetic equations. Provided a high-dimensional measure valued solution to the equation

$$\frac{\partial \mu}{\partial t} + \nabla \cdot (\mathcal{H}_{\mathcal{F}}[\mu]\mu) = 0, \quad \mu(0) = \mu_0 \in \mathcal{P}_c(\mathbb{R}^d)\,, \tag{4.7}$$

we will study the question whether its solution can be approximated by suitable projections in lower dimension.

Given a probability measure $\mu \in \mathcal{P}_1(\mathbb{R}^d)$, its projection into $\mathbb{R}^k$ by means of a matrix $M : \mathbb{R}^d \to \mathbb{R}^k$ is given by the *push-forward* measure $\mu_M := M\#\mu$,

$$\langle \mu_M, \varphi\rangle := \langle \mu, \varphi(M\cdot)\rangle \quad \text{for all } \varphi \in \mathrm{Lip}(\mathbb{R}^k). \tag{4.8}$$

Let us mention two explicit and relevant examples:

- If $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ is an atomic measure, we have $\langle \mu_M^N, \varphi\rangle = \langle \mu^N, \varphi(M\cdot)\rangle = \frac{1}{N}\sum_{i=1}^N \varphi(Mx_i)$. Therefore,

$$\mu_M^N = \frac{1}{N}\sum_{i=1}^N \delta_{Mx_i}\,. \tag{4.9}$$

- If μ is absolutely continuous with respect to the Lebesgue measure, i.e., it is a function in $L^1(\mathbb{R}^d)$, the calculation requires a bit more effort: Let us consider $M^\dagger$ the pseudo-inverse matrix of M. Recall that $M^\dagger = M^*(MM^*)^{-1}$ is a right inverse of M, and $M^\dagger M$ is the orthogonal projection onto the range of M^*. Moreover, $x = M^\dagger Mx + \xi_x$, where $\xi_x \in \ker M$ for all $x \in \mathbb{R}^d$. According to these observations, we write

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi(Mx)\mu(x)dx &= \int_{\mathbb{R}^d} \varphi(Mx)\mu(M^\dagger Mx + \xi_x)dx \\ &= \int_{\mathrm{ran}M^* \oplus \ker M} \varphi(Mx)\mu(M^\dagger Mx + \xi_x)dx \\ &= \int_{\mathrm{ran}M^*} \int_{\ker M} \varphi(Mv)\mu(M^\dagger Mv + v^\perp)dv^\perp dv\end{aligned}$$

Note now that $M_{|\mathrm{ran}M^*} : \mathrm{ran}M^* \to \mathrm{ran}M \approx \mathbb{R}^k$ is an isomorphism, hence $y = Mv$ implies the change of variables $dv = \det(M_{|\mathrm{ran}M^*})^{-1}dy = \det(MM^*)^{-1/2}dy$. Consequently, we have

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi(Mx)\mu(x)dx &= \int_{\mathbb{R}^d} \varphi(Mx)\mu(M^\dagger Mx + \xi_x)dx \\ &= \int_{\mathrm{ran}M^*} \int_{\ker M} \varphi(Mv)\mu(M^\dagger Mv + v^\perp)dv^\perp dv \\ &= \int_{\mathbb{R}^k} \left(\frac{1}{\det(MM^*)^{1/2}} \int_{\ker M} \mu(M^\dagger y + v^\perp)dv^\perp \right) \varphi(y)dy\,,\end{aligned}$$

and

$$\mu_M(y) = \frac{1}{\det(MM^*)^{1/2}} \int_{\ker M} \mu(M^\dagger y + v^\perp)dv^\perp.$$

According to the notion of push-forward, we can consider the measure valued function $\nu \in C([0,T], \mathcal{P}_c(\mathbb{R}^k))$, solution of the equation

$$\frac{\partial \nu}{\partial t} + \nabla \cdot (\mathcal{H}_{\mathcal{F}_M}[\nu]\nu) = 0, \quad \nu(0) = (\mu_0)_M \in \mathcal{P}_c(\mathbb{R}^k), \tag{4.10}$$

where $(\mu_0)_M = M\#\mu_0$ and $\mathcal{F}_M = (\{Mf_i, f_{ij}, i,j = 1,\dots,N\})_{N\in\mathbb{N}}$. As for the dynamical system (3.5), also equation (4.10) is fully defined on the lower-dimensional space $\mathbb{R}^k$ and depends on the original high-dimensional problem exclusively by means of the initial condition.

The natural question at this point is whether the solution ν of (4.10) provides information about the solution μ of (4.7). In particular, similarly to the result of Theorem 3.1, we will examine whether the approximation

$$\nu(t) \approx \mu_M(t), \quad t \in [0,T],$$

in Monge-Kantorovich-Rubinstein distance is preserved in finite time. We depict the expected result by the following diagram:

$$\begin{tikzcd}
\mu(0) \arrow[r, "t"] \arrow[d, "M"] & \mu(t) \arrow[d, "M"] \\
\nu(0) = (\mu_0)_M \arrow[r, "t"] & \nu(t) \approx \mu_M(t)
\end{tikzcd} .$$

This question will be addressed by approximation of the problem by atomic measures and by an application of Theorem 3.1 for the corresponding dynamical system, as concisely described by

$$\begin{tikzcd} \mu \arrow[r, "W_1(\mu,\mu^N)\lesssim\varepsilon"] \arrow[d, "M"] & \mu^N \arrow[d, "M"] \\ \nu \approx \mu_M \arrow[r, "W_1(\nu,\nu^N)\lesssim\varepsilon"] & \nu^N \approx \mu^N_M \end{tikzcd}$$

Let us now recall the framework and general assumptions for this analysis to be performed. We assume again that for all $N \in \mathbb{N}$ the family $\mathcal{F}_N = \{f_i, f_{ij} : i,j = 1,\dots N\}$ is composed of functions satisfying (1.2)-(1.4). Moreover, we assume that associated to $\mathcal{F} = (\mathcal{F}_N)_{N\in\mathbb{N}}$ and to

$$\dot{x}_i(t) = f_i(\mathcal{D}x(t)) + \sum_{j=1}^N f_{ij}(\mathcal{D}x(t))x_j(t), \tag{4.11}$$

we can define a mean-field equation

$$\frac{\partial \mu}{\partial t} + \nabla \cdot (\mathcal{H}[\mathcal{F}](\mu)\mu) = 0, \quad \mu(0) = \mu_0 \in \mathcal{P}_c(\mathbb{R}^d), \tag{4.12}$$

such that for any compactly supported measure valued weak solutions $\mu(t), \nu(t) \in C([0,T], \mathcal{P}_c(\mathbb{R}^d))$ of (4.1) we have the following stability

$$W_1(\mu(t),\nu(t)) \leq C(t) W_1(\mu(0),\nu(0)), \quad t \in [0,T], \tag{4.13}$$

where $C(t)$ is a positive increasing function of t, independent of the dimension d. We further require that corresponding assumptions, including stability, hold for the projected system (2.5) and kinetic equation (4.10). Then we have the following approximation result:

Theorem 4.1. *Let us assume that $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ and there exist points $\{x_1^0,\dots,x_N^0\} \subset \mathbb{R}^d$, for which the atomic measure $\mu_0^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i^0}$ approximates μ_0 up to $\varepsilon > 0$ in Monge-Kantorovich-Rubinstein distance, in the following sense*

$$W_1(\mu_0,\mu_0^N) \leq \varepsilon, \quad N = \mathcal{N}^{\overline{k}(\varepsilon)} \text{ for } \overline{k}(\varepsilon) \leq d \text{ and } \overline{k}(\varepsilon) \to d \text{ for } \varepsilon \to 0. \tag{4.14}$$

Requirement (4.14) *is in fact called the* delayed curse of dimension *as explained below in detail in Section 4.5. Depending on $\varepsilon > 0$ we fix also*

$$k = k(\varepsilon) = \mathcal{O}(\varepsilon^{-2}\log(N)) = \mathcal{O}(\varepsilon^{-2}\log(\mathcal{N})\overline{k}(\varepsilon)).$$

Moreover, let $M : \mathbb{R}^d \to \mathbb{R}^k$ be a linear mapping which is a continuous Johnson-Lindenstrauss embedding *as in* (3.4) *for continuous in time trajectories $x_i(t)$ of* (4.11) *with initial datum $x_i(0) = x_i^0$. Let $\nu \in C([0,T],\mathcal{P}_c(\mathbb{R}^k))$ be the weak solution of*

$$\frac{\partial \nu}{\partial t} + \nabla \cdot (\mathcal{H}[\mathcal{F}_M](\nu)\nu) = 0, \tag{4.15}$$

$$\nu(0) = (\mu_0)_M \in \mathcal{P}_c(\mathbb{R}^k), \tag{4.16}$$

where $(\mu_0)_M = M\#\mu_0$. Then

$$W_1(\mu_M(t),\nu(t)) \leq \mathcal{C}(t)\|M\|\varepsilon, \quad t \in [0,T], \tag{4.17}$$

where $\mathcal{C}(t)$ is an increasing function of t, with $\mathcal{C}(0) > 0$, which is at most polynomially growing with the dimension d.

Proof. Let us define $\nu^N(t)$ the solution to equation (4.15) with initial datum $\nu^N(0) = (\mu_0^N)_M$, or, equivalently, thanks to (4.9)

$$\nu^N(t) = \frac{1}{N}\sum_{i=1}^{n} \delta_{y_i(t)},$$

where $y_i(t)$ is the solution of

$$\begin{aligned} \dot{y}_i &= f_i(\mathcal{D}'y) + \sum_{j=1}^{N} f_{ij}(\mathcal{D}'y)y_j\,, \qquad i = 1,\dots,N\,,\\ y_i(0) &= Mx_i^0\,, \qquad i = 1,\dots,N\,. \end{aligned}$$

We estimate

$$W_1(\mu_M(t),\nu(t)) \le W_1(\mu_M(t),(\mu^N(t))_M) + W_1((\mu^N(t))_M,\nu^N(t)) + W_1(\nu^N(t),\nu(t)).$$

By using the definition of push-forward (4.8) and (4.14), the first term can be estimated by

$$\begin{aligned} W_1(\mu_M(t),(\mu^N(t))_M) &= \sup\{\langle \mu_M(t) - (\mu^N(t))_M, \varphi\rangle : \mathrm{Lip}(\varphi) \le 1\}\\ &= \sup\{\langle \mu(t) - \mu^N(t), \varphi(M\cdot)\rangle : \mathrm{Lip}(\varphi) \le 1\}\\ &\le \|M\| W_1(\mu(t),\mu^N(t)) \le \|M\| C(t)\varepsilon. \end{aligned}$$

We estimate now the second term

$$\begin{aligned} W_1((\mu^N(t))_M,\nu^N(t)) &= \sup\{\langle (\mu^N(t))_M - \nu^N(t), \varphi\rangle : \mathrm{Lip}(\varphi) \le 1\}\\ &= \sup\{\frac{1}{N}\sum_{i=1}^{N}(\varphi(Mx_i(t)) - \varphi(y_i(t))) : \mathrm{Lip}(\varphi) \le 1\}\\ &\le \frac{1}{N}\sum_{i=1}^{N} \|Mx_i(t) - y_i(t)\|_{\ell_2^k}. \end{aligned}$$

We recall the uniform approximation of Theorem 3.1,

$$\|Mx_i(t) - y_i(t)\|_{\ell_2^k} \le D(t)\varepsilon\,, \qquad i = 1,\dots,N,$$

where $D(t)$ is the time-dependent function on the right-hand-side of (3.7). Hence

$$W_1(\mu_M(t),(\mu^N(t))_M) \le D(t)\varepsilon.$$

We address now the upper estimate of the third term, by the assumed stability of the lower dimensional equation (4.10)

$$\begin{aligned} W_1(\nu^N(t),\nu(t)) &\le C(t)W_1(\nu^N(0),\nu(0))\\ &= C(t)W_1((\mu_0^N)_M,(\mu_0)_M)\\ &\le C(t)\|M\|W(\mu_0^N,\mu_0) \le C(t)\|M\|\varepsilon. \end{aligned}$$

We can fix $\mathcal{C}(t) = 2C(t)\|M\| + D(t)$, and, as observed in Theorem 3.3, we can assume without loss of generality that $\|M\| \le \sqrt{\frac{d}{k}}$. Hence, $\mathcal{C}(t)$ depends at most polynomially with respect to the dimension d. □

4.4. Approximation of probability measures by atomic measures and optimal integration. In view of the fundamental requirement (4.14) in Theorem 4.1, given $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$, we are interested to establish an upper bound to the best possible approximation in Monge-Kantorovich-Rubinstein distance by means of atomic measures $\mu_0^N = \frac{1}{N}\sum_{i=0}^{N-1} \delta_{x_i^0}$ with N atoms, i.e.,

$$\begin{aligned}\mathcal{E}_N(\mu_0) &:= \inf_{\mu_0^N = \frac{1}{N}\sum_{i=0}^{N-1}\delta_{x_i^0}} W_1(\mu_0, \mu_0^N) \qquad (4.18)\\ &= \inf_{\{x_0^0,\dots,x_{N-1}^0\}\subset\mathbb{R}^d} \sup\big\{\big|\int_{\mathbb{R}^d}\xi(x)d\mu_0(x) - \frac{1}{N}\sum_{i=0}^{N-1}\xi(x_i^0)\big| : \xi \in \mathrm{Lip}(\mathbb{R}^d), \mathrm{Lip}(\xi)\le 1\big\}.\end{aligned}$$

In fact, once we identify the optimal points $\{x_0^0,\dots,x_{N-1}^0\}$, we can use them as initial conditions $x_i(0) = x_i^0$ for the dynamical system (4.11), and by using the stability relationship (4.6), we obtain

$$W_1(\mu(t), \mu^N(t)) \le C(T) W_1(\mu_0, \mu_0^N), \quad t \in [0,T], \qquad (4.19)$$

where $\mu^N(t) = \frac{1}{N}\sum_{i=0}^{N-1}\delta_{x_i(t)}$, meaning that the solution of the partial differential equation (4.1) keeps optimally close to the particle solution of (4.11) also for successive time $t > 0$. Note that estimating (4.18) as a function of N is in fact a very classical problem in numerical analysis well-known as *optimal integration* with its high-dimensional behaviour being a relevant subject of the field of *Information Based Complexity* [40, 45].

The numerical integration of Lipschitz functions with respect to the Lebesgue measure and the study of its high-dimensional behaviour goes back to Bakhvalov [2], but much more is known nowadays. We refer to [29] and [32] for the state of the art of quantization of probability distributions.

The scope of this section is to recall some facets of these estimates and to reformulate them in terms of W_1 and $\mathcal{E}_N$. We emphasize that here and in what follows, we consider generic compactly supported probability measures μ, not necessarily absolutely continuous with respect to the Lebesgue measure. We start first by assuming $d = 1$, i.e., we work with a univariate measure $\mu \in \mathcal{P}_c(\mathbb{R})$ with support $\operatorname{supp}\mu \subset [a,b]$ and $\sigma := b - a > 0$. We define the points $x_0, \dots, x_{N-1}$ as the *quantiles* of the probability measure μ, i.e., $x_0 := a$ and

$$\frac{i}{N} = \int_{-\infty}^{x_i} d\mu(x), \quad i = 1, \dots, N-1. \qquad (4.20)$$

This is notationally complemented by putting $x_N := b$. Note that by definition $\int_{x_i}^{x_{i+1}} d\mu(x) = \frac{1}{N}, i = 0, \dots, N-1$, and we have

$$\begin{aligned}\left|\int_{\mathbb{R}}\xi(x)d\mu(x) - \frac{1}{N}\sum_{i=0}^{N-1}\xi(x_i)\right| &= \left|\sum_{i=0}^{N-1}\int_{x_i}^{x_{i+1}}(\xi(x)-\xi(x_i))d\mu(x)\right|\\ &\le \sum_{i=0}^{N-1}\int_{x_i}^{x_{i+1}}|\xi(x)-\xi(x_i)|\,d\mu(x) \qquad (4.21)\\ &\le \frac{\mathrm{Lip}(\xi)}{N}\sum_{i=0}^{N-1}(x_{i+1}-x_i) = \frac{\sigma\mathrm{Lip}(\xi)}{N}.\end{aligned}$$

Hence it is immediate to see that

$$\mathcal{E}_N(\mu) = \inf_{\mu^N=\frac{1}{N}\sum_{i=0}^{N-1}\delta_{x_i^0}} W_1(\mu,\mu^N) \le \frac{\sigma}{N}.$$

We would like to extend this estimate to higher dimension $d > 1$. However, for multivariate measures μ there is no such an easy upper bound, see [29] and [32] for very general statements, and for the sake of simplicity we restrict here the class of measures μ to certain special cases. As a typical situation, we address tensor product measures and sums of tensor products.

Lemma 4.2. *Let $\mu^1,\dots,\mu^d \in \mathcal{P}_1(\mathbb{R})$ with $W_1(\mu^j,\mu^{j,N_j}) \le \varepsilon_j$, $j = 1,\dots,d$ for some $N_1,\dots,N_d \in \mathbb{N}$, $\varepsilon_1,\dots,\varepsilon_d > 0$ and $\mu^{j,N_j} := \frac{1}{N_j}\sum_{i=0}^{N_j-1}\delta_{x_i^j}$. Let $N = \prod_{i=1}^d N_i$. Then*

$$W_1(\mu^1\otimes\dots\otimes\mu^d,\mu^N) \le \sum_{j=1}^d \varepsilon_j,$$

where

$$\mu^N := \frac{1}{N}\sum_{x\in X}\delta_x \quad \textit{and} \quad X := \prod_{j=1}^d\{x_0^j,\dots,x_{N_j-1}^j\}.$$

Proof. The proof is based on a simple argument using a telescopic sum. For $j = 1,\dots,d+1$ we put

$$V_j := \frac{1}{\prod_{i=j}^d N_i}\sum_{i_j=0}^{N_j-1}\cdots\sum_{i_d=0}^{N_d-1}\int_{\mathbb{R}^{j-1}}\xi(x_1,\dots,x_{j-1},x_{i_j}^j,\dots,x_{j_d}^d)d\mu^1(x_1)\dots d\mu^{j-1}(x_{j-1}).$$

Of course, if $j = 1$, then the integration over $\mathbb{R}^{j-1}$ is missing and if $j = d+1$ then the summation becomes empty. Now

$$\int_{\mathbb{R}^d}\xi(x)d\mu(x) - \frac{1}{\prod_{i=1}^d N_i}\sum_{i_1=0}^{N_1-1}\cdots\sum_{i_d=0}^{N_d-1}\xi(x_{i_1}^1,\dots,x_{i_d}^d) = \sum_{j=1}^d(V_{j+1}-V_j)$$

together with the estimate $|V_{j+1}-V_j| \le \varepsilon_j$ finishes the proof. □

Lemma 4.2 says, roughly speaking, that the tensor products of sampling points of univariate measures are good sampling points for the tensor product of the univariate measures. Next lemma deals with sums of measures.

Lemma 4.3. *Let $\mu_1,\dots,\mu_L \in \mathcal{P}_1(\mathbb{R}^d)$ with $W_1(\mu_l,\mu_l^N) \le \varepsilon_l$, $l = 1,\dots,L$ for some $N \in \mathbb{N}$, $\varepsilon_1,\dots,\varepsilon_L > 0$ and $\mu_l^N := \frac{1}{N}\sum_{i=0}^{N-1}\delta_{x_{l,i}}$. Then*

$$W_1\Big(\frac{\mu_1+\dots+\mu_L}{L},\mu^{LN}\Big) \le \frac{1}{L}\sum_{l=1}^L\varepsilon_l,$$

where

$$\mu^{LN} := \frac{1}{LN}\sum_{x\in X}\delta_x = \frac{1}{L}\sum_{l=1}^L\mu_l^N \quad \textit{and} \quad X := \bigcup_{l=1}^L\{x_{l,0},\dots,x_{l,N-1}\}.$$

Proof. We use the homogeneity of the Monge-Kantorovich-Rubinstein distance $W_1(a\mu, a\nu) = aW_1(\mu,\nu)$ for $\mu,\nu \in \mathcal{P}_1(\mathbb{R}^d)$ and $a \geq 0$ combined with its subadditivity $W_1(\mu_1+\mu_2,\nu_1+\nu_2) \leq W_1(\mu_1,\nu_1)+W_1(\mu_2,\nu_2)$ for $\mu_1,\mu_2,\nu_1,\nu_2 \in \mathcal{P}_1(\mathbb{R}^d)$. We obtain

$$W_1\Big(\frac{\mu_1+\cdots+\mu_L}{L},\frac{\mu_1^N+\cdots+\mu_L^N}{L}\Big) \leq \frac{1}{L}\sum_{l=1}^L W_1(\mu_l,\mu_l^N) \leq \frac{1}{L}\sum_{l=1}^L \varepsilon_l.$$

□

Next corollary follows directly from Lemma 4.2 and Lemma 4.3.

Corollary 4.4. *(i) Let $\mu^1,\dots,\mu^d \in \mathcal{P}_1(\mathbb{R})$ and $N_1,\dots,N_d \in \mathbb{N}$. Then*

$$\mathcal{E}_N(\mu^1\otimes\cdots\otimes\mu^d) \leq \sum_{j=1}^d \mathcal{E}_{N_j}(\mu^j), \quad \textit{where} \quad N := N_1\cdots N_d.$$

(ii) Let $\mu_1,\dots,\mu_L \in \mathcal{P}_1(\mathbb{R}^d)$ and $N \in \mathbb{N}$. Then

$$\mathcal{E}_{LN}\Big(\frac{\mu_1+\cdots+\mu_L}{L}\Big) \leq \frac{1}{L}\sum_{l=1}^L \mathcal{E}_N(\mu_l).$$

4.5. Delayed curse of dimension. Although Lemma 4.2, Lemma 4.3 and Corollary 4.4 give some estimates of the Monge-Kantorovich-Rubinstein distance between general and atomic measures, the number of atoms needed may still be too large to allow the assumption (4.14) in Theorem 4.1 to be fulfilled. Let us for example consider the case, where $\mu^1 = \cdots = \mu^d$ in Lemma 4.2 and $\varepsilon_1 = \cdots = \varepsilon_d =: \varepsilon$. Then, of course, $N_1 = \cdots = N_d =: \mathcal{N}$ and we observe, that the construction given in Lemma 4.2 gives an atomic measure, which approximates μ up to the error $d\varepsilon$ using $\mathcal{N}^d$ atoms, hence with an exponential dependence on the dimension d. This effect is another instance of the well-known phenomenon of the *curse of dimension*.

However, in many real-life high-dimensional applications the objects of study (in our case the measure $\mu \in \mathcal{P}_c(\mathbb{R}^d)$) concentrate along low-dimensional subspaces (or, more general, along low-dimensional manifolds) [5, 6, 17, 18, 19]. The number of atoms necessary to approximate these measures behaves in a much better way, allowing the application of (4.14) and Theorem 4.1. To clarify this effect, let us consider $\mu = \mu^1\otimes\cdots\otimes\mu^d$ with $\operatorname{supp}\mu^j \subset [a_j,b_j]$ and define $\sigma_j = b_j - a_j$. Let us assume, that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$ is a rapidly decreasing sequence. Furthermore, let $\varepsilon > 0$. Then we define $\overline{k} := \overline{k}(\varepsilon)$ to be the smallest natural number, such that

$$\sum_{k=\overline{k}(\varepsilon)+1}^d \sigma_k \leq \varepsilon/2$$

and put $N_k = 1$ for $k \in \{\overline{k}(\varepsilon)+1,\dots,d\}$. The numbers $N_1 = \cdots = N_{\overline{k}(\varepsilon)} = \mathcal{N}$ are chosen large enough so that

$$\frac{1}{\mathcal{N}}\sum_{k=1}^{\overline{k}(\varepsilon)} \sigma_k \leq \varepsilon/2.$$

Then Lemma 4.2 together with (4.20) state that there is an atomic measure μ^N with $N = \mathcal{N}^{\overline{k}(\varepsilon)}$ atoms, such that

$$W_1(\mu, \mu^N) \leq \sum_{k=1}^{d} \frac{\sigma_k}{N_k} \leq \varepsilon/2 + \varepsilon/2. \tag{4.22}$$

Hence, at the cost of assuming that the tensor product measure μ is concentrated along a $\overline{k}(\varepsilon)$-dimensional coordinate subspace, we can always approximate the measure μ with accuracy ε by using an atomic measure supported on points whose number depends exponentially on $\overline{k} = \overline{k}(\varepsilon) \ll d$. However, if we liked to have $\varepsilon \to 0$, then $\overline{k}(\varepsilon) \to d$ and again we are falling under the curse of dimension. This delayed kicking in of the need of a large number of points for obtaining high accuracy in the approximation (4.22) is in fact the so-called *delayed curse of dimension*, expressed by assumption (4.14), a concept introduced first by Curbera in [15], in the context of optimal integration with respect to Gaussian measures in high dimension.

Let us only remark, that the discussion above may be easily extended (with help of Lemma 4.3) to sums of tensor product measures. In that case we obtain as atoms the so-called *sparse grids*, cf. [10]. Using suitable change of variables, one could also consider measures concentrated around (smooth) low-dimensional manifolds, but this goes beyond the scope of this work, see [29] for a broader discussion.

Acknowledgments. We acknowledge the financial support provided by the START award “Sparse Approximation and Optimization in High Dimensions” no. FWF Y 432-N15 of the Fonds zur Förderung der wissenschaftlichen Forschung (Austrian Science Foundation). We would also like to thank Erich Novak for a discussion about optimal integration and for pointing us to some of the references given in Section 4.4.

REFERENCES

[1] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, J. Comput. Syst. Sci., 66 (2003), pp. 671–687.

[2] N. S. Bakhvalov, *On approximate computation of integrals*, Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem., 4 (1959), pp. 3–18.

[3] R. G. Baraniuk, M. Davenport, R. A. DeVore and M. Wakin, *A simple proof of the Restricted Isometry Property for random matrices*, Constr. Approx., 28 (2008), pp. 253–263.

[4] R. G. Baraniuk and M. B. Wakin, *Random projections of smooth manifolds*, Found. Comput. Math., 9 (2009), pp. 51–77.

[5] M. Belkin and P. Niyogi, *Laplacian eigenmaps and spectral techniques for embedding and clustering*, in: Advances in Neural Information Processing Systems 14 (NIPS 2001), MIT Press, Cambridge, 2001.

[6] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, *Neural Computation*, 6 (2003), pp. 1373–1396.

[7] P. Binev, A. Cohen, W. Dahmen, G. Petrova and P. Wojtaszczyk, *Convergence rates for greedy algorithms in reduced basis methods*, preprint, 2010.

[8] F. Bolley, J. A. Canizo and J. A. Carrillo, *Stochastic mean-field limit: non-Lipschitz forces and swarming*, Math. Models Methods Appl. Sci., to appear.

[9] A. Buffa, Y. Maday, A. T. Patera, C. Prudhomme and G. Turinici, *A priori convergence of the greedy algorithm for the parameterized reduced basis*, preprint.

[10] H. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.

[11] E. J. Candès, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l’Academie des Sciences, Paris, Serie I, 346 (2008), pp. 589–592.

[12] E. J. Candès, T. Tao and J. Romberg, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[13] J. A. Canizo, J. A. Carrillo and J. Rosado, *A well-posedness theory in measures for some kinetic models of collective motion*, Math. Models Methods Appl. Sci., to appear.

[14] J. A. Carrillo, M. Fornasier, G. Toscani and F. Vecil, *Particle, kinetic, hydrodynamic models of swarming*, in: Mathematical modeling of collective behavior in socio-economic and life-sciences, Birkhäuser, 2010.

[15] F. Curbera, *Delayed curse of dimension for Gaussian integration*, J. Complexity, 16 (2000), pp. 474–506.

[16] C. Cercignani, R. Illner and M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, Springer series in Applied Mathematical Sciences 106, Springer, 1994.

[17] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure denition of data: Diffusion maps, part I.*, Proc. of Nat. Acad. Sci., 102 (2005), pp. 7426–7431.

[18] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure denition of data: Diffusion maps, part II.*, Proc. of Nat. Acad. Sci., 102 (2005), pp. 7432–7438.

[19] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comp. Harm. Anal., 21 (2006), pp. 5–30.

[20] F. Cucker and S. Smale, *Emergent behavior in flocks*, IEEE Trans. Automat. Control, 52 (2007), pp 852–862.

[21] F. Cucker and S. Smale, *On the mathematics of emergence*, Japan J. Math., 2 (2007), pp. 197–227.

[22] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random. Struct. Algorithms, 22 (2003), pp. 60–65.

[23] T. Dijkema, C. Schwab and R. Stevenson, *An adaptive wavelet method for solving high-dimensional elliptic PDEs*, Constr. Approx., 30 (2009), pp. 423–455.

[24] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[25] S. Foucart, *A note on ensuring sparse recovery via ℓ_1-minimization*, Appl. Comput. Harmon. Anal., 29 (2010), pp. 97–103.

[26] M. Fornasier, *Domain decomposition methods for linear inverse problems with sparsity constraints*, Inverse Probl., 23 (2007), pp. 2505–2526.

[27] M. Fornasier, *Numerical methods for sparse recovery*, in: Theoretical Foundations and Numerical Methods for Sparse Recovery (ed. M. Fornasier), Volume 9 of Radon Series Comp. Appl. Math., deGruyter, pp. 93–200, 2010.

[28] M. Fornasier and H. Rauhut, *Compressive Sensing*, in: Handbook of Mathematical Methods in Imaging (ed. O. Scherzer), Springer, 2010.

[29] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Lecture Notes in Mathematics, 1730, Springer-Verlag, Berlin, 2000.

[30] M. Griebel and S. Knapek, *Optimized tensor-product approximation spaces*, Constr. Approx., 16 (2000), pp. 525–540.

[31] M. Griebel and P. Oswald, *Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems*, Adv. Comput. Math., 4 (1995), pp. 171–206.

[32] P. M. Gruber, *Optimum quantization and its applications*, Adv. Math. 186 (2004), pp. 456–497.

[33] S.-Y. Ha and E. Tadmor, *From particle to kinetic and hydrodynamic descriptions of flocking*, Kinetic and Related models, 1 (2008), pp. 315–335.

[34] M. Iwen and M. Maggioni, *Approximation of points on low-dimensional manifolds via compressive measurements*, in preparation.

[35] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contem. Math., 26 (1984), pp. 189–206.

[36] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol. 26 (1970), pp. 399–415.

[37] F. Krahmer and R. Ward, *New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property*, preprint, 2010.

[38] Y. Maday, A. T. Patera and G. Turinici, *Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations*, C. R. Acad. Sci., Paris, Ser. I, Math., 335 (2002), pp. 289–294.

[39] R. C. B. Nadler, S. Lafon and I. Kevrekidis, *Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 113–127.

[40] E. Novak and H. Woźniakowski, *Tractability of Multivariate Problems Volume II: Standard Information for Functionals*, Eur. Math. Society, EMS Tracts in Mathematics, Vol 12, 2010.

[41] M. R. D'Orsogna, Y. L. Chuang, A. L. Bertozzi and L. Chayes, *Self-propelled particles*

with soft-core interactions: patterns, stability, and collapse, Phys. Rev. Lett. 96 (2006).

[42] H. Rauhut, *Compressive sensing and structured random matrices*, in: Theoretical Foundations and Numerical Methods for Sparse Recovery (ed. M. Fornasier), Volume 9 of Radon Series Comp. Appl. Math., deGruyter, pp. 1–92, 2010.

[43] G. Rozza, D. B. P. Huynh and A. T. Patera, *Reduced basis approximation and a posteriori error estimation for afinely parametrized elliptic coercive partial diferential equations, application to transport and continuum mechanics*, Arch. Comput Method E, 15 (2008), pp. 229–275.

[44] S. Sen, *Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems*, Numer. Heat Tr. B-Fund, 54 (2008), pp. 369–389.

[45] J. F. Traub, G. W. Wasilkowski, H. Woźniakowski, *Information-based Complexity, Computer Science and Scientific Computing*, Academic Press, Inc., Boston, MA, 1988.

[46] K. Veroy, C. Prudhomme, D. V. Rovas and A. T. Patera, *A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations*, in: Proceedings of the 16th AIAA Computational Fluid Dynamics Conference, 2003.

[47] C. Villani, *Topics in Optimal transportation*, Graduate Studies in Mathematics, 58, American Mathematical Society, Providence, RI, 2003.

[48] M. B. Wakin, *Manifold-based signal recovery and parameter estimation from compressive measurements*, preprint, 2008.

Found Comput Math (2012) 12:229–262
DOI 10.1007/s10208-012-9115-y

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Learning Functions of Few Arbitrary Linear Parameters in High Dimensions

Massimo Fornasier · Karin Schnass · Jan Vybiral

Received: 17 August 2010 / Revised: 17 November 2011 / Accepted: 11 January 2012 /
Published online: 29 February 2012

Abstract Let us assume that f is a continuous function defined on the unit ball of $\mathbb{R}^d$, of the form $f(x) = g(Ax)$, where A is a $k \times d$ matrix and g is a function of k variables for $k \ll d$. We are given a budget $m \in \mathbb{N}$ of possible point evaluations $f(x_i)$, $i = 1, \dots, m$, of f, which we are allowed to query in order to construct a uniform approximating function. Under certain smoothness and variation assumptions on the function g, and an *arbitrary* choice of the matrix A, we present in this paper

1. a sampling choice of the points $\{x_i\}$ drawn at random for each function approximation;
2. algorithms (Algorithm 1 and Algorithm 2) for computing the approximating function, whose complexity is at most polynomial in the dimension d and in the number m of points.

Dedicated to Ronald A. DeVore on his 70th birthday.

Communicated by Emmanuel Candès.

M. Fornasier (✉)
Faculty of Mathematics, Technische Universität München, Boltzmannstraße 3, 85748 Garching,
Germany
e-mail: massimo.fornasier@ma.tum.de

K. Schnass · J. Vybiral
Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences,
Altenbergerstraße 69, 4040 Linz, Austria

K. Schnass
e-mail: karin.schnass@oeaw.ac.at

J. Vybiral
e-mail: jan.vybiral@oeaw.ac.at

Due to the arbitrariness of A, the sampling points will be chosen according to suitable random distributions, and our results hold with overwhelming probability. Our approach uses tools taken from the *compressed sensing* framework, recent Chernoff bounds for sums of positive semidefinite matrices, and classical stability bounds for invariant subspaces of singular value decompositions.

Keywords High-dimensional function approximation · Compressed sensing · Chernoff bounds for sums of positive semidefinite matrices · Stability bounds for invariant subspaces of singular value decompositions

Mathematics Subject Classification (2010) 65D15 · 03D32 · 68Q30 · 60B20 · 60G50

1 Introduction

1.1 Learning High-Dimensional Functions from Few Samples

In large-scale data analysis and learning, several real-life problems can be formulated as capturing or approximating a function defined on $\Omega \subset \mathbb{R}^d$ with dimension d very large, from relatively few given samples or queries. The usual assumption on the class of functions to be recovered is smoothness. The more regular a function is, the more accurately and the more efficiently it can be numerically approximated. However, in the field of *information-based complexity* this kind of problem is generally *intractable*; i.e., it does not have polynomial complexity. To clarify this poor approximation phenomenon, assume

$$\mathcal{F}_d := \left\{f : [0,1]^d \to \mathbb{R}, \left\|D^\alpha f\right\|_\infty \le 1, \alpha \in \mathbb{N}_0^d\right\}$$

to be the class of smooth functions we would like to approximate. We define the sampling operator $S_n = \phi \circ N$, where $N : \mathcal{F}_d \to \mathbb{R}^n$ is a suitable measurement operator and $\phi : \mathbb{R}^n \to L_\infty([0,1]^d)$ a recovery map. For example, N can take n samples $f(x_i)$, $i = 1, \dots, n$ of f, and ϕ can be a suitable interpolation operator. The approximation error provided by such a sampling operator is given by

$$e(S_n) := \sup_{f \in \mathcal{F}_d} \left\|f - S_n(f)\right\|_\infty.$$

With this notion we further define the approximation numbers

$$e(n,d) := \inf_{S_n} e(S_n),$$

indicating the performance of the best sampling method, and

$$n(\varepsilon, d) := \inf\left\{n : e(n,d) \le \varepsilon\right\}, \tag{1}$$

which is the minimal number of samples we need for the best sampling method to achieve a uniform accuracy $\varepsilon \in (0,1)$.

1.2 Intractability Results

Recent results by Novak and Woźniakowski [24] state that for a uniform approximation over $\mathcal{F}_d$ we have $e(n,d)=1$ for all $n \leq 2^{\lfloor d/2 \rfloor}-1$ or $n(\varepsilon,d) \geq 2^{\lfloor d/2 \rfloor}$ for all $\varepsilon \in (0,1)$. Hence, the number of samples to approximate even a C^∞-function grows exponentially with the dimension d. This result seems to obliterate any hope for an efficient solution of the learning problem in high dimension, and this phenomenon is sometimes referred to as the *curse of dimensionality*.

Nevertheless, very often the high-dimensional functions which we can expect as solutions to real-life problems exhibit more structure and eventually are much better behaved with respect to the approximation problem. Several models currently appear in the literature for which the approximation problem is *tractable*; i.e., the approximation error does not grow exponentially with respect to the dimension d.

According to the behavior of the *information complexity* $n(\varepsilon,d)$, cf. (1), for small $\varepsilon > 0$ and large $d \in \mathbb{N}$, one speaks of

- *polynomial tractability*: if $n(\varepsilon,d)$ depends polynomially on ε^{-1} and d
- *strong polynomial tractability*: if $n(\varepsilon,d)$ depends polynomially only on ε^{-1}
- *weak tractability*: if $\lim_{\varepsilon^{-1}+d\to\infty} \frac{\log n(\varepsilon,d)}{\varepsilon^{-1}+d}=0$

We point to [23, Chaps. 1 and 2] for further notions of tractability and many references.

In the next two subsections we will recount a few relevant approaches leading in some cases to (some sort of) tractability.

1.3 Functions of Few Variables

A function $f:[0,1]^d \to \mathbb{R}$ of d variables (d large) may be a sum of functions, which only depend on k variables (k small):

$$f(x_1,\dots,x_d)=\sum_{\ell=1}^{m} g_\ell(x_{i_1},\dots,x_{i_k}). \tag{2}$$

In optimization such functions are called *partially separable*. This model arises for instance in physics, when we consider problems involving interaction potentials, such as the Coulomb potential in electronic structure computations, or in social and economical models describing multiagent dynamics. Once k is fixed and $d \to \infty$, the learning problem of such functions is tractable, even if the g_ℓ are not very smooth. We specifically refer to the recent work of DeVore et al. [13], which describes an adaptive method for the recovery of high-dimensional functions in this class, for $m=1$.

This model can be extended to functions which are only approximatively depending on few variables, by considering the unit ball $\mathcal{H}_{d,\gamma}$ of the weighted Sobolev space of functions $f:[0,1]^d \to \mathbb{R}$ with

$$\|f\|_{d,\gamma}^2 := \sum_{u\subset[d]} \gamma_{d,u}^{-1} \int_{[0,1]^d} \left(\frac{\partial^{|u|}}{\partial x_u} f(x)\right)^2 \mathrm{d}x \leq 1, \tag{3}$$

where $[d] := \{1, \ldots, d\}$, and $\gamma := \{\gamma_{d,u}\}$ are non-negative weights; the definition $\frac{0}{0} := 0$ and the choice of $\gamma_{d,u} = 0$ leads us again to the model (2). A study of the tractability of this class, for various weights, can be found in [23].

1.4 Functions of One Linear Parameter in High Dimensions

One of the weaknesses of the model classes introduced above is that they are very coordinate biased. It would be desirable to have results for a class of basis changes which would make the model basis-independent. A general model assumes that

$$f(x) = g(Ax), \tag{4}$$

for A an arbitrary $k \times d$ matrix. While solutions to these unconstrained problems have so far been elusive, the special case of

$$f(x) = g(a \cdot x), \tag{5}$$

where a is a stochastic vector, i.e., $a = (a_1, \ldots, a_d)$, $a_j \geq 0$, $\sum_{j=1}^d a_j = 1$, and $g : [0, 1] \to \mathbb{R}$ is a $\mathcal{C}^s$ function for $s > 1$, has been fully addressed with an optimal recovery method in [11].

The aim of this work is to find an appropriate formulation of the general model (4), which generalizes both the model of k active coordinates as well as the model of one stochastic vector, and to analyze the tractability of the corresponding approximation problem. The rest of the paper is organized as follows. After introducing some basic notation, the next section is dedicated to the motivation and discussion of the generalized model. As an introduction to our formulation and solution approach, we then proceed to analyze the simple case of one active direction in Sect. 3, under milder assumptions on the vector $a = (a_1, \ldots, a_d)$, before finally addressing the fully generalized problem in Sect. 4. The last section is dedicated to the discussion of further extensions of our approach, to be addressed in successive papers.

1.5 Notation

In the following we will deal exclusively with real matrices, and we denote the space of $n \times m$ real matrices by $M_{n\times m}$. The entries of a matrix X are denoted by lower case letters and the corresponding indices, i.e., $X_{ij} = x_{ij}$. The transposed matrix $X^T \in M_{m\times n}$ of a matrix $X \in M_{n\times m}$ is the matrix with entries $x_{ij}^T = x_{ji}$. For $X \in M_{n\times m}$ we can write its (reduced) *singular value decomposition* [19] as

$$X = U \Sigma V^T$$

with $U \in M_{n\times p}$, $V \in M_{m\times p}$, $p \leq \min(n, m)$, matrices with orthonormal columns, and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in M_{p\times p}$ a diagonal matrix where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are the *singular values*. For specific matrices X we write the singular value decomposition

$$X = U(X)\Sigma(X)V(X)^T = U_X \Sigma_X V_X^T.$$

For symmetric, positive semidefinite matrices, i.e., $X = X^T$ and $v^T X v \geq 0$ for all vectors v, we can take $V = U$, and the singular value decomposition is equivalent to the eigenvalue decomposition. Note also that $\sigma_i(X) = \sqrt{\lambda_i(X^T X)}$, where $\lambda_i(X^T X)$ is the ith largest eigenvalue of the matrix $X^T X$ (actually, this holds for $n \geq m$, whereas we may want to consider XX^T instead of $X^T X$ if $m > n$). The rank of $X \in M_{n\times m}$ denoted by $\operatorname{rank}(X)$ is the number of nonzero singular values. We define the Frobenius norm of a matrix X as

$$\|X\|_F := \left(\sum_{ij} |x_{ij}|^2\right)^{1/2}.$$

It is also convenient to introduce the ℓ_p^n vector norms

$$\|x\|_{\ell_p^n} := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, \quad 0 < p < \infty.$$

We denote the identity matrix by $I_n \in M_{n\times n}$. The symbol $B_{\mathbb{R}^n}$ stands for the unit ball and $B_{\mathbb{R}^n}(r)$ for the ball of radius $r > 0$ in $\mathbb{R}^n$. The unit sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$. Finally, $\mathcal{L}^n$ indicates the Lebesgue measure in $\mathbb{R}^n$.

2 The General Model $f(x) = g(Ax)$ and Its Simplifications

The first approach one may be tempted to consider for a generalization of (5) is to ask that $f : [0,1]^d \to \mathbb{R}$ is of the form $f(x) = g(Ax)$, where A is a $k \times d$ stochastic matrix with orthonormal rows, i.e., $a_{ij} \geq 0$, $\sum_{j=1}^d a_{ij} = 1$ for all $i = 1, \dots, k$, $AA^T = I_k$, and $g : A([0,1]^d) \to \mathbb{R}$ is a $\mathcal{C}^s$ function for $s > 1$. However, there are two main problems with this formulation. The conditions of stochasticity and orthonormality of the rows of A together are very restrictive—the only matrices satisfying both of them are those having only one non-negative entry per column—and the domain of g cannot be chosen generically as $[0,1]^k$ but depends on A; i.e., it is the k-dimensional polytope $A([0,1]^d)$. Thus we will first return to the unconstrained model in (4) and give up the conditions of stochasticity and orthonormality. This introduces rotational invariance for the rows of A, and the quadrant defined by $[0,1]^d$ is no longer set apart as search space. Consequently and to avoid the complications arising with the polytope $A([0,1]^d)$, we will therefore focus on functions defined on the Euclidean ball.

To be precise, we consider functions $f : B_{\mathbb{R}^d}(1+\bar{\epsilon}) \to \mathbb{R}$ of the form (4), where A is an arbitrary $k \times d$ matrix whose rows are in ℓ_q^d, for some $0 < q \leq 1$,

$$\left(\sum_{j=1}^d |a_{ij}|^q\right)^{1/q} \leq C_1.$$

Further, we assume that the function g is defined on the image of $B_{\mathbb{R}^d}(1+\bar{\epsilon})$ under the matrix A and is twice continuously differentiable on this domain, i.e.,

$g \in C^2(AB_{\mathbb{R}^d}(1+\bar{\epsilon}))$, and

$$\max_{|\alpha|\leq 2}\big\| D^\alpha g\big\|_\infty \leq C_2.$$

For $\mu_{\mathbb{S}^{d-1}}$ the uniform surface measure on the sphere $\mathbb{S}^{d-1}$ we define the matrix

$$H^f := \int_{\mathbb{S}^{d-1}} \nabla f(x)\nabla f(x)^T \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x). \tag{6}$$

From the identity $\nabla f(x) = A^T \nabla g(Ax)$ we obtain

$$H^f = A^T \cdot \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x) \cdot A, \tag{7}$$

and therefore the rank of H^f is k or less. We will require H^f to be well conditioned; i.e., its singular values must satisfy $\sigma_1(H^f) \geq \cdots \geq \sigma_k(H^f) \geq \alpha > 0$.

The parameters in our model are the dimension d (large), the linear parameter dimension k (small), the non-negative constants $C_1, C_2, 0 < q \leq 1$, and $0 < \alpha \leq kC_2^2$.

We now show that this model can be simplified as follows. First we see that giving up the orthonormality condition on the rows of A was actually unnecessary. Let us consider the singular value decomposition of $A = U\Sigma V^T$, hence we rewrite

$$f(x) = g(Ax) = \tilde{g}(\tilde{A}x), \quad \tilde{A}\tilde{A}^T = I_k,$$

where $\tilde{g}(y) = g(U\Sigma y)$ and $\tilde{A} = V^T$. In particular, by simple direct computations,

- $\sup_{|\alpha|\leq 2}\|D^\alpha \tilde{g}\|_\infty \leq \sup_{|\alpha|\leq 2}\|D^\alpha g\|_\infty \cdot \max\{\sqrt{k}\sigma_1(A), k\sigma_1(A)^2\}$ and
- $(\sum_{j=1}^d |\tilde{a}_{ij}|^q)^{1/q} \leq C_1\sigma_k(A)^{-1}k^{1/q-1/2}$.

Hence, by possibly considering different constants $\tilde{C}_1 = k^{1/q-1/2}\sigma_k(A)^{-1}C_1$ and $\tilde{C}_2 = \max\{\sqrt{k}\sigma_1(A), k\sigma_1(A)^2\}C_2$, we can always assume that $AA^T = I_k$, meaning A is row-orthonormal. Note that for a row-orthonormal matrix A, (7) tells us that the singular values of H^f are the same as those of H_g, where

$$H_g := \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x).$$

The following simple result states that our model is almost well defined. As we will see later, the conditions on A and f will be sufficient for the unique identification of f by approximation up to any accuracy, but not necessarily for the unique identification of A and g.

Lemma 2.1 *Assume that $f(x) = g(Ax) = \tilde{g}(\tilde{A}x)$ with $A, \tilde{A}$ two $k \times d$ matrices such that $AA^T = I_k = \tilde{A}\tilde{A}^T$ and that H^f has rank k. Then $\tilde{A} = \mathcal{O}A$ for some $k \times k$ orthonormal matrix $\mathcal{O}$.*

Proof Because A and $\tilde{A}$ are row-orthonormal the singular values of H_g and $H_{\tilde{g}}$ are the same as those of H^f; i.e., we have $H_g = U\Sigma U^T$ and $H_{\tilde{g}} = \tilde{U}\Sigma\tilde{U}^T$, where Σ is a $k \times k$ diagonal matrix containing the singular values of H^f in nonincreasing order and $U, \tilde{U}$ are orthonormal $k \times k$ matrices. Inserting this into (7) we get

$$H^f = A^T H_g A = A^T U \Sigma U^T A = \tilde{A}^T H_{\tilde{g}} \tilde{A} = \tilde{A}^T \tilde{U} \Sigma \tilde{U}^T \tilde{A}.$$

$U^T A$ and $\tilde{U}^T \tilde{A}$ are both row-orthonormal, so we have two singular value decompositions of H^f. Because the singular vectors are unique up to an orthonormal transform, we have $\tilde{U}^T \tilde{A} = V U^T A$ for some orthonormal matrix V or $\tilde{A} = \mathcal{O} A$ for $\mathcal{O} = \tilde{U} V U^T$, which is by construction orthonormal. □

With these observations in mind, let us now restate the problem and summarize our requirements. We restrict the learning problem to functions $f : B_{R^d}(1+\bar{\epsilon}) \to \mathbb{R}$ of the form $f(x) = g(Ax)$, where $A \in M_{k\times d}$ and $AA^T = I_k$. As we are interested in recovering f from a small number of samples, the accuracy will depend on the smoothness of g. In order to get simple convergence estimates, we require $g \in C^2(B_{R^k}(1+\bar{\epsilon}))$. These choices determine two positive constants C_1, C_2 for which

$$\left(\sum_{j=1}^{d} |a_{ij}|^q\right)^{1/q} \le C_1, \tag{8}$$

and

$$\sup_{|\alpha|\le 2} \left\| D^\alpha g \right\|_\infty \le C_2. \tag{9}$$

For the problem to be well conditioned we require the matrix H^f to be positive definite,

$$\sigma_1\bigl(H^f\bigr) \ge \cdots \ge \sigma_k\bigl(H^f\bigr) \ge \alpha, \tag{10}$$

for a fixed constant $\alpha > 0$ (actually later we may simply choose $\alpha = \sigma_k(H^f)$).

Remark 1 Let us briefly comment on condition (10) in the most simple case $k = 1$, by showing that such a condition is actually necessary in order to formulate a *tractable* algorithm for the uniform approximation of f from point evaluations.

The optimal choice of α is given by

$$\alpha = \int_{\mathbb{S}^{d-1}} \left|g'(a\cdot x)\right|^2 \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x) = \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} \bigl(1-|y|^2\bigr)^{\frac{d-3}{2}} \mathrm{d}y, \tag{11}$$

cf. Theorem 3.7. Furthermore, we consider the function $g \in C^2([-1-\bar{\epsilon}, 1+\bar{\epsilon}])$ given by $g(y) = 8(y-1/2)^3$ for $y \in [1/2, 1+\bar{\epsilon}]$ and zero otherwise. Notice that, for every $a \in \mathbb{R}^d$ with $\|a\|_{\ell_2^d} = 1$, the function $f(x) = g(a\cdot x)$ vanishes everywhere on $\mathbb{S}^{d-1}$ outside of the cap $\mathcal{U}(a, 1/2) := \{x \in \mathbb{S}^{d-1} : a\cdot x \ge 1/2\}$ (see Fig. 1). The $\mu_{\mathbb{S}^{d-1}}$ measure of $\mathcal{U}(a, 1/2)$ obviously does not depend on a and is known to be exponentially small in d [21]; see also Sect. 3.3. Furthermore, it is known that there is a constant $c > 0$ and unit vectors $a^1, \dots, a^K$, such that the sets $\mathcal{U}(a^1, 1/2), \dots, \mathcal{U}(a^K, 1/2)$ are

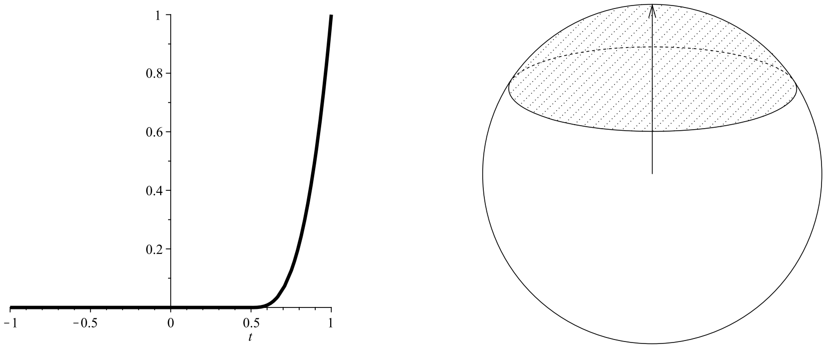


Fig. 1 The function g and the spherical cap $\mathcal{U}(a, 1/2)$

mutually disjoint and $K \geq e^{cd}$. Finally, we observe that $\max_{x\in\mathbb{S}^{d-1}} |f(x)| = f(a) = g(1) = 1$.

We conclude that *any* algorithm making use only of the structure of $f(x) = g(a \cdot x)$ and the condition (9) must use exponentially many sampling points in order to distinguish between $f(x) \equiv 0$ and $f(x) = g(a^i \cdot x)$ for some of the a^i's as constructed above. Hence, some additional conditions like (8) and (10) are actually necessary to avoid the curse of dimensionality and to achieve at least some sort of tractability. Let us observe that $\alpha = \alpha(d)$ decays exponentially with d for the function g considered above. We shall further discuss the role of α in Sect. 3.3.

Contrary to the approach in [11], our strategy to learn functions of the type (4) is to first find an approximation $\hat{A}$ to A. Once this is known, we will give a pointwise definition of the function $\hat{g}$ on $B_{R^k}(1)$ such that $\hat{f}(x) := \hat{g}(\hat{A}x)$ is a good approximation to f on $B_{R^d}(1)$. This will be done in such a way that the evaluation of $\hat{g}$ at one point will require only one function evaluation of f. Consequently, an approximation of $\hat{g}$ on its domain $B_{R^k}(1)$ using standard techniques, like sampling on a regular grid and spline-type approximations, will require a number of function evaluations of f depending only on the desired accuracy and k, but not on d. Thus we will restrict our analysis to the problem of finding $\hat{A}$, defining $\hat{g}$, and the amount of queries necessary to do that.

3 The One-Dimensional Case $k = 1$

For an easy introduction, we start by addressing our recovery method again in the simplest case of a *ridge function*,

$$f(x) = g(a \cdot x), \tag{12}$$

where $a = (a_1, \dots, a_d) \in \mathbb{R}^d$ is a row vector, $\|a\|_{\ell_2^d} = 1$, and g is a function from the image of $B_{\mathbb{R}^d}(1 + \bar{\epsilon})$ under a to $\mathbb{R}$, i.e., $g : B_R(1 + \bar{\epsilon}) \to \mathbb{R}$.

The ridge function terminology was introduced in the 1970s by Logan and Shepp [22] in connection with the mathematics of computer tomography, but these functions have been considered for some time under the name of *plane waves*. See, for example, [12, 20]. Ridge functions and ridge function approximation are studied in statistics, often under the name of projection pursuit. Projection pursuit algorithms approximate a function of d variables by functions of the form

$$f(x) \approx \sum_{j=1}^{\ell} g_j(a_j \cdot x). \tag{13}$$

Hence the recovery of f in (12) from few samples can be seen as an instance of the projection pursuit problem. For a survey on some approximation-theoretic questions concerning ridge functions and their connections to neural networks, see [27] and references therein, and the work of Candès and Donoho on ridgelet approximation [5–7].

For further clarity of notation, in the following we will assume a to be a row vector, i.e., a $1 \times d$ matrix, while other vectors, $x, \xi, \varphi, \ldots$, are always assumed to be column vectors. Hence the symbol $a \cdot x$ stands for the product of the $1 \times d$ matrix a with the $d \times 1$ vector x.

3.1 The Algorithm

As in [11] a basic ingredient of the algorithm is a version of Taylor's theorem giving access to the vector a. For $\xi \in B_{R^d}$, $\varphi \in B_{\mathbb{R}^d}(r)$, $\epsilon, r \in \mathbb{R}_+$, with $r\epsilon \le \bar{\epsilon}$, we have, by Taylor's expansion, the identity

$$\bigl[g'(a \cdot \xi)a\bigr] \cdot \varphi = \frac{\partial f}{\partial \varphi}(\xi) = \frac{f(\xi + \epsilon\varphi) - f(\xi)}{\epsilon} - \frac{\epsilon}{2}\bigl[\varphi^T \nabla^2 f(\zeta)\varphi\bigr], \tag{14}$$

for a suitable $\zeta(\xi, \varphi) \in B_{\mathbb{R}^d}(1 + \bar{\epsilon})$. From our assumptions (8) and (9), the term $[\varphi^T \nabla^2 f(\zeta)\varphi]$ is uniformly bounded as soon as φ is bounded. We will consider the above equality for several directions φ_i and at several sampling points ξ_j.

To be more precise, we define two sets $\mathcal{X}$, Φ of points. The first,

$$\mathcal{X} = \bigl\{\xi_j \in \mathbb{S}^{d-1} : j = 1, \ldots, m_{\mathcal{X}}\bigr\}, \tag{15}$$

contains the $m_{\mathcal{X}}$ sampling points and is drawn at random in $\mathbb{S}^{d-1}$ according to the probability measure $\mu_{\mathbb{S}^{d-1}}$. For the second, containing the m_Φ derivative directions, we have

$$\Phi = \left\{\varphi_i \in B_{\mathbb{R}^d}(\sqrt{d}/\sqrt{m_\Phi}) : \varphi_{i\ell} = \frac{1}{\sqrt{m_\Phi}} \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2, \end{cases} \right.$$
$$\left. i = 1, \ldots, m_\Phi, \text{ and } \ell = 1, \ldots, d \right\}. \tag{16}$$

Actually, we identify Φ with the $m_\Phi \times d$ matrix whose rows are the vectors φ_i. To write the $m_{\mathcal{X}} \times m_\Phi$ instances of (14) in a concise way, we collect the directional derivatives $g'(a\cdot\xi_j)a$, $j=1,\dots,m_{\mathcal{X}}$ as columns in the $d\times m_{\mathcal{X}}$ matrix X, i.e.,

$$X=\big(g'(a\cdot\xi_1)a^T,\dots,g'(a\cdot\xi_{m_{\mathcal{X}}})a^T\big), \tag{17}$$

and we define the $m_{\Phi}\times m_{\mathcal{X}}$ matrices Y and $\mathcal{E}$ entrywise by

$$y_{ij}=\frac{f(\xi_j+\epsilon\varphi_i)-f(\xi_j)}{\epsilon}, \tag{18}$$

and

$$\varepsilon_{ij}=\frac{\epsilon}{2}\big[\varphi_i^T\nabla^2 f(\zeta_{ij})\varphi_i\big]. \tag{19}$$

We denote by y_j the columns of Y and by ε_j the columns of $\mathcal{E}$, $j=1,\dots,m_{\mathcal{X}}$. With these matrices we can write the following factorization:

$$\Phi X=Y-\mathcal{E}. \tag{20}$$

The algorithm we propose to approximate the vector a is now based on the fact that the matrix X has a very special structure, i.e., $X=a^T\mathcal{G}^T$, where $\mathcal{G}=(g'(a\cdot\xi_1),\dots,g'(a\cdot\xi_{m_{\mathcal{X}}}))^T$. In other words, every column x_j is a scaled copy of the vector a^T and *compressible* if a is *compressible*. We define a vector a as compressible informally by saying that it can be well approximated in ℓ_p-norm by a sparse vector. Actually, any vector a with small ℓ_q-norm can be approximated in ℓ_p by its best K-term approximation $a_{[K]}$ according to the following well-known estimate:

$$\sigma_K(x)_{\ell_p^d}:=\|a-a_{[K]}\|_{\ell_p^d}\le\|a\|_{\ell_q^d}K^{1/p-1/q},\quad p\ge q. \tag{21}$$

Thus by changing the viewpoint to get

$$Y=\Phi X+\mathcal{E},$$

we see that due to the random construction of Φ we actually have a *compressed sensing* problem, and known theory tells us that we can recover a stable approximation $\hat{x}_j$ to x_j via ℓ_1-minimization (see Theorem 3.2 for the precise statement). To get an approximation of a we then simply have to set $\hat{a}=\hat{x}_j/\|\hat{x}_j\|_{\ell_2^d}$ for j such that $\|\hat{x}_j\|_{\ell_2^d}$ is maximal. From these informal ideas we derive the following algorithm.

Algorithm 1

- Given $m_{\Phi},m_{\mathcal{X}}$, draw at random the sets Φ and $\mathcal{X}$ as in (15) and (16), and construct Y according to (18).
- Set $\hat{x}_j=\Delta(y_j):=\arg\min_{y_j=\Phi z}\|z\|_{\ell_1^d}$.
- Find

$$j_0=\arg\max_{j=1,\dots,m_{\mathcal{X}}}\|\hat{x}_j\|_{\ell_2^d}. \tag{22}$$

- Set $\hat{a}=\hat{x}_{j_0}/\|\hat{x}_{j_0}\|_{\ell_2^d}$.
- Define $\hat{g}(y):=f(\hat{a}^Ty)$ and $\hat{f}(x):=\hat{g}(\hat{a}\cdot x)$.

The quality of the final approximation clearly depends on the error between $\hat{x}_j$ and x_j, which can be controlled through the number of *compressed sensing measurements* m_Φ, and the size of $\hat{a} \approx \max_j \|x_j\|_{\ell_2^d} = \max_j |g'(a \cdot \xi_j)|$, which is related to the number of random samples $m_{\mathcal{X}}$. If (11) is satisfied with α large, we shall show in Lemma 3.6 with the help of Hoeffding's inequality that $\max_j \|x_j\|_{\ell_2^d} = \max_j |g'(a \cdot \xi_j)|$ is also large with high probability. If the value of α is unknown and small, the values of $\|\hat{x}_j\|_{\ell_2^d}$ produced by Algorithm 1 could be small as well and, as discussed after the formula (11), no reliable and tractable approximation procedure is possible.

To be exact, in the next section we will prove the following approximation result.

Theorem 3.1 *Let* $0 < s < 1$ *and* $\log d \le m_\Phi \le [\log 6]^{-2} d$. *Then there is a constant* c_1' *such that using* $m_{\mathcal{X}} \cdot (m_\Phi + 1)$ *function evaluations of* f, *Algorithm* 1 *defines a function* $\hat{f} : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \to \mathbb{R}$ *that, with probability*

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + 2e^{-\frac{2 m_{\mathcal{X}} s^2 \alpha^2}{C_2^4}}\right), \tag{23}$$

will satisfy

$$\|f - \hat{f}\|_\infty \le 2C_2(1 + \bar{\epsilon}) \frac{\nu_1}{\sqrt{\alpha(1-s)} - \nu_1}, \tag{24}$$

where

$$\nu_1 = C' \left(\left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon}{\sqrt{m_\Phi}} \right) \tag{25}$$

and C' *depends only on* C_1 *and* C_2 *from* (8) *and* (9).

Remark 2 1. We shall fix ν_1 as defined by (25) for the rest of this section. Furthermore, we suppose that the selected parameters (s, ϵ, and m_Φ) are such that $\nu_1 < \sqrt{\alpha(1-s)}$ holds. Refer to Remark 4(ii) to see how we can circumvent in practice the case that this condition may not hold, clearly invalidating the approximation (24).

2. In order to show a concrete application of the previous result, let us consider, for simplicity, a class of uniformly smooth functions g such that $|g'(0)| \neq 0$; hence, by Proposition 3.8, $\alpha = \alpha(g) > 0$ is independent of the dimension d. If additionally we choose $q = 1$, $m_\Phi < d$, and $\epsilon > 0$ such that $m_\Phi(\epsilon + \sqrt{\log(d/m_\Phi)})^{-2} = \mathcal{O}(\delta^{-2}\alpha^{-1})$, $\delta > 0$, for $\delta, \alpha \to 0$ and $m_{\mathcal{X}} = \mathcal{O}(\alpha^{-2})$ for $\alpha \to 0$, then, according to Theorem 3.1, we obtain the uniform error estimate

$$\|f - \hat{f}\|_\infty = \mathcal{O}(\delta), \quad \delta \to 0,$$

with high probability. Notice that, if $1/\log(d) > \delta > 0$, then the number of evaluation points $m_{\mathcal{X}} \cdot (m_\Phi + 1) = \mathcal{O}((\delta \cdot \alpha)^{-3})$, for $\delta, \alpha \to 0$, is actually independent of the dimension d.

3.2 The Analysis

We will first show that $\hat{x}_j$ is a good approximation to x_j for all j. This follows by the results from the framework of *compressed sensing* [3, 8, 10, 14, 16–18]. In particular, we state the following useful result, which is a specialization of Theorem 1.2 from [36] to the case of Bernoulli matrices.

Theorem 3.2 *Assume that Φ is an $m \times d$ random matrix with all entries being independent Bernoulli variables scaled with $1/\sqrt{m}$, see, e.g.*, (16).

(i) *Let $0 < \delta < 1$. Then there are two positive constants $c_1, c_2 > 0$, such that the matrix Φ has the restricted isometry property*

$$(1-\delta)\|x\|^2_{\ell_2^d} \le \|\Phi x\|^2_{\ell_2^m} \le (1+\delta)\|x\|^2_{\ell_2^d} \tag{26}$$

for all $x \in \mathbb{R}^d$ such that $\#\operatorname{supp}(x) \le c_2 m/\log(d/m)$ with probability at least

$$1 - e^{-c_1 m}. \tag{27}$$

(ii) *Let us suppose that $d > [\log 6]^2 m$. Then there are positive constants $C, c_1', c_2' > 0$, such that, with probability at least*

$$1 - e^{-c_1' m} - e^{-\sqrt{md}}, \tag{28}$$

the matrix Φ has the following property. For every $x \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}^m$ and every natural number $K \le c_2' m/\log(d/m)$ we have

$$\big\|\Delta(\Phi x + \varepsilon) - x\big\|_{\ell_2^d} \le C\big(K^{-1/2}\sigma_K(x)_{\ell_1^d} + \max\big\{\|\varepsilon\|_{\ell_2^m}, \sqrt{\log d}\,\|\varepsilon\|_{\ell_\infty^m}\big\}\big), \tag{29}$$

where

$$\sigma_K(x)_{\ell_1^d} := \inf\big\{\|x - z\|_{\ell_1^d} : \#\operatorname{supp} z \le K\big\}$$

is the best K-term approximation of x.

Remark 3 (i) The first part of Theorem 3.2 is well known; see, e.g., [3] or [16, p. 15] and references therein.

(ii) The second part of Theorem 3.2 is relatively new. It follows from Theorem 2.3 of [36] combined with Theorem 3.5 of [13], and the first part of Theorem 3.2. Without the explicit bound of the probability (28), it also appears as Theorem 1.2 in [36].

Applied to the situation at hand, we immediately derive the following corollary.

Corollary 3.3 (i) *Let $d > [\log 6]^2 m_\Phi$. Then with probability at least*

$$1 - \big(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}}\big),$$

all the vectors $\hat{x}_j = \Delta(y_j)$, $j = 1, \dots, m_{\mathcal{X}}$ *calculated in Algorithm* 1 *satisfy*

$$\|x_j - \hat{x}_j\|_{\ell_2^d} \le C\left(\left[\frac{m_\Phi}{\log(d/m_\Phi)}\right]^{1/2-1/q} + \max\left\{\|\varepsilon_j\|_{\ell_2^{m_\Phi}}, \sqrt{\log d}\|\varepsilon_j\|_{\ell_\infty^{m_\Phi}}\right\}\right) \quad (30)$$

where C *depends only on* C_1 *and* C_2 *from* (8) *and* (9).

(ii) *If furthermore* $m_\Phi \ge \log d$ *holds*, *then with the same probability also*

$$\|x_j - \hat{x}_j\|_{\ell_2^d} \le C'\left(\left[\frac{m_\Phi}{\log(d/m_\Phi)}\right]^{1/2-1/q} + \frac{\epsilon}{\sqrt{m_\Phi}}\right), \quad (31)$$

where C' *depends again only on* C_1 *and* C_2 *from* (8) *and* (9).

Proof We apply Theorem 3.2 to the equation $y_j = \Phi x_j + \varepsilon_j$ and $K \le c_2' m_\Phi / \log(d/m_\Phi)$. To do so, we must estimate the best K-term approximation error of $\sigma_K(x_j)_{\ell_1^d}$ and the size of the errors ε_j. We start by bounding $\sigma_K(x_j)_{\ell_1^d}$. Recall that due to the construction of X every column is a scaled copy of the vector a^T, i.e., $x_j = g'(a \cdot \xi_j)a^T$, so we have by (21)

$$K^{-1/2}\sigma_K(x_j)_{\ell_1^d} \le \left|g'(a \cdot \xi_j)\right| \cdot \|a\|_{\ell_q^d} \cdot K^{1/2-1/q} \le C_1 C_2 \left[\frac{m_\Phi}{\log(d/m_\Phi)}\right]^{1/2-1/q}. \quad (32)$$

This finishes the proof of the first part.

To prove the second part, we estimate the size of the errors using (19),

$$\begin{aligned}\|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} &= \frac{\epsilon}{2} \cdot \max_{i=1,\dots,m_\Phi} \left|\varphi_i^T \nabla^2 f(\zeta_{ij})\varphi_i\right| \\ &= \frac{\epsilon}{2m_\Phi} \cdot \max_{i=1,\dots,m_\Phi} \left|\sum_{k,l=1}^d a_k a_l g''(a \cdot \zeta_{ij})\right| \\ &\le \frac{\epsilon\|g''\|_\infty}{2m_\Phi}\left(\sum_{k=1}^d |a_k|\right)^2 \le \frac{\epsilon\|g''\|_\infty}{2m_\Phi}\left(\sum_{k=1}^d |a_k|^q\right)^{2/q} \le \frac{C_1^2 C_2}{2m_\Phi}\epsilon, \quad (33)\end{aligned}$$

$$\|\varepsilon_j\|_{\ell_2^{m_\Phi}} \le \sqrt{m_\Phi}\|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} \le \frac{C_1^2 C_2}{2\sqrt{m_\Phi}}\epsilon, \quad (34)$$

leading to

$$\max\left\{\|\varepsilon_j\|_{\ell_2^{m_\Phi}}, \sqrt{\log d}\|\varepsilon_j\|_{\ell_\infty^{m_\Phi}}\right\} \le \frac{C_1^2 C_2}{2\sqrt{m_\Phi}}\epsilon \cdot \max\left\{1, \sqrt{\frac{\log d}{m_\Phi}}\right\}.$$

Together with our assumption $m_\Phi \ge \log d$, this finishes the proof. □

Next we need a technical lemma to relate the error between the normalized version of $\hat{x}_j$ and a to the size of $\|\hat{x}_j\|_{\ell_2^d}$.

Lemma 3.4 (Stability of subspaces: one-dimensional case) *Let us fix* $\hat{x} \in \mathbb{R}^d$, $a \in \mathbb{S}^{d-1}$, $0 \neq \gamma \in \mathbb{R}$, *and* $n \in \mathbb{R}^d$ *with norm* $\|n\|_{\ell_2^d} \leq \nu_1 < |\gamma|$. *If we assume* $\hat{x} = \gamma a + n$, *then*

$$\left\| \operatorname{sign}\gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} \leq \frac{2\nu_1}{\|\hat{x}\|_{\ell_2^d}}. \tag{35}$$

Proof Applying the triangular inequality and its reverse form several times and using $a \in \mathbb{S}^{d-1}$, we get

$$\begin{aligned} \left\| \operatorname{sign}\gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} &\leq \left\| \operatorname{sign}\gamma \frac{\hat{x}}{\|\hat{x}\|_{\ell_2^d}} - \frac{|\gamma|a}{\|\hat{x}\|_{\ell_2^d}} \right\|_{\ell_2^d} + \left\| \frac{|\gamma|a}{\|\hat{x}\|_{\ell_2^d}} - a \right\|_{\ell_2^d} \\ &\leq \frac{\nu_1}{\|\hat{x}\|_{\ell_2^d}} + \left| \frac{|\gamma|}{\|\hat{x}\|_{\ell_2^d}} - 1 \right| \leq \frac{2\nu_1}{\|\hat{x}\|_{\ell_2^d}}. \end{aligned}$$ □

Applied to our situation where $\hat{x}_j = g'(a \cdot \xi_j)a^T + n_j$, we see that the bound in (35) is best for $\|\hat{x}_j\|_{\ell_2^d}$ maximal, which justifies our definition of $\hat{a}$ in Algorithm 1.

As a last ingredient for the proof of Theorem 3.1 we need a lower bound for $\max_{j=1,\dots,m_\mathcal{X}} \|\hat{x}\|_{\ell_2^d}$. Since we have $\max_j \|\hat{x}_j\|_{\ell_2^d} \geq \max_j |g'(a \cdot \xi_j)| - \max_j \|\hat{x}_j - x_j\|_{\ell_2^d} \geq \max_j |g'(a \cdot \xi_j)| - \nu_1$ we just have to show that, with high probability, our random sampling of the gradient via the ξ_j provided a good maximum. To do this we will use Hoeffding's inequality, which we recall below for the reader's convenience.

Proposition 3.5 (Hoeffding's inequality) *Let* $X_1, \dots, X_m$ *be independent random variables. Assume that the* X_j *are almost surely bounded, i.e., there exist finite scalars* a_j, b_j *such that*

$$\mathbb{P}\{X_j - \mathbb{E}X_j \in [a_j, b_j]\} = 1,$$

for $j = 1, \dots, m$. *Then we have*

$$\mathbb{P}\left\{ \left| \sum_{j=1}^m X_j - \mathbb{E}\left(\sum_{j=1}^m X_j \right) \right| \geq t \right\} \leq 2e^{-\frac{2t^2}{\sum_{j=1}^m (b_j - a_j)^2}}.$$

Let us now apply Hoeffding's inequality to the random variables $X_j = |g'(a \cdot \xi_j)|^2$.

Lemma 3.6 *Let us fix* $0 < s < 1$. *Then with probability* $1 - 2e^{-\frac{2m_\mathcal{X} s^2 \alpha^2}{C_2^4}}$ *we have*

$$\max_{j=1,\dots,m_\mathcal{X}} |g'(a \cdot \xi_j)| \geq \sqrt{\alpha(1-s)},$$

where $\alpha := \mathbb{E}_\xi(|g'(a \cdot \xi_j)|^2)$.

Proof By our assumptions (10) and (9) we have

$$\mathbb{E}X_j = \mathbb{E}_\xi(|g'(a \cdot \xi_j)|^2) = \int_{\mathbb{S}^{d-1}} |g'(a \cdot \xi)|^2 \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(\xi) \geq \alpha > 0,$$

and

$$X_j - \mathbb{E}X_j \in \left[-\alpha, C_2^2 - \alpha\right].$$

Hence, by Hoeffding's inequality we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^{m_{\mathcal{X}}} \left|g'(a \cdot \xi_j)\right|^2 - m_{\mathcal{X}}\alpha\right| \geq s m_{\mathcal{X}}\alpha\right\} \leq 2e^{-\frac{2m_{\mathcal{X}}s^2\alpha^2}{C_2^4}}. \tag{36}$$

Using (36) we immediately obtain

$$\frac{1}{m_{\mathcal{X}}} \sum_{j=1}^{m_{\mathcal{X}}} \left|g'(a \cdot \xi_j)\right|^2 \geq \alpha(1-s), \tag{37}$$

with probability $1 - 2e^{-\frac{2m_{\mathcal{X}}s^2\alpha^2}{C_2^4}}$. If $|g'(a \cdot \xi_j)|^2 < \alpha(1-s)$ for all $j = 1, \dots, m_{\mathcal{X}}$, then (37) would be violated. Hence for the maximum we have

$$\max_{j=1,\dots,m_{\mathcal{X}}} \left|g'(a \cdot \xi_j)\right| \geq \sqrt{\alpha(1-s)}.$$

□

Finally we have all the tools ready to prove Theorem 3.1.

Proof of Theorem 3.1 Lemma 3.6 ensures that

$$\left|g'(a \cdot \xi_{j_0})\right| \geq \sqrt{\alpha(1-s)}$$

with probability $1 - 2e^{-\frac{2m_{\mathcal{X}}s^2\alpha^2}{C_2^4}}$. Therefore, Corollary 3.3 together with Lemma 3.4 show that with probability at least

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + 2e^{-\frac{2m_{\mathcal{X}}s^2\alpha^2}{C_2^4}}\right),$$

$\hat{a}$ as defined in Algorithm 1 satisfies

$$\left\|\operatorname{sign}\left(g'(a \cdot \xi_{j_0})\right)\hat{a} - a\right\|_{\ell_2^d} \leq \frac{2\nu_1}{\sqrt{\alpha(1-s)} - \nu_1} \tag{38}$$

for the unknown sign of $g'(a \cdot \xi_{j_0})$.

Using this estimate we can prove that $\hat{f}$ as defined in Algorithm 1 is a good approximation to f. For $x \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$ we have

$$\begin{aligned}
\left|f(x) - \hat{f}(x)\right| &= \left|g(a \cdot x) - \hat{g}(\hat{a} \cdot x)\right| \\
&= \left|g(a \cdot x) - f\left(\hat{a}^T \cdot \hat{a} \cdot x\right)\right| \\
&= \left|g(a \cdot x) - g\left(a \cdot \hat{a}^T \cdot \hat{a} \cdot x\right)\right|
\end{aligned}$$

$$\begin{aligned}&\le C_2\big|a\cdot x - a\cdot\big[\hat{a}^T\hat{a}\big]\cdot x\big|\\&= C_2\big|a\cdot\big(I_d-\hat{a}^T\hat{a}\big)x\big|.\end{aligned}$$

Because $\hat{a}(I_d-\hat{a}^T\hat{a})=0$ and therefore $\operatorname{sign}(g'(a\cdot\xi_{j0}))\hat{a}(I_d-\hat{a}^T\hat{a})=0$, we can further estimate

$$\begin{aligned}\big|f(x)-\hat{f}(x)\big| &\le C_2\big|a\cdot\big(I_d-\hat{a}^T\hat{a}\big)x\big|\\&= C_2\big|\big(a-\operatorname{sign}\big(g'(a\cdot\xi_{j0})\big)\hat{a}\big)\cdot\big(I_d-\hat{a}^T\hat{a}\big)x\big|\\&\le C_2\big\|a-\operatorname{sign}\big(g'(a\cdot\xi_{j0})\big)\hat{a}\big\|_{\ell_2^d}\cdot\|x\|_{\ell_2^d}\\&\le 2C_2(1+\bar{\epsilon})\frac{\nu_1}{\sqrt{\alpha(1-s)}-\nu_1}.\end{aligned}$$ □

Remark 4 Here we present a few comments on this result.

(i) Our recovery method differs from the one proposed by Cohen, Daubechies, DeVore, Kerkyacharian, and Picard [11]. In their approach, the domain is taken to be $[0,1]^d$ and they make heavy use of the additional assumption $\sum_{j=1}^d a_j = 1$ with $a_j \ge 0$. This allows them to derive an almost completely deterministic and adaptive strategy for sampling the function f in order to first find an approximation to g and only then address the approximation to a. Here we follow somehow the opposite order, first approximating a and then finding a uniform approximation to g and, eventually, to f as well. Notice further that not having additional information on a, which is fully arbitrary in our case, we need to use a random sampling scheme which eventually gives a result holding with high probability.

(ii) Note that Theorem 3.1 gives an a priori estimate of the success probability and approximation error of Algorithm 1. If the problem parameters q, C_1, C_2, and α are known, they can be used to choose m_Φ and $m_\mathcal{X}$ big enough to have, say, a prescribed desired accuracy δ with probability at least $1-p$.

However, once Algorithm 1 has been run we have the following *a posteriori* estimate. With probability at least $1-(e^{-c_1' m_\Phi}+e^{-\sqrt{m_\Phi d}})$ we obtain

$$\|f-\hat{f}\|_\infty \le C_2(1+\bar{\epsilon})\frac{2\nu_1}{\|x_{j0}\|_{\ell_2^d}}.$$

Hence, the ratio $\frac{2\nu_1}{\|x_{j0}\|_{\ell_2^d}} \ll 1$ defines an a posteriori indicator that the number of samples $m_\mathcal{X}$ and m_Φ has been properly calibrated; otherwise, more points will be drawn until such a condition is obtained.

(iii) The parameter ϵ is chosen at the very beginning in the Taylor expansion (14) and, from a purely theoretical point of view, could be chosen arbitrarily small. Unfortunately, this may affect the numerical stability in the approximation in (14) of the derivative $\frac{\partial f}{\partial\varphi}(\xi)$ by means of a finite difference. Hence, the parameter ϵ should not be taken too small in practice. Up to some extent this may be compensated by choosing a larger number of points m_Φ in (25), as in our expression for ν_1 in (25) ϵ appears in a ratio of the form $\frac{\epsilon}{\sqrt{m_\Phi}}$. We return in more detail to this point in Sect. 5.1. In recent numerical experiments associated to the work [31], we have been experiencing very stable reconstructions with reasonable choices, e.g., $\epsilon \approx 0.1$. Hence, we do not consider this issue of any practical relevance or difficulty.

3.3 Discussion on Tractability

The approximation performances of our learning strategy are basically determined by the optimal value of α (see, e.g., (10)), which is achieved by the choice

$$\alpha := \int_{\mathbb{S}^{d-1}} \big|g'(a \cdot x)\big|^2 \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x). \tag{39}$$

For symmetry reasons this quantity does not depend on the particular choice of a.

The rotation-invariant probability measure $\mu_{\mathbb{S}^{d-1}}$ on $\mathbb{S}^{d-1}$ is induced on the sphere by the (left) Haar measure on the Lie group of all orientation-preserving rotations. For a given $k \times d$ matrix U such that $UU^T = I_k$ (i.e., with orthonormal rows) we define the measure μ_k on the unit ball $B_{\mathbb{R}^k}$ in $\mathbb{R}^k$ induced by the projection of $\mu_{\mathbb{S}^{d-1}}$ via U; i.e., for any Borel set $B \subset B_{\mathbb{R}^k}$ we define

$$\mu_k(B) := U_\# \mu_{\mathbb{S}^{d-1}}(B) := \mu_{\mathbb{S}^{d-1}}\big(U^{\leftarrow}(B)\big). \tag{40}$$

Since $\mu_{\mathbb{S}^{d-1}}$ is rotation-invariant, μ_k does not depend on the particular matrix U, and is itself a rotation-invariant measure on $B_{\mathbb{R}^k}$. Hence for any summable function $h : B_{\mathbb{R}^k} \to \mathbb{R}$, for any $k \times k$ orthogonal matrix $\mathcal{O}$ such that $\mathcal{O}\mathcal{O}^T = I_k = \mathcal{O}^T \mathcal{O}$, and for any $k \times d$ matrix U such that $UU^T = I_k$, we have the identities

$$\int_{B_{\mathbb{R}^k}} h(\mathcal{O}y) \, \mathrm{d}\mu_k(y) = \int_{B_{\mathbb{R}^k}} h(y) \, \mathrm{d}\mu_k(y) = \int_{\mathbb{S}^{d-1}} h(Ux) \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x). \tag{41}$$

The following result is well known. We refer to [30, Sect. 1.4.4] for the case of $\mathbb{C}^n$. The proof given there also works literally in the real case.

Theorem 3.7 *Let $1 \le k < d$ be natural numbers. Then the measure μ_k defined in* (40) *is given by*

$$\mathrm{d}\mu_k(y) = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \big(1 - \|y\|^2_{\ell_2^k}\big)^{\frac{d-2-k}{2}} \, \mathrm{d}y.$$

Notice that as $d \to \infty$, and for fixed k, the measure μ_k becomes more and more concentrated around 0, in the sense that, for $\varepsilon > 0$ fixed,

$$\mu_k\big(B_{\mathbb{R}^k}(\varepsilon)\big) \to 1, \quad \text{for } d \to \infty,$$

very rapidly (typically exponentially). By using the explicit form of the measure μ_k we can compute

$$\begin{aligned}\mu_k\big(B_{\mathbb{R}^k}(\varepsilon)\big) &= 1 - \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)}\int_{B_{\mathbb{R}^k}\setminus B_{\mathbb{R}^k}(\varepsilon)}\big(1-\|y\|^2_{\ell_2^k}\big)^{\frac{d-2-k}{2}}\,\mathrm{d}y \\ &= 1 - \frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)}\int_\varepsilon^1 \big(1-r^2\big)^{\frac{d-2-k}{2}} r^{k-1}\,\mathrm{d}r \\ &\ge 1 - \frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)} e^{-\frac{d-2-k}{2}\varepsilon^2}.\end{aligned}$$

By Stirling's approximation $\frac{2\Gamma(d/2)}{\Gamma(k/2)\Gamma((d-k)/2)} \approx \sqrt{\frac{d^{d-1}}{\pi k^{k-1}(d-k)^{d-k-1}}}$; thus for k and ε constant,

$$\mu_k\big(B_{\mathbb{R}^k}(\varepsilon)\big) \to 1$$

exponentially fast as $d \to \infty$. For $k = 1$, this phenomenon can be summarized informally by saying that the surface measure of the unit sphere in high dimension is concentrated around the equator [21]. Hence in the case $d \gg k$ we may want to take into account possible rescaling, i.e., working with spheres of larger radii, in order to eventually consider properties of g (actually the matrix H_g) on larger subsets of $\mathbb{R}^k$ (see also Remark 4). Without loss of generality, by keeping in mind this possible rescaling, we can assume to work with the unit sphere.

For $k = 1$, we observe that α as in (39) is determined by the interplay between the variation properties of g and the measure μ_1. As just mentioned above, the most relevant feature of μ_1 is that it concentrates around zero exponentially fast as $d \to \infty$. Hence, the asymptotic behavior of α exclusively depends on the behavior of the function g' in a neighborhood of 0.

To illustrate this phenomenon more precisely, we present the following result.

Proposition 3.8 *Let us fix $M \in \mathbb{N}$ and assume that $g : B_\mathbb{R} \to \mathbb{R}$ is C^{M+2}-differentiable in an open neighborhood $\mathcal{U}$ of* 0 *and $\frac{d^\ell}{dx^\ell} g(0) = 0$ for $\ell = 1, \ldots, M$. Then*

$$\alpha(d) = \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^1 \big|g'(y)\big|^2\big(1-y^2\big)^{\frac{d-3}{2}}\,\mathrm{d}y = \mathcal{O}\big(d^{-M}\big), \quad \textit{for } d \to \infty.$$

Proof First of all, we compute the ℓth moment of the measure $\frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \times (1-y^2)^{\frac{d-3}{2}}\mathcal{L}^1$:

$$\frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)}\int_{-1}^1 y^\ell\big(1-y^2\big)^{\frac{d-3}{2}}\,\mathrm{d}y = \frac{[1+(-1)^\ell]\Gamma(d/2)\Gamma((1+\ell)/2)}{2\sqrt{\pi}\,\Gamma((d+\ell)/2)}. \tag{42}$$

Notice that all the odd moments vanish. By a Taylor expansion of g' around 0 and by taking into account that $\frac{d^\ell}{dx^\ell} g(0) = 0$ for $\ell = 1, \dots, M$, we obtain

$$g'(y) = \sum_{\ell=1}^{M+1} \frac{1}{(\ell-1)!} \frac{d^\ell}{dx^\ell} g(0) y^{\ell-1} + \mathcal{O}\bigl(y^{M+1}\bigr) = \frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0) y^M + \mathcal{O}\bigl(y^{M+1}\bigr).$$

Hence,

$$\bigl|g'(y)\bigr|^2 = \left(\frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0)\right)^2 y^{2M} + \mathcal{O}\bigl(y^{2M+1}\bigr),$$

and

$$\begin{aligned}
\alpha(d) &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y \\
&= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \biggl(\int_{\mathcal{U}} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y \\
&\quad + \int_{B_{\mathbb{R}} \setminus \mathcal{U}} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y\biggr) \\
&= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \biggl(\left(\frac{1}{M!} \frac{d^{M+1}}{dx^{M+1}} g(0)\right)^2 \int_{\mathcal{U}} y^{2M} \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y \\
&\quad + \int_{\mathcal{U}} \mathcal{O}\bigl(y^{2M+2}\bigr)\bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y + \int_{B_{\mathbb{R}} \setminus \mathcal{U}} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y\biggr).
\end{aligned}$$

We consider the $(2M+2)$th moment in the expression above because the previous one is odd and therefore vanishes. Now, the term $\int_{B_{\mathbb{R}} \setminus \mathcal{U}} |g'(y)|^2 (1-y^2)^{\frac{d-3}{2}} dy$ goes to zero exponentially fast for $d \to 0$. Using (42) we immediately obtain

$$\begin{aligned}
\alpha(d) &= \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y \\
&= \mathcal{O}\left(\frac{\Gamma(d/2)\Gamma((1+2M)/2)}{\Gamma((d+2M)/2)}\right), \quad d \to \infty.
\end{aligned}$$

By Stirling's approximation, for which $\Gamma(z) = \sqrt{\frac{2\pi}{z}} (\frac{z}{e})^z + \mathcal{O}(1 + 1/z)$, for $z \to \infty$, we obtain

$$\frac{\Gamma(d/2)\Gamma((1+2M)/2)}{\Gamma((d+2M)/2)} \approx d^{(d-1)/2} (1+2M)^M (d+2M)^{-(\frac{d+1}{2}+M)}, \quad d \to \infty.$$

This eventually yields

$$\alpha(d) = \frac{\Gamma(d/2)}{\pi^{1/2}\Gamma((d-1)/2)} \int_{-1}^{1} \bigl|g'(y)\bigr|^2 \bigl(1-y^2\bigr)^{\frac{d-3}{2}} \,\mathrm{d}y = \mathcal{O}\bigl(d^{-M}\bigr), \quad d \to \infty. \qquad \square$$

The number $m_{\mathcal{X}} \times (m_\Phi + 1)$ of points we need to achieve a prescribed accuracy in the error estimate (24) of Theorem 3.1 depends on α. Proposition 3.8 ensures that, if $g'(y)$ does not vanish for $y \to 0$ superpolynomially, then the dependence of α (and therefore of the error estimate and the number $m_{\mathcal{X}} \times (m_\Phi + 1)$ of points) on d is at most polynomial. According to this observation we distinguish three classes of ridge functions:

(1) For $0 < q \le 1$, $C_1 > 1$ and $C_2 \ge \alpha_0 > 0$, we define

$$\begin{aligned}\mathcal{F}_d^1 := \mathcal{F}_d^1(\alpha_0, q, C_1, C_2) := \bigl\{ f : B_{\mathbb{R}^d} \to \mathbb{R} :\\ \exists a \in \mathbb{R}^d,\ \|a\|_{\ell_2^d} = 1,\ \|a\|_{\ell_q^d} \le C_1 \text{ and}\\ \exists g \in C^2(B_{\mathbb{R}}),\ \bigl|g'(0)\bigr| \ge \alpha_0 > 0 : f(x) = g(a \cdot x) \bigr\}.\end{aligned}$$

(2) For a neighborhood $\mathcal{U}$ of 0, $0 < q \le 1$, $C_1 > 1$, $C_2 \ge \alpha_0 > 0$ and $N \ge 2$, we define

$$\begin{aligned}\mathcal{F}_d^2 := \mathcal{F}_d^2(\mathcal{U}, \alpha_0, q, C_1, C_2, N) := \bigl\{ f : B_{\mathbb{R}^d} \to \mathbb{R} :\\ \exists a \in \mathbb{R}^d,\ \|a\|_{\ell_2^d} = 1,\ \|a\|_{\ell_q^d} \le C_1 \text{ and } \exists g \in C^2(B_{\mathbb{R}}) \cap C^N(\mathcal{U})\\ \exists 0 \le M \le N-1,\ \bigl|g^{(M)}(0)\bigr| \ge \alpha_0 > 0 : f(x) = g(a \cdot x) \bigr\}.\end{aligned}$$

(3) For a neighborhood $\mathcal{U}$ of 0, $0 < q \le 1$, $C_1 > 1$ and $C_2 \ge \alpha_0 > 0$, we define

$$\begin{aligned}\mathcal{F}_d^3 := \mathcal{F}_d^3(\mathcal{U}, \alpha_0, q, C_1, C_2) := \bigl\{ f : B_{\mathbb{R}^d} \to \mathbb{R} :\\ \exists a \in \mathbb{R}^d,\ \|a\|_{\ell_2^d} = 1,\ \|a\|_{\ell_q^d} \le C_1 \text{ and } \exists g \in C^2(B_{\mathbb{R}}) \cap C^\infty(\mathcal{U})\\ \bigl|g^{(M)}(0)\bigr| = 0 \text{ for all } M \in \mathbb{N} : f(x) = g(a \cdot x) \bigr\}.\end{aligned}$$

Theorem 3.1 and Proposition 3.8 immediately imply the following tractability result for these function classes.

Corollary 3.9 *The problem of learning functions f in the classes $\mathcal{F}_d^1$ and $\mathcal{F}_d^2$ from point evaluations is* strongly polynomially tractable *and* polynomially tractable, *respectively.*

On the one hand, let us notice that if in the class $\mathcal{F}_d^3$ we remove the condition $\|a\|_{\ell_q^d} \le C_1$, then the discussion on the functions described in Remark 1 shows that the problem actually becomes *intractable*. On the other hand, we conjecture that the restriction imposed by a condition such as $\|a\|_{\ell_q^d} \le C_1$ should instead give the problem some sort of tractability. Unfortunately, our learning method and approximation estimates in Theorem 3.1 do not provide any information about the tractability of the problem for functions in the class $\mathcal{F}_d^3$.

4 The General Case $k \geq 1$

In this section we generalize our approach to the case $k \geq 1$; i.e., we consider *k-ridge functions*

$$f(x) = g(Ax). \tag{43}$$

Obviously, the sum of k ridge functions (as appearing for example in (13)) is a k-ridge function, and the same holds true for the product.

We will proceed as in the one-dimensional case, first giving the basic ideas, which motivate the recovery algorithm, and then stating and proving our main theorem. Remember that we assume that A is a $k \times d$ matrix such that $AA^T = I_k$, and $g : B_{\mathbb{R}^k}(1+\bar{\epsilon}) \to \mathbb{R}$ is a C^2 function.

4.1 The Algorithm

As before, we consider a version of Taylor's theorem giving access to the matrix A. For $\xi \in B_{\mathbb{R}^d}$, $\varphi \in B_{\mathbb{R}^d}(r)$, $\epsilon, r \in \mathbb{R}_+$, with $r\epsilon \leq \bar{\epsilon}$, we have the identity

$$\bigl[\nabla g(A\xi)^T A\bigr]\varphi = \frac{f(\xi + \epsilon\varphi) - f(\xi)}{\epsilon} - \frac{\epsilon}{2}\bigl[\varphi^T \nabla^2 f(\zeta)\varphi\bigr], \tag{44}$$

for a suitable $\zeta(\xi, \varphi) \in B_{\mathbb{R}^d}(1+\bar{\epsilon})$, and from (9) the term $[\varphi^T \nabla^2 f(\zeta)\varphi]$ is again uniformly bounded as soon as φ is bounded.

As in the one-dimensional case we now consider (44) for the m_Φ directions in the set Φ and at the $m_\mathcal{X}$ sampling points in the set $\mathcal{X}$, where $\mathcal{X}, \Phi$ are defined as in (15) and (16), respectively. Again we collect the directional derivatives $\nabla g(A\xi_j)^T A$, $j = 1, \ldots, m_\mathcal{X}$ as columns in the $d \times m_\mathcal{X}$ matrix X, i.e.,

$$X = \bigl(A^T \nabla g(A\xi_1), \ldots, A^T \nabla g(A\xi_{m_\mathcal{X}})\bigr), \tag{45}$$

and using the matrices Y and $\mathcal{E}$ as defined in (18) and (19), we can write the following factorization:

$$\Phi X = Y - \mathcal{E}. \tag{46}$$

Similarly to the one-dimensional case we find that the matrix X has a special structure, which we will exploit for the algorithm, i.e., $X = A^T \mathcal{G}^T$, where $\mathcal{G} = (\nabla g(A\xi_1)^T | \cdots | \nabla g(A\xi_{m_\mathcal{X}})^T)^T$. The columns of X are now no longer scaled copies of one compressible vector, but are linear combinations of k compressible vectors, i.e., the rows of the matrix A. Thus compressed sensing theory again tells us that we can stably recover the columns of X from the columns of Y via ℓ_1-minimization and consequently get a good approximation $\hat{X}$ to X.

Furthermore, since A has rank k, as long as $\mathcal{G}^T$ has full rank, X will also have rank k; also the column span of the right singular vectors of $X^T = USV^T$ will coincide with the row span of A, i.e., $A^T A = VV^T$. Moreover, V^T gives us an alternative representation of f as follows:

$$f(x) = g(Ax) = g\bigl(AA^T Ax\bigr) = g\bigl(AVV^T x\bigr) =: \tilde{g}\bigl(V^T x\bigr),$$

where $\tilde{g}(y) := g(AVy) = f(Vy)$. If $\hat{X}$ is a good approximation of X, then we can expect the first k right singular vectors of $\hat{X}$ to have almost the same span as that of X and thus of A, which inspires the following algorithm.

Algorithm 2

- Given $m_\Phi, m_{\mathcal{X}}$, draw at random the sets Φ and $\mathcal{X}$ as in (15) and (16), and construct Y according to (18)
- Set $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j = \Phi z} \|z\|_{\ell_1^d}$, for $j = 1, \dots, m_{\mathcal{X}}$, and $\hat{X} = (\hat{x}_1, \dots, \hat{x}_{m_{\mathcal{X}}})$
- Compute the singular value decomposition of

$$\hat{X}^T = \begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix}, \tag{47}$$

  where $\hat{\Sigma}_1$ contains the k largest singular values
- Set $\hat{A} = V_1^T$
- Define $\hat{g}(y) := f(\hat{A}^T y)$ and $\hat{f}(x) := \hat{g}(\hat{A}x)$

The quality of the final approximation of f by means of $\hat{f}$ depends on two kinds of accuracies:

1. The error between $\hat{X}$ and X, which can be controlled through the number of compressed sensing measurements m_Φ;
2. The stability of the span of V^T, simply characterized by how well the singular values of X or equivalently $\mathcal{G}$ are separated from 0, which is related to the number of random samples $m_{\mathcal{X}}$.

To be precise, in the next section we will prove the following approximation result.

Theorem 4.1 *Let* $\log d \le m_\Phi \le [\log 6]^2 d$. *Then there is a constant* c_1' *such that using* $m_{\mathcal{X}} \cdot (m_\Phi + 1)$ *function evaluations of* f, *Algorithm* 2 *defines a function* $\hat{f} : B_{\mathbb{R}^d}(1 + \bar{\epsilon}) \to \mathbb{R}$ *that, with probability*

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}} + k e^{\frac{-m_{\mathcal{X}} \alpha s^2}{2kC_2^2}} \right), \tag{48}$$

will satisfy

$$\|f - \hat{f}\|_\infty \le 2C_2 \sqrt{k}(1 + \bar{\epsilon}) \frac{\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}, \tag{49}$$

where

$$\nu_2 = C\left(k^{1/q} \left[\frac{m_\Phi}{\log(d/m_\Phi)} \right]^{1/2 - 1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}} \right),$$

and C *depends only on* C_1 *and* C_2 *(cf.* (8) *and* (9)*).*

4.2 The Analysis

We will first show that $\hat{X}$ is a good approximation to X by applying Theorem 3.2 columnwise. This leads to the following corollary.

Corollary 4.2 *Let* $\log d \le m_\Phi < [\log 6]^2 d$. *Then with probability*

$$1 - \left(e^{-c_1' m_\Phi} + e^{-\sqrt{m_\Phi d}}\right)$$

the matrix $\hat{X}$ *as calculated in Algorithm 2 satisfies*

$$\|X - \hat{X}\|_F \le C\sqrt{m_{\mathcal{X}}}\left(k^{1/q}\left[\frac{m_\Phi}{\log(d/m_\Phi)}\right]^{1/2-1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}}\right), \tag{50}$$

where C *depends only on* C_1 *and* C_2 (*cf.* (8) *and* (9)).

Proof The proof works essentially like that of Corollary 3.3. We decompose

$$\|X - \hat{X}\|_F^2 = \sum_{j=1}^{m_{\mathcal{X}}} \|x_j - \hat{x}_j\|_{\ell_2^d}^2.$$

The best K-term approximation of x_j may be estimated using

$$\|x_j\|_{\ell_q^d} = \left\|A^T \nabla g(A\xi_j)\right\|_{\ell_q^d} \le C_2\left(\sum_{v=1}^{d}\left(\sum_{u=1}^{k}|a_{uv}|\right)^q\right)^{1/q} \le C_1 C_2 k^{1/q},$$

which leads to

$$K^{-1/2}\sigma_K(x_j)_{\ell_1^d} \le \|x_j\|_{\ell_q^d} K^{1/2-1/q} \le C_1 C_2 k^{1/q} K^{1/2-1/q}.$$

The norms of ε_j may be estimated similarly to the proof of Corollary 3.3 as

$$\|\varepsilon_j\|_{\ell_2^{m_\Phi}} \le \frac{C_1^2 C_2 k^2 \epsilon}{2\sqrt{m_\Phi}} \quad \text{and} \quad \|\varepsilon_j\|_{\ell_\infty^{m_\Phi}} \le \frac{C_1^2 C_2 k^2 \epsilon}{2 m_\Phi}.$$

Putting all these estimates (with the choice $K \approx m_\Phi/\log(d/m_\Phi)$) into Theorem 3.2 we obtain the result. □

Remark 5 The construction $\hat{x}_j = \Delta(y_j) := \arg\min_{y_j = \Phi z}\|z\|_{\ell_1^d}$, for $j = 1, \dots, m_{\mathcal{X}}$, and $\hat{X} = (\hat{x}_1, \dots, \hat{x}_{m_{\mathcal{X}}})$ and Corollary 4.2 are not a unique possible approach to approximate X. As we are expecting X to be a k-rank matrix for $k \ll \min\{d, m_{\mathcal{X}}\}$, one might want to consider also *nuclear norm minimization*, i.e., the minimization of the ℓ_1-norm of singular values, as a possible way of accessing X from m_Φ random measurements, as in the work [15, 26, 28]. However, presently no estimates of the type (29) are available in this context. Hence we postpone an analysis based on these methods fully tailored to matrices to further research.

Next we need the equivalent of Lemma 3.4 to relate the error between the subspaces defined by the largest right singular values of $\hat{X}$ and X, respectively, to the error $\|X-\hat{X}\|_F$. We will develop the necessary tools in the following subsection.

4.2.1 Stability of the Singular Value Decomposition

Given two matrices B and $\hat{B}$ with corresponding singular value decompositions

$$B=\begin{pmatrix} U_1 & U_2 \end{pmatrix}\begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

and

$$\hat{B}=\begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix}\begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix}\begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix},$$

where it is understood that two corresponding submatrices, e.g., $U_1, \hat{U}_1$, have the same size, we would like to bound the difference between V_1 and $\hat{V}_1$ by the error $\|B-\hat{B}\|_F$. As a consequence of Wedin's perturbation bound [34] (see also [32, Sect. 7]), we have the following useful result.

Theorem 4.3 (Stability of subspaces: Wedin's bound) *If there is an* $\bar{\alpha}>0$ *such that*

$$\min_{\ell,\hat{\ell}}\bigl|\sigma_{\hat{\ell}}(\hat{\Sigma}_1)-\sigma_\ell(\Sigma_2)\bigr|\geq\bar{\alpha}, \tag{51}$$

and

$$\min_{\hat{\ell}}\bigl|\sigma_{\hat{\ell}}(\hat{\Sigma}_1)\bigr|\geq\bar{\alpha}, \tag{52}$$

then

$$\bigl\|V_1V_1^T-\hat{V}_1\hat{V}_1^T\bigr\|_F\leq\frac{2}{\bar{\alpha}}\|B-\hat{B}\|_F. \tag{53}$$

The conditions (51) and (52) are separation conditions. The first says that the singular values of Σ_1 are separated from those of Σ_2, although, strictly speaking, the separation is between Σ_1 and $\hat{\Sigma}_2$. However, if $\|B-\hat{B}\|_F$ is sufficiently small compared to $\bar{\alpha}$, then Weyl's inequality [35]

$$\bigl|\sigma_\ell(B)-\sigma_\ell(\hat{B})\bigr|\leq\|B-\hat{B}\|_F$$

guarantees that the two separations are essentially equivalent. The second condition says that the singular values of Σ_1 or $\hat{\Sigma}_1$ must be far away from 0.

Applied to our situation, where X has rank k and thus $\Sigma_2=0$, we obtain

$$\bigl\|V_1V_1^T-\hat{V}_1\hat{V}_1^T\bigr\|_F\leq\frac{2\sqrt{m_{\mathcal{X}}}\nu_2}{\sigma_k(\hat{X}^T)}, \tag{54}$$

and further, since $\sigma_k(\hat{X}^T) \geq \sigma_k(X^T) - \|X - \hat{X}\|_F$, we obtain

$$\left\| V_1 V_1^T - \hat{V}_1 \hat{V}_1^T \right\|_F \leq \frac{2\sqrt{m_{\mathcal{X}}}\nu_2}{\sigma_k(X^T) - \sqrt{m_{\mathcal{X}}}\nu_2}. \tag{55}$$

As a final ingredient we need to estimate the kth singular value of X. The next subsection will provide us with a generalization of Hoeffding's inequality that can be used to show that, with high probability, on a random draw of the sampling points ξ_j the kth singular value of X is separated from zero.

4.2.2 Spectral Estimates and Sums of Random Semidefinite Matrices

The following theorem generalizes Hoeffding's inequality to sums of random semidefinite matrices and was recently proved by Tropp in [33, Corollary 5.2 and Remark 5.3], improving results from [1] and using techniques from [29] and [25].

Theorem 4.4 (Matrix Chernoff) *Consider $X_1, \dots, X_m$ independent random, positive semidefinite matrices of dimension $k \times k$. Moreover, suppose*

$$\sigma_1(X_j) \leq C, \tag{56}$$

almost surely. Compute the singular values of the sum of the expectations

$$\mu_{\max} = \sigma_1\left(\sum_{j=1}^m \mathbb{E}X_j\right) \quad \textit{and} \quad \mu_{\min} = \sigma_k\left(\sum_{j=1}^m \mathbb{E}X_j\right). \tag{57}$$

Then

$$\mathbb{P}\left\{\sigma_1\left(\sum_{j=1}^m X_j\right) - \mu_{\max} \geq s\mu_{\max}\right\} \leq k\left(\frac{(1+s)}{e}\right)^{-\frac{\mu_{\max}(1+s)}{C}}, \tag{58}$$

for all $s > (e-1)$, and

$$\mathbb{P}\left\{\sigma_k\left(\sum_{j=1}^m X_j\right) - \mu_{\min} \leq -s\mu_{\min}\right\} \leq ke^{-\frac{\mu_{\min}s^2}{2C}}, \tag{59}$$

for all $s \in (0,1)$.

Applied to the matrix X^T this theorem leads to the following estimate of the singular values of X^T.

Lemma 4.5 *For any $s \in (0,1)$ we have that*

$$\sigma_k\big(X^T\big) \geq \sqrt{m_{\mathcal{X}}\alpha(1-s)} \tag{60}$$

with probability $1 - ke^{\frac{-m_{\mathcal{X}}\alpha s^2}{2kC_2^2}}$.

Proof The proof is based on an application of Theorem 4.4. First note that

$$X^T = \mathcal{G}A = U_{\mathcal{G}}\Sigma_{\mathcal{G}}\big[V_{\mathcal{G}}^T A\big],$$

hence $\Sigma_{X^T} = \Sigma_{\mathcal{G}}$. Moreover,

$$\sigma_i(\mathcal{G}) = \sqrt{\sigma_i\big(\mathcal{G}^T\mathcal{G}\big)}, \quad \text{for all } i = 1, \dots, k.$$

Thus, to get information about the singular values of X^T it is sufficient to study those of

$$\mathcal{G}^T\mathcal{G} = \sum_{j=1}^{m_{\mathcal{X}}} \nabla g(A\xi_j)\nabla g(A\xi_j)^T.$$

We further notice that

$$\sigma_1\big(\nabla g(A\xi_j)\nabla g(A\xi_j)^T\big) \le \left(\sum_{\ell,\ell'=1}^{k} \big|\nabla g(A\xi_j)_\ell \nabla g(A\xi_j)_{\ell'}\big|^2\right)^{1/2} \le kC_2^2 := C.$$

Hence $X_j = \nabla g(A\xi_j)\nabla g(A\xi_j)^T$ is a random positive semidefinite matrix that is almost surely bounded. Moreover,

$$\mathbb{E}X_j = \mathbb{E}_\xi \nabla g(A\xi_j)\nabla g(A\xi_j)^T = \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T \, \mathrm{d}\mu_{\mathbb{S}^{d-1}}(x) = H_g.$$

Hence, remembering that the singular values of H_g are equivalent to those of H^f, by condition (10) we have $\mu_{\max} = m_{\mathcal{X}}\sigma_1(H_g) \le m_{\mathcal{X}}kC_2^2$ and $\mu_{\min} = m_{\mathcal{X}}\sigma_k(H_g) \ge m_{\mathcal{X}}\alpha > 0$. In particular,

$$m_{\mathcal{X}}k^2C_2 \ge \mu_{\max} \ge \mu_{\min} \ge m_{\mathcal{X}}\alpha > 0.$$

By an application of Theorem 4.4 we conclude that

$$\begin{aligned}\sigma_k\big(X^T\big) = \sigma_k(\mathcal{G}) &= \sqrt{\sigma_k\left(\sum_{j=1}^{m_{\mathcal{X}}} \nabla g(A\xi_j)\nabla g(A\xi_j)^T\right)} \\ &\ge \sqrt{\mu_{\min}(1-s)} \ge \sqrt{m_{\mathcal{X}}\alpha(1-s)},\end{aligned}$$

with probability

$$1 - ke^{-\frac{\mu_{\min}s^2}{2kC_2^2}} \ge 1 - ke^{\frac{-m_{\mathcal{X}}\alpha s^2}{2kC_2^2}},$$

for all $s \in (0, 1)$. □

Finally we have collected all the results necessary to prove Theorem 4.1.

Proof of Theorem 4.1 Combining Corollary 4.2, Theorem 4.3, and Lemma 4.5 shows that with probability at least

$$1-\Big(e^{-c_1' m_\Phi}+e^{-\sqrt{m_\Phi d}}+ke^{\frac{-m_{\mathcal{X}}\alpha s^2}{2kC_2^2}}\Big),$$

for the first k right singular vectors of $\hat{X}$ and X we have

$$\big\|V_1V_1^T-\hat{V}_1\hat{V}_1^T\big\|_F\le\frac{2\nu_2}{\sqrt{\alpha(1-s)}-\nu_2}.$$

Recalling from the proof of Lemma 4.5 that the (first k) right singular vectors V_1^T of X^T have the form $V_1^T=V_{\mathcal{G}}^TA$ then shows that $\hat{A}$ as defined in Algorithm 2 satisfies

$$\begin{aligned}\big\|A^TA-\hat{A}^T\hat{A}\big\|_F&=\big\|A^TV_{\mathcal{G}}V_{\mathcal{G}}^TA-\hat{V}_1\hat{V}_1^T\big\|_F\\&=\big\|V_1V_1^T-\hat{V}_1\hat{V}_1^T\big\|_F\le\frac{2\nu_2}{\sqrt{\alpha(1-s)}-\nu_2}.\end{aligned}$$

Using this estimate we can prove that $\hat{f}$ as defined in Algorithm 2 is a good approximation to f. Since A is row-orthogonal we have $A=AA^TA$ and therefore

$$\begin{aligned}\big|f(x)-\hat{f}(x)\big|&=\big|g(Ax)-\hat{g}(\hat{A}x)\big|\\&=\big|g(Ax)-g\big(A\hat{A}^T\hat{A}x\big)\big|\\&\le C_2\sqrt{k}\big\|Ax-A\hat{A}^T\hat{A}x\big\|_{\ell_2^k}\\&=C_2\sqrt{k}\big\|A\big(A^TA-\hat{A}^T\hat{A}\big)x\big\|_{\ell_2^k}\\&\le C_2\sqrt{k}\big\|\big(A^TA-\hat{A}^T\hat{A}\big)\big\|_F\|x\|_{\ell_2^d}\\&\le 2C_2\sqrt{k}(1+\bar{\epsilon})\frac{\nu_2}{\sqrt{\alpha(1-s)}-\nu_2}.\end{aligned}$$ □

Remark 6 (i) Note that Theorem 4.1 is again an a priori estimate of the success probability and approximation error of Algorithm 2. Once Algorithm 2 has been run we have the following a posteriori estimate. With probability at least $1-(e^{-c_1' m_\Phi}+e^{-\sqrt{m_\Phi d}})$ we have that

$$\|f-\hat{f}\|_\infty\le 2C_2\sqrt{km_{\mathcal{X}}}(1+\bar{\epsilon})\frac{\nu_2}{\sigma_k(\hat{X}^T)}.$$

(ii) We further observe that Theorem 4.1 does not straightforwardly reduce to Theorem 3.1 for $k=1$, because in the one-dimensional case we used the simpler maximum strategy as in (22) instead of the singular value decomposition (47).

4.3 Discussion on Tractability

Recall that the push-forward measure $\mu_k = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)}(1 - \|y\|^2_{\ell_2^k})^{\frac{d-2-k}{2}}\mathcal{L}^k$ of $\mu_{\mathbb{S}^{d-1}}$ on the unit ball $B_{\mathbb{R}^k}$ was determined in Theorem 3.7 as the measure for which

$$\begin{aligned} H_g &= \int_{\mathbb{S}^{d-1}} \nabla g(Ax)\nabla g(Ax)^T \,\mathrm{d}\mu_{\mathbb{S}^{d-1}}(x) \\ &= \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \nabla g(y)\nabla g(y)^T \bigl(1 - \|y\|^2_{\ell_2^k}\bigr)^{\frac{d-2-k}{2}}\,\mathrm{d}y. \end{aligned}$$

As an instructive example, let us apply this formula to the case when g is a radial function, i.e.,

$$g(y) = g_0\bigl(\|y\|_{\ell_2^k}\bigr),$$

for a function $g_0 : [0, 1] \to \mathbb{R}$ sufficiently smooth, and $g_0'(0) = 0$.

A direct calculation shows that $\nabla g(y) = \frac{g_0'(r)}{r} \cdot y$, where $r = \|y\|_{\ell_2^k}$, and

$$\nabla g(y)\nabla g(y)^T = \frac{g_0'(r)^2}{r^2} yy^T.$$

Hence,

$$(H_g)_{ij} = \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \frac{g_0'(\|y\|_{\ell_2^k})^2}{\|y\|^2_{\ell_2^k}} y_i y_j \bigl(1 - \|y\|^2_{\ell_2^k}\bigr)^{\frac{d-2-k}{2}}\,\mathrm{d}y.$$

If $i \neq j$, the integral vanishes due to the symmetry of $B_{\mathbb{R}^k}$. If $i = j$, we get again by symmetry

$$\begin{aligned} (H_g)_{ii} &= \frac{\Gamma(d/2)}{\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} \frac{g_0'(\|y\|_{\ell_2^k})^2}{\|y\|^2_{\ell_2^k}} y_i^2 \bigl(1 - \|y\|^2_{\ell_2^k}\bigr)^{\frac{d-2-k}{2}}\,\mathrm{d}y \\ &= \frac{\Gamma(d/2)}{k\pi^{k/2}\Gamma((d-k)/2)} \int_{B_{\mathbb{R}^k}} g_0'\bigl(\|y\|_{\ell_2^k}\bigr)^2 \bigl(1 - \|y\|^2_{\ell_2^k}\bigr)^{\frac{d-2-k}{2}}\,\mathrm{d}y \\ &= \frac{2\Gamma(d/2)}{k\Gamma((d-k)/2)\Gamma(k/2)} \int_0^1 g_0'(r)^2 \bigl(1 - r^2\bigr)^{\frac{d-2-k}{2}} r^{k-1}\,\mathrm{d}r \\ &=: \alpha(k, d). \end{aligned}$$

Hence, $H_g = \alpha(k, d)I_k$. Similarly to Proposition 3.8, we can expand g_0' into a Taylor series,

$$g_0'(r) = \sum_{\ell=2}^{N-1} \frac{g_0^{(\ell)}(0)}{(\ell-1)!} r^{\ell-1} + \mathcal{O}\bigl(r^N\bigr).$$

If we assume that $g_0^{(\ell)}(0)=0$, for all $\ell=1,\dots,M$, but $g_0^{(M+1)}(0)\neq 0$, then we obtain

$$g_0'(r)^2=\left(\frac{g_0^{(M+1)}(0)}{M!}\right)^2 r^{2M}+\mathcal{O}\bigl(r^{2M+1}\bigr),$$

and, by Stirling's approximation,

$$\begin{aligned}\alpha(k,d)&=\mathcal{O}\left(\frac{\Gamma(d/2)}{\Gamma((d-k)/2)}\int_0^1 r^{2M+k-1}\bigl(1-r^2\bigr)^{\frac{d-k-2}{2}}\,\mathrm{d}r\right)\\&=\mathcal{O}\left(\frac{\Gamma(d/2)}{\Gamma(d/2+M)}\right)\\&=\mathcal{O}\bigl(d^{-M}\bigr),\quad d\to\infty.\end{aligned}$$

From these computations we deduce that, using our method, learning functions $f(x)=g(Ax)$, where g is radial (or nearly radial), usually have polynomial complexity with respect to the dimension d.

5 Extensions and Generalizations

We assumed throughout the paper that the function f is defined on the unit ball $B_{\mathbb{R}^d}$ of $\mathbb{R}^d$. To be able to approximate the derivatives of f even on the boundary of $B_{\mathbb{R}^d}$, we actually assumed that f is also defined on an $\bar{\epsilon}$ neighborhood of the unit ball. Furthermore, we assumed that the function values may be measured exactly without any error. The main aim of this section is to discuss the possibilities and limitations of our method. First, we discuss the numerical stability of our approach with respect to noise. Second, we deal with functions defined on a convex body $\Omega\subset\mathbb{R}^d$. It is our intention here only to sketch (although rigorously) further interesting research directions, so we limit our discussion to the case of $k=1$.

5.1 Stability Under Noisy Measurements

Let us assume that the function evaluation in (14) can be performed only with certain precision. We again collect the $m_{\mathcal{X}}\times m_{\Phi}$ instances of (14) as

$$\Phi X=Y-\mathcal{E}+\frac{\mathcal{W}}{\epsilon},\tag{61}$$

where the (i,j)th entry of $\mathcal{W}$ (denoted by w_{ij}) is the difference between the exact value of $f(\xi_j+\epsilon\varphi_i)-f(\xi_j)$ and its value measured with noise. This leads to a compressed sensing setting

$$Y=\Phi X+\mathcal{E}-\frac{\mathcal{W}}{\epsilon}.\tag{62}$$

Applying Theorem 3.2 we obtain a substitute for Corollary 3.3 with $\mathcal{E}$ replaced by $\mathcal{E}-\mathcal{W}/\epsilon$. Therefore, we would like to estimate the norm of w_j (the jth column of $\mathcal{W}$) in

$\ell_2^{m_\Phi}$ and $\ell_\infty^{m_\Phi}$. If we merely assume that the noise is bounded (i.e., $|w_{ij}| \le \nu$), the best possible estimate is $\|w_j\|_{\ell_2^{m_\Phi}} \le \nu\sqrt{m_\Phi}$. We observe that the more sampling points we take, the greater the level of noise. This effect of noise amplification is known as *noise folding* [2] and, unfortunately, it corrupts the estimate (31). See also [11, Sect. 4] for a discussion in a related context.

Let us therefore sketch a different approach. We make the rather natural assumption that w_{ij} is a random noise.

The analogue of Theorem 3.2 for the recovery of x from noisy measurements $y = \Phi x + \omega$, where $\omega = (\omega_1, \dots, \omega_m)$ are independent identically distributed (i.i.d.) Gaussian variables with mean zero and variance σ^2, was given in the work of Candès and Tao [9]. They proposed a certain ℓ_1-regularization problem, whose solution (called the *Dantzig selector*) satisfies

$$\|x - \hat{x}\|_{\ell_2^d}^2 \le C^2 \cdot 2\log d \cdot \left(\sigma^2 + \sum_{i=1}^{d} \min\bigl(x_i^2, \sigma^2\bigr)\right).$$

In particular, if x is a k-sparse vector, then $\|x - \hat{x}\|_{\ell_2^d} \le C \cdot \sqrt{2\log d} \cdot \sqrt{k+1} \cdot \sigma$. This estimate scales very favorably with d (only as $\sqrt{\log d}$) and, moreover, depends only on the sparsity of x, and no longer on the number of measurements m_Φ. Therefore, there is no noise folding in this case.

Equation (62) requires a combination of Theorem 3.2 and the result of Candès and Tao. Namely, we would like to reconstruct x if $y = \Phi x + \varepsilon + \omega$ is given, where ε is a deterministic error and ω is a vector of i.i.d. Gaussian variables. A detailed analysis of this issue goes beyond the scope of this paper. Nevertheless, let us present some numerical evidence of the numerical stability of our approach in the presence of random noise.

We consider the function

$$f(x) = \max\Bigl(\Bigl[1 - 5\sqrt{(x_3 - 1/2)^2 + (x_4 - 1/2)^2}\Bigr]^3, 0\Bigr), \quad x \in \mathbb{R}^{1000} \tag{63}$$

in dimension $d = 1000$. We use a variant of Algorithm 1 based on ℓ_1-minimization to identify the active coordinates of f; see [31] for details. We assume that function evaluations were distorted by Gaussian error $\nu\omega$ with $\omega \approx \mathcal{N}(0,1)$ and $\nu \in \{0.1, 0.01, 0.001\}$. We chose $\epsilon = 0.1$ in the approximation (14). For each number of points $m_{\mathcal{X}} \in \{6\ell, \ell = 1, \dots, 10\}$ (x-axis) and each number of directions $m_\Phi \in \{20\ell, \ell = 1, \dots, 10\}$ (y-axis) we produced 100 trials. As shown in Fig. 2, the success rates of recovery go from white (no success) to black (100 successful recoveries).

We conclude from Fig. 2 that there is a smooth increase of the rate of successful recovery with decreasing noise power and a fully stable recovery behavior.

5.2 Convex Bodies

A careful inspection of our method shows that it may be generalized to arbitrary convex bodies. Let us describe the necessary modifications and give an overview of

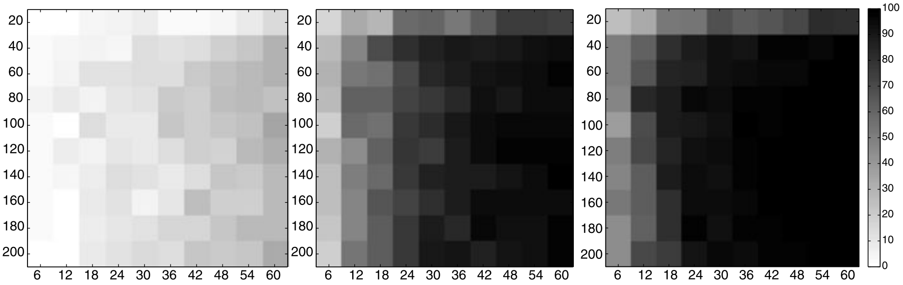


Fig. 2 Recovery of active coordinates of $f(x)$ given by (63) with $\nu = 0.1$, $\nu = 0.01$, and $\nu = 0.001$, *from left to right*, respectively. Note that the success rates of recovery for the noise-free setting are hardly distinguishable from the last picture above ($\nu = 0.001$)

the results for the case $k = 1$. First, one has to replace (6) by

$$H^f := \int_\Omega \nabla f(x) \nabla f(x)^T \, \mathrm{d}\mu_\Omega(x). \tag{64}$$

Here, μ_Ω is a probability measure on Ω, and the points in $\mathcal{X}$ (cf. (15)) are selected at random with respect to μ_Ω. For $\Omega = B_{\mathbb{R}^d}$, we simply selected $\mu_\Omega = \mu_{\mathbb{S}^{d-1}}$ to be the normalized surface measure on $\mathbb{S}^{d-1}$. This corresponded to the fact that $a \in \mathbb{S}^{d-1}$ was arbitrary and therefore a priori no direction was preferred. To be able to evaluate the derivatives of f even on the boundary of Ω, we assume that f is actually defined on an $\bar{\epsilon}$ neighborhood of Ω, namely on the set $\Omega + \bar{\epsilon} := \{x \in \mathbb{R}^d : \mathrm{dist}(\Omega, x) \le \bar{\epsilon}\}$. The function g is assumed to be defined on the image of $\Omega + \bar{\epsilon}$ under the mapping $x \to a \cdot x$, i.e., on an interval. We again assume (9).

Surprisingly, these are all the modifications necessary to proceed with the identification of $\hat{a}$, and (38) holds true under these circumstances.

The proof of Theorem 3.1 was based on the fact that, for every $y \in B_\mathbb{R}$, we can easily find an element $x_y \in B_{\mathbb{R}^d}$, such that $\hat{a} \cdot x_y = y$. It is enough to consider $x_y = \hat{a}^T y$. In the case of a general convex set Ω, we first need to define, for any $\hat{a} \in \mathbb{S}^{d-1}$ fixed, a function $x_\cdot : \hat{a}(\Omega + \bar{\epsilon}) \to \Omega + \bar{\epsilon}$ given by $y \mapsto x_y$, and such that

$$\hat{a} \cdot x_y = y.$$

In particular, for all $y \in \hat{a}(\Omega + \bar{\epsilon})$ we need to find

$$x_y \in \Omega + \bar{\epsilon} \cap \left\{ x \in \mathbb{R}^d : \hat{a} \cdot x = y \right\}.$$

Since both $\Omega + \bar{\epsilon}$ and the solution space $\{x \in \mathbb{R}^d : \hat{a} \cdot x = y\}$ are closed convex sets in $\mathbb{R}^d$, one could use an alternating projection algorithm for finding x_y [4]. Thus, we can assume that, at least algorithmically, this map can be computed. Moreover, and alternatively, since the operation described above, i.e., finding $x_y \in B_{\mathbb{R}^d}$ such that $\hat{a} \cdot x_y = y$, must be executed as many times as we need to define, e.g., an appropriate spline approximation of $\hat{g}$, we may proceed as follows. First we find $x_{\max}, x_{\min} \in B_{\mathbb{R}^d}$, such that $\hat{a} \cdot x_{\max} = \max_{x \in B_{\mathbb{R}^d}} \hat{a} \cdot x$ and $\hat{a} \cdot x_{\min} = \min_{x \in B_{\mathbb{R}^d}} \hat{a} \cdot x$. Then any other

x_y such that $y = \hat{a} \cdot x_y$ is computed very quickly by $x_y = \lambda_y x_{\min} + (1 - \lambda_y) x_{\max}$ for some $\lambda_y \in [0, 1]$.

With this modification, Theorem 3.1 also holds true, with the definition of $\hat{g}$ given in Algorithm 1 now replaced by

$$\hat{g}(y) := f(x_y), \quad y \in \hat{a}(\Omega + \bar{\epsilon})$$

and (24) replaced by

$$\|f - \hat{f}\|_\infty \le 2C_2\big(\operatorname{diam}(\Omega) + 2\bar{\epsilon}\big)\frac{\nu_1}{\sqrt{\alpha(1-s)} - \nu_1}.$$

Unfortunately—and this seems to be the main drawback of this approach—the diameter of Ω, $\operatorname{diam}(\Omega) = \max_{x,x' \in \Omega} \|x - x'\|_{\ell_2^d}$, may grow with d. This is especially the case, when $\Omega = [-1, 1]^d$, which gives $\operatorname{diam}(\Omega) = \sqrt{2d}$.

5.3 An Approach Through the Minkowski Functional

To get better results for specific convex bodies (i.e., $\Omega = [-1, 1]^d$), we propose another approach. We stress very clearly that up to now this is only to be understood as an open direction for further research.

We assume that Ω is a closed convex set which is *absorbing* and *balanced*, i.e.,

- for every $x \in \mathbb{R}^d$, there is a $t = t(x) > 0$, such that $tx \in \Omega$
- $\alpha\Omega := \{\alpha x : x \in \Omega\} \subset \Omega$ for every $\alpha \in [-1, 1]$

Then we can define its Minkowski functional as

$$p_\Omega(x) := \inf\{r > 0 : x/r \in \Omega\}, \quad x \in \mathbb{R}^d.$$

It is well known that this expression is actually a norm and Ω is its unit ball. Hence,

$$\sup_{x,x' \in \Omega} p_\Omega\big(x - x'\big) \le 2. \tag{65}$$

This allows us to replace the inequality

$$\big|(a - \hat{a}) \cdot (x_y - x)\big| \le \|a - \hat{a}\|_2 \cdot \|x_y - x\|_2$$

by

$$\big|(a - \hat{a}) \cdot (x_y - x)\big| \le \|a - \hat{a}\|'_\Omega \cdot \|x_y - x\|_\Omega.$$

Here, $\|\cdot\|_\Omega = p_\Omega(\cdot)$ and $\|\cdot\|'_\Omega$ is its dual norm. According to (65), this solves the problem of the factor $\operatorname{diam}(\Omega)$—the diameter of Ω with respect to $\|\cdot\|_\Omega$ is always bounded by 2. Unfortunately, the problem is transferred to the second factor, namely $\|a - \hat{a}\|'_\Omega$. For this, one would need the analogue of Theorem 3.2 with the ℓ_2^d-norm in (29) replaced by $\|\cdot\|'_\Omega$. While any treatment of this general case is clearly beyond the scope of this paper and remains a topic for further investigation, we can briefly sketch what happens in the special case $\Omega = [-1, 1]^d$. Then we simply have $\|\cdot\|_\Omega = \|\cdot\|_{\ell_\infty^d}$ and $\|\cdot\|'_\Omega = \|\cdot\|_{\ell_1^d}$. To estimate $\|a - \hat{a}\|_{\ell_1^d}$ we would have to combine Lemma 3.1 in [11] with (38) and would get again a result that does not depend on the dimension d.

Acknowledgements Massimo Fornasier would like to thank Ronald A. DeVore for his kind and warm hospitality at Texas A&M University and for the very exciting daily joint discussions which later inspired part of this work. We acknowledge the financial support provided by the START-award "Sparse Approximation and Optimization in High Dimensions" of the Fonds zur Förderung der wissenschaftlichen Forschung (FWF, Austrian Science Foundation). We would also like to thank the anonymous referees for their very valuable comments and remarks.

References

1. R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, *IEEE Trans. Inf. Theory* **48**(3), 569–579 (2002).
2. E. Arias-Castro, Y.C. Eldar, Noise folding in compressed sensing, *IEEE Signal Process. Lett.* **18**(8), 478–481 (2011).
3. R.G. Baraniuk, M. Davenport, R.A. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, *Constr. Approx.* **28**(3), 253–263 (2008).
4. H. Bauschke, H. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Rev.* **38**(3), 367–426 (1996).
5. E.J. Candès, Harmonic analysis of neural networks, *Appl. Comput. Harmon. Anal.* **6**(2), 197–218 (1999).
6. E.J. Candès, Ridgelets: estimating with ridge functions, *Ann. Stat.* **31**(5), 1561–1599 (2003).
7. E.J. Candès, D.L. Donoho, Ridgelets: a key to higher-dimensional intermittency? *Philos. Trans. R. Soc., Math. Phys. Eng. Sci.* **357**(1760), 2495–2509 (1999).
8. E.J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Commun. Pure Appl. Math.* **59**(8), 1207–1223 (2006).
9. E.J. Candès, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Stat.* **35**(6), 2313–2351 (2007).
10. A. Cohen, W. Dahmen, R.A. DeVore, Compressed sensing and best k-term approximation, *J. Am. Math. Soc.* **22**(1), 211–231 (2009).
11. A. Cohen, I. Daubechies, R.A. DeVore, G. Kerkyacharian, D. Picard, Capturing ridge functions in high dimensions from point queries, *Constr. Approx.* **35**(2), 225–243 (2012).
12. R. Courant, D. Hilbert, *Methods of Mathematical Physics, II* (Interscience, New York, 1962).
13. R.A. DeVore, G. Petrova, P. Wojtaszczyk, Instance optimality in probability with an ℓ_1-minimization decoder, *Appl. Comput. Harmon. Anal.* **27**(3), 275–288 (2009).
14. D.L. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006).
15. M. Fazel, Matrix rank minimization with applications, Ph.D. thesis, Stanford University, Palo Alto, CA, 2002.
16. M. Fornasier, Numerical methods for sparse recovery, in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, ed. by M. Fornasier, Radon Series on Computational and Applied Mathematics (De Gruyter, Berlin, 2010).
17. M. Fornasier, H. Rauhut, Compressive sensing, in *Handbook of Mathematical Methods in Imaging*, vol. 1, ed. by O. Scherzer (Springer, Berlin, 2010), pp. 187–229.
18. S. Foucart, A note on ensuring sparse recovery via ℓ_1-minimization, *Appl. Comput. Harmon. Anal.* **29**(1), 97–103 (2010).
19. G. Golub, C.F. van Loan, *Matrix Computations*, 3rd edn. (Johns Hopkins University Press, Baltimore, 1996).
20. F. John, *Plane Waves and Spherical Means Applied to Partial Differential Equations* (Interscience, New York, 1955).
21. M. Ledoux, *The Concentration of Measure Phenomenon* (American Mathematical Society, Providence, 2001).
22. B.F. Logan, L.A. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* **42**(4), 645–659 (1975).
23. E. Novak, H. Woźniakowski, *Tractability of Multivariate Problems, Volume I: Linear Information*, EMS Tracts in Mathematics, vol. 6 (Eur. Math. Soc., Zürich, 2008).
24. E. Novak, H. Woźniakowski, Approximation of infinitely differentiable multivariate functions is intractable, *J. Complex.* **25**, 398–404 (2009).
25. R.I. Oliveira, Sums of random Hermitian matrices and an inequality by Rudelson, *Electron. Commun. Probab.* **15**, 203–212 (2010).

26. S. Oymak, K. Mohan, M. Fazel, B. Hassibi, A simplified approach to recovery conditions for low-rank matrices, in *Proc. Intl. Symp. Information Theory (ISIT)* (2011).
27. A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numer.* **8**, 143–195 (1999).
28. B. Recht, M. Fazel, P. Parillo, Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization, *SIAM Rev.* **52**(3), 471–501 (2010).
29. M. Rudelson, R. Vershynin, Sampling from large matrices: an approach through geometric functional analysis, *J. ACM* **54**(4), 21 (2007) 19 pp.
30. W. Rudin, *Function Theory in the Unit Ball of* $\mathbb{C}^n$ (Springer, New York, 1980).
31. K. Schnass, J. Vybíral, Compressed learning of high-dimensional sparse functions, in *Proc. ICASSP11* (2011).
32. G.W. Stewart, Perturbation theory for the singular value decomposition, in *SVD and Signal Processing, II*, ed. by R.J. Vacarro (Amsterdam, Elsevier, 1991).
33. J.A. Tropp, User-friendly tail bounds for sums of random matrices, *Found. Comput. Math.* (2011). doi:10.1007/s10208-011-9099-z.
34. P.-A. Wedin, Perturbation bounds in connection with singular value decomposition, *BIT* **12**, 99–111 (1972).
35. H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), *Math. Ann.* **71**, 441–479 (1912).
36. P. Wojtaszczyk, ℓ_1 minimisation with noisy data, Preprint (2011).

Available online at www.sciencedirect.com

ScienceDirect

Journal of Approximation Theory 194 (2015) 35–61

Journal of Approximation Theory

www.elsevier.com/locate/jat

Full length article

On some aspects of approximation of ridge functions

Anton Kolleck[a], Jan Vybíral[b,*]

[a] *Mathematical Institute, Technical University Berlin, Strasse des 17. Juni 136, D-10623 Berlin, Germany*
[b] *Department of Mathematical Analysis, Charles University, Sokolovská 83, 186 00, Prague 8, Czech Republic*

Received 27 May 2014; received in revised form 2 January 2015; accepted 22 January 2015
Available online 7 February 2015

Communicated by Paul Nevai

Abstract

We present effective algorithms for uniform approximation of multivariate functions satisfying some prescribed inner structure. We extend, in several directions, the analysis of recovery of ridge functions $f(x) = g(\langle a, x\rangle)$ as performed earlier by one of the authors and his coauthors. We consider ridge functions defined on the unit cube $[-1, 1]^d$ as well as recovery of ridge functions defined on the unit ball from noisy measurements. We conclude with the study of functions of the type $f(x) = g(\|a - x\|_{l_2^d}^2)$.

MSC: 65D15; 41A25

Keywords: Ridge functions; High-dimensional function approximation; Noisy measurements; Compressed sensing; Dantzig selector

1. Introduction

Functions depending on a large number of variables play nowadays a crucial role in many areas, including parametric and stochastic PDEs, bioinformatics, financial mathematics, data analysis and learning theory. Together with an extensive computational power being used in

[*] Corresponding author.
E-mail addresses: kolleck@math.tu-berlin.de (A. Kolleck), vybiral@karlin.mff.cuni.cz (J. Vybíral).

http://dx.doi.org/10.1016/j.jat.2015.01.003
0021-9045/© 2015 Elsevier Inc. All rights reserved.

these applications, results on basic numerical aspects of these functions become crucial. Unfortunately, multivariate problems often suffer from the *curse of dimension*, i.e. the minimal number of operations necessary to achieve (an approximation of) a solution grows exponentially with the underlying dimension of the problem. Although this effect was observed many times in the literature, we refer to [27] for probably the most impressive result of this kind—namely that even the uniform approximation of infinitely-differentiable functions is intractable in high dimensions.

In the area of *Information-Based Complexity* it was possible to achieve a number of positive results on tractability of multivariate problems by imposing an additional (structural) assumption on the functions under study. The best studied concepts in this area include tensor product constructions and different concepts of anisotropy and weights. We refer to the series of monographs [26,28,29] for an extensive treatment of these and related problems. We pursue the direction initiated by Cohen, Daubechies, DeVore, Kerkyacharian and Picard in [11] and further developed in a series of recent papers [18,20,25]. This line of study is devoted to *ridge functions*, which are multivariate functions f taking the form $f(x) = g(\langle a, x\rangle)$ for some univariate function g and a non-zero vector $a \in \mathbb{R}^d$. We refer also to [15,32,33] for a related approach.

Functions of this type are by no means new in mathematics. They appear for example very often in statistics in the frame of the so-called *single index models*. They play also an important role in approximation theory, where their simple structure motivated the question if a general function could be well approximated by sums of ridge functions. The pioneering work in this field is [24], where the term “ridge function” was first introduced, and also [22], where the fundamentality of ridge functions was investigated. Ridge functions appeared also in mathematical analysis of neural networks [4,31] and as the basic building blocks of *ridgelets* of Candès and Donoho [6]. A survey on approximation by (sums of) ridge functions was given in [30].

The biggest difference between our setting and the usual approach of statistical learning and data analysis is that we suppose that the sampling points of f can be freely chosen, and are not given in advance. This happens, for instance, if sampling of the unknown function at a point is realized by a (costly) PDE solver.

Most of the techniques applied so far in recovery of ridge functions are based on the simple formula

$$\nabla f(x) = g'(\langle a, x\rangle) \cdot a. \tag{1.1}$$

One way how to use (1.1) is to approximate the gradient of f at a point with non-vanishing $g'(\langle a, x\rangle)$. By (1.1), it is then collinear with a. Once a is recovered, one can use any one-dimensional sampling method to approximate g.

Another way to approximate a is inspired by the technique of *compressed sensing* [8,16]. Taking directional derivatives of f at x results into

$$\frac{\partial f(x)}{\partial \varphi} = \langle \nabla f(x), \varphi\rangle = g'(\langle a, x\rangle)\langle a, \varphi\rangle,$$

i.e. it gives access to the scalar product of a with a chosen vector φ. If we assume that most of the coordinates of a are zero (or at least very small) and choose the directions $\varphi_1, \dots, \varphi_m$ at random, one can recover a effectively by the algorithms of sparse recovery.

Our aim is to fill some gaps left so far in the analysis done in [18]. Although the possibility of extending the analysis also to functions defined on other domains than the unit ball was mentioned already in [18], no steps in this direction were done there. We study in detail ridge functions defined on the unit cube $[-1, 1]^d$. The crucial component of our analysis is the use of the sign of a vector $\operatorname{sign}(x)$, which is defined componentwise. Although the mapping $x \to \operatorname{sign}(x)$

is obviously not continuous, the mapping (for $a \in \mathbb{R}^d$ fixed)

$$x \to \langle a, \operatorname{sign}(x)\rangle$$

is continuous at a (and takes the value $\|a\|_{l_1^d}$ there). This observation allows to imitate the approach of [18] and to adapt it to this setting. Let us remark, that all our approximation schemes recover first an approximation of the vector $a \in \mathbb{R}^d$. Afterwards, the problem becomes essentially one-dimensional and a good approximation of f by a limited number of sampling points can then be recovered by many classical methods, i.e. by spline approximation. We will therefore concentrate on an effective recovery of an approximation of a and the approximation of f will be given only implicitly.

Another topic only briefly discussed in [18] was the recovery of ridge functions from noisy measurements. Furthermore, our analysis as well as the approach of [18] or even the classical results of [3] are based on approximation of first (or higher) order derivatives by differences, which poses naturally the question on numerical stability of the presented algorithms. We present an algorithm based on the Dantzig selector of [9], which allows for recovery of a ridge function also in this setting. It turns out that in the case of a small step size $h > 0$, the first order differences cannot be evaluated with high enough precision. On the other hand, for a large step size h the first order differences do not approximate the first order derivatives well enough. Typically, there is therefore an $h > 0$, for which an optimal degree of approximation is achieved.

The next topic we discuss is the robustness of the methods developed. We show that (without much additional effort) it can be applied also for uniform recovery of translated radial functions $f(x) = g(\|a - x\|_{l_2^d}^2)$, which are constant along co-centered spheres instead of parallel hyperplanes. Similarly to the model of ridge functions, both the center $a \in \mathbb{R}^d$ and the univariate function g are unknown.

Finally, we close the paper with a number of numerical simulations of the algorithms presented. They highlight the surprising fact that their accuracy *improves* with increasing dimension. This is essentially based on the use of the *concentration of measure* phenomenon in the underlying theory and goes in line with similar observations made in the area of compressed sensing.

The paper is structured as follows. Section 2 collects some necessary notation and certain basic facts on sparse recovery from the area of *compressed sensing*. Section 3 extends the analysis of [18] to the setting of ridge functions defined on the unit cube. Section 4 treats the recovery of ridge functions defined on the unit ball from noisy measurements. Section 5 studies the translated radial functions $f(x) = g(\|a - x\|_{l_2^d}^2)$ and Section 6 closes with numerical examples.

2. Preliminaries

In this section we collect some notation and give an overview of results from the area of compressed sensing, which we shall need later on.

2.1. Notation

For a given vector $x \in \mathbb{R}^d$ and $0 \le p \le \infty$ we define

$$\|x\|_{l_p^d} := \begin{cases} \Big(\sum_{i=1}^d |x_i|^p\Big)^{\frac{1}{p}} & \text{if } 0 < p < \infty, \\ \#\{i \mid x_i \neq 0\} & \text{if } p = 0, \\ \max_{i=i,\dots,d} |x_i| & \text{if } p = \infty, \end{cases}$$

where $\#A$ denotes the cardinality of the set A.

This notation is further complemented by putting for $0 < p < \infty$

$$\|x\|_{l^d_{p,\infty}} := \max_{k=1,\dots,d} k^{\frac{1}{p}} x_{(k)},$$

where $x_{(k)}, k = 1, \dots, d$ denotes the non-increasing rearrangement of the absolute entries of x, i.e. $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(d)} \geq 0$ and $x_{(j)} = |x_{\sigma(j)}|$ for some permutation $\sigma : \{1, \dots, d\} \to \{1, \dots, d\}$ and all $j = 1, \dots, d$.

It is a very well known fact, that $\|\cdot\|_{\ell^d_p}$ is a norm for $1 \leq p \leq \infty$ and a quasi-norm if $0 < p < 1$. Also $\|\cdot\|_{\ell^d_{p,\infty}}$ is a quasi-norm for every $0 < p < \infty$. If $p = 2$, the space ℓ^d_2 is a Hilbert space with the usual inner product given by

$$\langle x, y\rangle = x^T y = \sum_{i=1}^{d} x_i y_i, \; x, y \in \mathbb{R}^d.$$

If $1 \leq s \leq d$ is a natural number, then a vector $x \in \mathbb{R}^d$ is called *s-sparse* if it contains at most s nonzero entries, i.e. $\|x\|_{l^d_0} \leq s$. The set of all s-sparse vectors is denoted by

$$\Sigma^d_s := \{x \in \mathbb{R}^d \mid \|x\|_{l^d_0} \leq s\}.$$

Finally, the best s-term approximation of a vector x describes how well x can be approximated by s-sparse vectors.

Definition 2.1. The *best s-term approximation* of a given vector $x \in \mathbb{R}^d$ with respect to the l^d_1-norm is given by

$$\sigma^d_s(x)_1 := \min_{z \in \Sigma^d_s} \|x - z\|_{l^d_1}.$$

2.2. *Results from compressed sensing*

Next we recall some basic concepts and results from compressed sensing which we will use later. Compressed sensing emerged in [8,7,16] as a method of recovery of sparse vectors x from a small set of linear measurements $y = \Phi x$. Since then, a vast literature on the subject appeared, concentrating on various aspects of the theory, and its applications. As it is not our aim to develop the theory of compressed sensing, but rather to use it in approximation theory, we shall restrict ourselves to the most important facts needed later on. We refer to [2,12,17,19] for recent overviews of the field and more references.

We focus on the recovery of vectors from noisy measurements, i.e. we want to recover the vectors $x \in \mathbb{R}^d$ from $m < d$ linear measurements of the form

$$y = \Phi x + e + z, \tag{2.1}$$

where $\Phi \in \mathbb{R}^{m \times d}$ is the measurement matrix and the noise is a composition of two factors, namely of the deterministic noise $e \in \mathbb{R}^m$ and the random noise $z \in \mathbb{R}^m$. Typically, we will assume that e is small (with respect to some ℓ^m_p norm) and that the components of z are generated independently according to a Gaussian distribution with small variance.

Obviously, some conditions have to be posed on Φ, so that the recovery of x from the measurements y given by (2.1) is possible. The most usual one in the theory of compressed sensing is that the matrix Φ satisfies the *restricted isometry property*.

Definition 2.2. The matrix $\Phi \in \mathbb{R}^{m\times d}$ satisfies the *restricted isometry property* (RIP) of order $s \le d$ if there exists a constant $0 < \delta < 1$ such that

$$(1-\delta)\|x\|_{l_2^d}^2 \le \|\Phi x\|_{l_2^m}^2 \le (1+\delta)\|x\|_{l_2^d}^2$$

holds for all s-sparse vectors $x \in \Sigma_s^d$. The smallest constant δ for which this inequality holds is called the restricted isometry constant and we will denote it by δ_s.

In general it is very hard to show that a given matrix satisfies this RIP or not. This is in particular the main reason why we will use random matrices, since it turns out that those matrices satisfy the RIP with overwhelming high probability. We present a version of such a statement, which comes from [1].

Theorem 2.3. *For every $0 < \delta < 1$ there exist constants $C_1, C_2 > 0$ depending on δ such that the random matrix $\Phi \in \mathbb{R}^{m\times d}$ with entries generated independently as*

$$\varphi_{ij} = \frac{1}{\sqrt{m}}\begin{cases} +1 & \textit{with probability } 1/2,\\ -1 & \textit{with probability } 1/2\end{cases} \tag{2.2}$$

satisfies the RIP of order s for each $s \le (C_2 m)/\log(d/m)$ with RIP constant $\delta_s \le \delta$ with probability at least

$$1 - 2e^{-C_1 m}.$$

Let us remark that log stands for the natural logarithm throughout the paper. A matrix Φ generated by (2.2) is called *normalized Bernoulli matrix*. For the sake of simplicity, we work with Bernoulli sensing matrices, but note that most of the statements presented below remain true for other classes of random matrices, c.f. [14, Section 5].

Next we present several recovery results for our starting problem (2.1). The first result of this kind deals with the case of exact measurements (i.e. $e = z = 0$) and uses the so called l_1^d-minimizer, cf. [10, Theorem 4.4] or [5, Theorem 1.2].

Theorem 2.4. *Let $\Phi \in \mathbb{R}^{m\times d}$ satisfy the RIP of order $2s$ with constant $\delta_{2s} \le \delta < 1/3$. Let $x \in \mathbb{R}^d$ and let us denote $y = \Phi x$. Finally, let $\Delta_{l_1^d}(y) \in \mathbb{R}^d$ be the solution of the minimization problem*

$$\min_{w\in\mathbb{R}^d} \|w\|_{l_1^d} \quad \textit{subject to } \Phi w = y. \tag{2.3}$$

Then it holds

$$\|x - \Delta_{l_1^d}(y)\|_{l_1^d} = \|x - \Delta_{l_1^d}(\Phi x)\|_{l_1^d} \le C_0 \sigma_s^d(x)_1$$

with constant C_0 depending only on δ.

This theorem implies that s-sparse vectors are recovered exactly by the l_1^d-minimizer (2.3) in the noise-free setting, since $\sigma_s^d(x)_1 = 0$ holds for every $x \in \Sigma_s^d$. To deal with the deterministic noise e, we shall need some more information about the geometrical properties of Bernoulli matrices. In particular, we will make use of Theorems 3.5 and 4.1 of [14], cf. also [23].

Theorem 2.5. *Let $\Phi \in \mathbb{R}^{m\times d}$ be a normalized Bernoulli matrix and let $d \geq (\log 6)^2 m$. Let $U_J = \{y \in \mathbb{R}^m : \|y\|_J \leq 1\}$, where*

$$\|y\|_J = \max\left\{\sqrt{m}\|y\|_{l_\infty^m}; \sqrt{\frac{m}{\log(d/m)}}\|y\|_{l_2^m}\right\}.$$

(i) *Then there exists an absolute constant $C_3 > 0$ such that with probability at least $1 - e^{-\sqrt{dm}}$ for every $y \in U_J$ there is an $x \in \mathbb{R}^d$, such that $\Phi x = y$ and $\|x\|_{l_1^d} \leq C_3$.*

(ii) *Let $\delta > 0$ and let C_1 and C_2 be the constants from Theorem 2.3. Then there exists an absolute constant C_3 and a constant C_4 depending on δ such that, with probability at least $1-2e^{-C_1 m}-e^{-\sqrt{md}}$, for each $y \in U_J$ there exists a vector $x \in \mathbb{R}^d$ with $\Phi x = y$, $\|x\|_{l_1^d} \leq C_3$ and $\|x\|_{l_2^d} \leq C_4\sqrt{\log(d/m)/m}$.*

We will use those two theorems to handle the deterministic noise e. Further we need a similar result to handle the random noise z, therefore we recall the *Dantzig selector* from [9].

Definition 2.6 (*Dantzig Selector*). For a matrix $\Phi \in \mathbb{R}^{m\times d}$ and constants $\lambda_d, \sigma > 0$ the *Dantzig selector* $\Delta_{DS}(y) \in \mathbb{R}^d$ of an input vector $y \in \mathbb{R}^m$ is defined as the solution of the minimization problem

$$\min_{w\in\mathbb{R}^d} \|w\|_{l_1^d} \quad \text{subject to } \|\Phi^T(y - \Phi w)\|_{l_\infty^d} \leq \lambda_d \sigma. \tag{2.4}$$

Remark 2.7. In what follows we shall use several parameters as the description of the typical frame of compressed sensing. First, we take $m \leq d$ to be natural numbers and denote by $\Phi \in \mathbb{R}^{m\times d}$ the normalized Bernoulli matrix (2.2). We put $\delta := 1/6$ and denote by C_1 and C_2 the constants appearing in Theorem 2.3. Next, we assume that the natural numbers $s \leq m \leq d$ satisfy

$$d \geq (\log 6)^2 m \quad \text{and} \quad 3s \leq (C_2 m)/\log(d/m). \tag{2.5}$$

Hence, by Theorem 2.3, Φ has (with high probability) the RIP of order $3s$ with a constant at most $1/6$.

Now we can use Theorem 1.3 of [9] to handle the random noise z.

Theorem 2.8. *Let s, m, d be natural numbers satisfying (2.5) and let $\Phi \in \mathbb{R}^{m\times d}$ be a normalized Bernoulli matrix. Let*

$$y = \Phi x + z$$

for $x \in \mathbb{R}^d$ with $\|x\|_{p,\infty} \leq R$, $0 < p \leq 1$, and $z \in \mathbb{R}^m$ with independent entries $z_i \sim \mathcal{N}(0, \sigma^2)$. Then there exists a constant C_5 such that the Dantzig selector (with $\lambda_d = \sqrt{2\log d}$) satisfies

$$\|\Delta_{DS}(y) - x\|_{l_2^d}^2 \leq \min_{1\leq s_*\leq s} 2C_5 \log d \left(s_*\sigma^2 + R^2 s_*^{-2(1/p-1/2)}\right)$$

with high probability.

Combining Theorems 2.5 and 2.8 we get the following new result.

Theorem 2.9. *Let s, m, d be natural numbers satisfying (2.5) and let $\Phi \in \mathbb{R}^{m\times d}$ be a normalized Bernoulli matrix. For $x \in \mathbb{R}^d$ and $e, z \in \mathbb{R}^m$ with $\|x\|_{l^d_{1,\infty}} \le R$, $\|e\|_{l^d_2} \le c\,\varepsilon\sqrt{\log(d/m)}$, $\|e\|_{l^d_\infty} \le c\,\varepsilon$ and $z_i \sim \mathcal{N}(0,\sigma^2)$ for some constants $R, \sigma, \varepsilon, c > 0$ let*

$$y = \Phi x + e + z.$$

Then there exist constants C_5, C_6, C_7 such that the Dantzig selector Δ_{DS} (with $\lambda_d = \sqrt{2\log d}$) applied to y satisfies the estimate

$$\|\Delta_{DS}(y) - x\|_{l^d_2} \le \left(\min_{1\le s_*\le s} 2C_5 \log d\left(s_*\sigma^2 + \tilde{R}^2 s_*^{-1}\right)\right)^{\frac{1}{2}} + C_7\frac{\varepsilon\sqrt{m}}{\sqrt{s}}$$

with high probability, where $\tilde{R} = 2(R + 2C_6\varepsilon\sqrt{m})$.

Proof. It follows from the assumptions that $\|e\|_J \le c\,\varepsilon\sqrt{m}$. Then we use Theorem 2.5 (ii) to find a vector $u \in \mathbb{R}^d$, such that

$$\begin{aligned}
&\Phi u = e,\\
&\|u\|_{l^d_1} \le C_3\|e\|_J \le C_3\, c\,\varepsilon\sqrt{m},\\
&\|u\|_{l^d_2} \le C_4\sqrt{\log(d/m)/m}\,\|e\|_J \le C_4\, c\,\varepsilon\sqrt{\log(d/m)}.
\end{aligned}$$

Further we apply the triangle inequality for the $\|\cdot\|_{1,\infty}$ quasinorm (see, for instance, Lemma 2.7 in [19]) to get

$$\begin{aligned}
\|x+u\|_{l^d_{1,\infty}} &\le 2\big(\|x\|_{l^d_{1,\infty}} + \|u\|_{l^d_{1,\infty}}\big) \le 2\big(\|x\|_{l^d_{1,\infty}} + \|u\|_{l^d_1}\big)\\
&\le 2\left(R + C_6\varepsilon\sqrt{m}\right) =: \tilde{R}.
\end{aligned}$$

Finally, applying Theorem 2.8 (with $p = 1$) we get

$$\begin{aligned}
\|\Delta_{DS}(y) - x\|_{l^d_2} &= \|\Delta_{DS}(\Phi x + e + z) - x\|_{l^d_2}\\
&\le \|\Delta_{DS}(\Phi(x+u) + z) - (x+u)\|_{l^d_2} + \|u\|_{l^d_2}\\
&\le \left(\min_{1\le s_*\le s} 2C_5\log d\left(s_*\sigma^2 + \tilde{R}^2 s_*^{-1}\right)\right)^{\frac{1}{2}} + C_7\frac{\varepsilon\sqrt{m}}{\sqrt{s}},
\end{aligned}$$

which finishes the proof. □

3. Approximation of ridge functions defined on cubes

Motivated by [11], we consider in this section uniform approximation of ridge functions of the form

$$f\colon [-1,1]^d \to \mathbb{R}, \qquad x \mapsto g(\langle a, x\rangle). \tag{3.1}$$

We assume that both the *ridge vector* $a \in \mathbb{R}^d$ and the univariate function g (also called *ridge profile*) are unknown.

First, we note that the problem is invariant with respect to scaling. Suppose that f is a ridge function with representation $f(x) = g(\langle a, x\rangle)$. Then for any scalar $\lambda \in \mathbb{R}\setminus\{0\}$ we put $\tilde{g}(x) := g(\frac{1}{\lambda}x)$ and $\tilde{a} := \lambda a$ to get another representation of f in the form of (3.1), namely

$$\tilde{g}(\langle\tilde{a}, x\rangle) = \tilde{g}(\langle\lambda a, x\rangle) = g\left(\left\langle\frac{1}{\lambda}\lambda a, x\right\rangle\right) = g(\langle a, x\rangle) = f(x).$$

Thus we can pose a scaling condition on a without any loss of generality. Furthermore, if $g'(0) \neq 0$, we can switch from a to $-a$, and obtain a ridge representation of f with $g'(0) > 0$.

In [18], the scaling condition $\|a\|_{l_2^d} = 1$ was assumed. This fitted together with both the scalar product structure used in the definition of f, as well as with the geometry of the domain of f used in [18], namely the Euclidean unit ball.

It is easy to observe that it will be more convenient for us to work with the ℓ_1^d-norm of a. Indeed, let us consider the case where the ridge profile $g(t) = t$ is known, i.e. that we have $f(x) = \langle a, x\rangle$ for some (unknown) $a \in \mathbb{R}^d$, and let us assume that we have an l_1^d-approximation $\hat{a}$ of a with $\|a - \hat{a}\|_{l_1^d} \leq \varepsilon$. Then Hölder's inequality gives us

$$\|\hat{f} - f\|_\infty = \sup_{x\in[-1,1]^d} |\langle a - \hat{a}, x\rangle| \leq \sup_{x\in[-1,1]^d} \|a - \hat{a}\|_{l_1^d}\|x\|_{l_\infty^d} \leq \varepsilon.$$

In what follows we shall therefore assume that

$$\|a\|_{l_1^d} = 1 \tag{3.2}$$

and that g is a univariate function defined on $I = [-1, 1]$. We further assume that g and g' are Lipschitz continuous with constants $c_0, c_1 > 0$, i.e.

$$|g(t_1) - g(t_2)| \leq c_0|t_1 - t_2|, \tag{3.3}$$

$$|g'(t_1) - g'(t_2)| \leq c_1|t_1 - t_2| \tag{3.4}$$

holds for all $t_1, t_2 \in I = [-1, 1]$. Finally, we assume that

$$g'(0) > 0 \tag{3.5}$$

as it is known, cf. [25], that approximation of ridge functions may be intractable if this condition is left out.

3.1. Approximation scheme without sparsity

In this part we evolve an approximation scheme for ridge functions with an arbitrary ridge vector $a \in \mathbb{R}^d$, merely assuming the right normalization (3.2). After this we consider the same problem with an additional sparsity condition on a, where we will use results from compressed sensing to reduce the number of samples.

Motivated by the formula (1.1) for $x = 0$

$$\nabla f(0) = g'(0)a, \tag{3.6}$$

we set for a small constant $h > 0$ and $i \in \{1, \ldots, d\}$

$$\tilde{a}_i := \frac{f(he_i) - f(0)}{h}, \tag{3.7}$$

where $e_1, \ldots, e_d$ are the usual canonical basis vectors of $\mathbb{R}^d$. As expected, it turns out that $\tilde{a}_i$ is a good approximation of $g'(0)a_i$ as the mean value theorem gives

$$\tilde{a}_i = \frac{f(he_i) - f(0)}{h} = \frac{g(h\langle a, e_i\rangle) - g(0)}{h} = g'(\xi_{h,i})a_i$$

for some $\xi_{h,i} \in (-|ha_i|, |ha_i|)$. And for the ℓ_1^d-approximation we obtain

$$\begin{aligned}\|\tilde{a} - g'(0)a\|_{l_1^d} &= \sum_{i=1}^d |\tilde{a}_i - g'(0)a_i| = \sum_{i=1}^d |g'(\xi_{h,i}) - g'(0)||a_i| \\ &\le \sum_{i=1}^d c_1|\xi_{h,i}||a_i| \le \sum_{i=1}^d c_1|ha_i||a_i| = c_1 h \sum_{i=1}^d |a_i|^2 \\ &\le c_1 h. \end{aligned} \tag{3.8}$$

Thus $\tilde{a}$ is a good approximation to $g'(0)a$ and since we want an approximation to a and we know that a is l_1^d-normalized we set

$$\hat{a} := \frac{\tilde{a}}{\|\tilde{a}\|_{l_1^d}}.$$

Now we have to estimate the difference between a and $\hat{a}$. We will use a variant of [18, Lemma 3.4].

Lemma 3.1. *Let $x \in \mathbb{R}^d$ with $\|x\|_{l_1^d} = 1$, $\tilde{x} \in \mathbb{R}^d \setminus \{0\}$ and $\lambda \in \mathbb{R}$. Then it holds*

$$\left\| \operatorname{sign}(\lambda)\frac{\tilde{x}}{\|\tilde{x}\|_{l_1^d}} - x \right\|_{l_1^d} \le \frac{2\|\tilde{x} - \lambda x\|_{l_1^d}}{\|\tilde{x}\|_{l_1^d}}.$$

Remark 3.2. We leave out the proof, which follows closely the proof of [18, Lemma 3.4]. The proof uses only triangle inequality and the lemma therefore remains true for any norm on $\mathbb{R}^d$.

Applying this lemma to our case it holds with (3.8) and the assumption (3.5)

$$\|\operatorname{sign}(g'(0))\hat{a} - a\|_{l_1^d} = \|\hat{a} - a\|_{l_1^d} \le \frac{2c_1 h}{\|\tilde{a}\|_{l_1^d}}. \tag{3.9}$$

Although we now know that $\hat{a}$ is a good approximation of a, it is still not clear how to define the uniform approximation $\hat{f}$ of f. The naive approach (used with success in [18] for ridge functions defined on the Euclidean unit ball) is to sample f along $\hat{a}$, i.e. to put $\hat{g}(t) := f(t\hat{a})$, and then define $\hat{f}(x) := \hat{g}(\langle \hat{a}, x\rangle)$. But when trying to estimate $\|f - \hat{f}\|_\infty$, we would need to ensure that $\langle \hat{a}, a\rangle$ is close to 1. This was indeed the case in [18], where an estimate on $\|\hat{a} - a\|_{\ell_2^d}$ was obtained, but it is not true any more in our setting of ℓ_1^d approximation.

On the other hand, because of the normalization of a, we have

$$\langle a, \operatorname{sign}(a)\rangle = \sum_{i=1}^d a_i \cdot \operatorname{sign}(a_i) = \sum_{i=1}^d |a_i| = \|a\|_{l_1^d} = 1,$$

where we defined the *sign* of a vector $x \in \mathbb{R}^d$ entrywise, i.e.

$$\operatorname{sign}(x) := (\operatorname{sign}(x_i))_i \in \mathbb{R}^d.$$

Note that this function is discontinuous, hence $\operatorname{sign}(a)$ and $\operatorname{sign}(\hat{a})$ can be far from each other, even if the difference $\|a - \hat{a}\|_{l_1^d}$ is small. Nevertheless their scalar product with a is nearly the same as Hölder's inequality gives

$$\begin{aligned}|\langle a, \operatorname{sign}(a) - \operatorname{sign}(\hat{a})\rangle| &= |\langle a, \operatorname{sign}(a)\rangle - \langle \hat{a}, \operatorname{sign}(\hat{a})\rangle - \langle a - \hat{a}, \operatorname{sign}(\hat{a})\rangle| \\ &\leq \|a - \hat{a}\|_{l_1^d} \|\operatorname{sign}(\hat{a})\|_{l_\infty^d} = \|a - \hat{a}\|_{l_1^d}. \end{aligned} \tag{3.10}$$

Thus we define

$$\hat{g}\colon [-1, 1] \to \mathbb{R},\ t \mapsto f\big(t \cdot \operatorname{sign}(\hat{a})\big) \tag{3.11}$$

and

$$\hat{f}(x) = \hat{g}(\langle \hat{a}, x\rangle). \tag{3.12}$$

Let us summarize our approximation algorithm as follows.

Algorithm A

- *Input:* Ridge function $f(x) = g(\langle a, x\rangle)$ with (3.2)-(3.5) and $h > 0$ small
- Put $\tilde{a}_i := \dfrac{f(he_i) - f(0)}{h}, i = 1, \ldots, d$
- Put $\hat{a} := \dfrac{\tilde{a}}{\|\tilde{a}\|_{l_1^d}}$
- Put $\hat{g}(t) = f(t \cdot \operatorname{sign}(\hat{a}))$ and $\hat{f}(x) = \hat{g}(\langle \hat{a}, x\rangle)$
- *Output:* $\hat{f}$

We formulate the approximation properties of Algorithm A as the following theorem.

Theorem 3.3. *Let $f\colon [-1, 1]^d \to \mathbb{R}$ be a ridge function with $f(x) = g(\langle a, x\rangle)$ for some $a \in \mathbb{R}^d$ with (3.2) and a differentiable function $g\colon [-1, 1] \to \mathbb{R}$ with (3.3)–(3.5). For $h > 0$ we construct the function $\hat{f}$ as described in Algorithm A. Then*

$$\|f - \hat{f}\|_\infty \leq 2c_0\|\hat{a} - a\|_{l_1^d} \leq \frac{4c_0c_1h}{g'(0) - c_1h}, \tag{3.13}$$

where the last inequality only holds if $g'(0) - c_1h$ is positive.

Proof. First, we show that

$$\|\hat{a} - a\|_{l_1^d} \leq \frac{2hc_1}{\|\tilde{a}\|_{l_1^d}} \leq \frac{2hc_1}{g'(0) - c_1h}, \tag{3.14}$$

where the last inequality only holds if $g'(0) - c_1h$ is positive.

Due to (3.9), we only have to show the last inequality of (3.14). With $\tilde{a}_i = g'(\xi_{h,i})a_i$ for some $\xi_{h,i} \in (-|ha_i|, |ha_i|) \subset [-h, h]$, it follows by the triangle inequality and (3.8)

$$\|\tilde{a}\|_{l_1^d} \geq \|g'(0)a\|_{l_1^d} - \|\tilde{a} - g'(0)a\|_{l_1^d} \geq g'(0) - c_1h,$$

which proves (3.14) and the second inequality in (3.13).

To prove the first inequality in (3.13), we use (3.3) and (3.10) to show that $\hat{g}$ is a good uniform approximation of g on $[-1, 1]$. We obtain

$$\begin{aligned}|g(t) - \hat{g}(t)| &= |g(t) - g(\langle a, t \cdot \operatorname{sign}(\hat{a})\rangle)| \leq c_0|t - t\langle a, \operatorname{sign}(\hat{a})\rangle| \\ &= c_0|t|\,\big|\langle a, \operatorname{sign}(a) - \operatorname{sign}(\hat{a})\rangle\big| \leq c_0\|a - \hat{a}\|_{l_1^d} \end{aligned} \tag{3.15}$$

for each $t \in [-1, 1]$. Finally, we combine this estimate with the definition of $\hat{f}$ as given in (3.12) and arrive at

$$\begin{aligned} |\hat{f}(x) - f(x)| &= |\hat{g}(\langle \hat{a}, x\rangle) - g(\langle a, x\rangle)| \\ &\le |\hat{g}(\langle \hat{a}, x\rangle) - g(\langle \hat{a}, x\rangle)| + |g(\langle \hat{a}, x\rangle) - g(\langle a, x\rangle)| \\ &\le c_0\|a - \hat{a}\|_{l_1^d} + c_0|\langle a - \hat{a}, x\rangle| \le 2c_0\|a - \hat{a}\|_{l_1^d}. \quad \Box \end{aligned} \tag{3.16}$$

Remark 3.4. (i) Algorithm A needs $d + 1$ function evaluations to identify $\hat{a}$. Once $\hat{a}$ was identified, an arbitrary one-dimensional sampling method can be used to identify $\hat{f}$ or, better said, its approximation. As there is a vast literature on that subject, we do not go into details and refer to [13] and the references in there.

(ii) The estimate (3.13) depends heavily on the value of $g'(0)$. Especially, the approximation becomes difficult, when this value gets smaller and (3.13) becomes void if $g'(0) = 0$. This is a very well known aspect of approximation of ridge functions, which was studied in great detail in [25]. We refer also to a slightly weaker condition used in [18].

(iii) If $\|a\|_{l_2^d}$ is small, the following improvement of (3.13) becomes of interest. First, we observe that (3.8) can be improved to $\|\tilde{a} - g'(0)a\|_{l_1^d} \le c_1 h\|a\|_{l_2^d}^2$, which results to

$$\|\hat{a} - a\|_{l_1^d} \le \frac{2c_1 h\|a\|_{l_2^d}^2}{\|\tilde{a}\|_{l_1^d}}.$$

Finally, this allows to improve (3.13) to

$$\|f - \hat{f}\|_\infty \le \frac{4c_0 c_1 h\|a\|_{l_2^d}^2}{g'(0) - c_1 h\|a\|_{l_2^d}^2}.$$

3.2. Approximation with sparsity

In this subsection we assume that the ridge vector $a \in \mathbb{R}^d$ is not only ℓ_1^d-normalized, but satisfies also some sparsity condition, i.e. most of the entries of a are zero or at least very small. We will use techniques of compressed sensing to address the approximation of the ridge vector a. Subsequently we obtain an approximation of f in the same way as before. The main advantage of Algorithm B presented below is that, for sparse vectors a, it achieves nearly the same error bound as Algorithm A, using a much smaller number of sampling points, cf. Remark 3.6.

Let $\Phi \in \mathbb{R}^{m\times d}$ be a normalized Bernoulli matrix and let $\varphi_1, \dots, \varphi_m$ be its rows. Taking their scalar product with the quantities in (3.6), we obtain

$$\frac{\partial f}{\partial \varphi_j}(0) = \langle \nabla f(0), \varphi_j\rangle = g'(0)\langle a, \varphi_j\rangle. \tag{3.17}$$

We use again first order differences as an approximation of the directional derivatives in (3.17), i.e. we set

$$\tilde{b}_j := \frac{f(h\varphi_j) - f(0)}{h}.$$

As in the previous section the mean value theorem gives the existence of some $\xi_{h,j}$ with $|\xi_{h,j}| \le |h| \cdot |\langle a, \varphi_j\rangle|$ such that

$$\tilde{b}_j = g'(\xi_{h,j})\langle a, \varphi_j\rangle.$$

In this sense, we expect $\tilde{b}_j$ to be a good approximation of $g'(0)\langle a, \varphi_j\rangle$ and $\tilde{b}$ to be a good approximation of $g'(0)\Phi a$. Hence, we recover $\tilde{a}$ through ℓ_1-minimization. From this point on, we may continue as before. Let us summarize this procedure as Algorithm B.

Algorithm B

- *Input:* Ridge function $f(x) = g(\langle a, x\rangle)$ with (3.2)–(3.5), $h > 0$ small and $m \le d/(\log 6)^2$
- Take $\Phi \in \mathbb{R}^{m\times d}$ a normalized Bernoulli matrix, cf. (2.2)
- Put $\tilde{b}_j := \dfrac{f(h\varphi_j) - f(0)}{h}, \quad j = 1, \dots, m$
- Put $\tilde{a} := \Delta_{l_1^d}(\tilde{b}) = \operatorname*{arg\,min}_{w\in\mathbb{R}^d} \|w\|_{l_1^d}$ s.t. $\Phi w = \tilde{b}$
- Put $\hat{a} := \dfrac{\tilde{a}}{\|\tilde{a}\|_{l_1^d}}$
- Put $\hat{g}(t) = f(t \cdot \operatorname{sign}(\hat{a}))$ and $\hat{f}(x) = \hat{g}(\langle \hat{a}, x\rangle)$
- *Output:* $\hat{f}$

Theorem 3.5. *Let $f\colon [-1, 1]^d \to \mathbb{R}$ be a ridge function with $f(x) = g(\langle a, x\rangle)$ for some vector $a \in \mathbb{R}^d$ with (3.2) and some differentiable function $g\colon [-1, 1] \to \mathbb{R}$ with (3.3)–(3.5). Let $d \ge (\log 6)^2 m$ and $h > 0$ be fixed. Then there exist some constants $C_0', C_1, C_2, c > 0$, such that for every positive integer s with $2s \le (C_2 m)/\log(d/m)$ the function $\hat{f}$ constructed in Algorithm B satisfies*

$$\|f - \hat{f}\|_\infty \le 2c_0\|\hat{a} - a\|_{l_1^d} \le 2c_0 \operatorname{err}(a, \hat{a}), \tag{3.18}$$

where

$$\operatorname{err}(a, \hat{a}) := C_0' \cdot \frac{g'(0)\cdot \sigma_s^d(a)_1 + ch}{g'(0)(1 - \sigma_s^d(a)_1) - 2ch},$$

with probability at least $1 - e^{-\sqrt{md}} - e^{-C_1 m}$ provided the denominator in the expression for $\operatorname{err}(a, \hat{a})$ is positive.

Proof. The first inequality in (3.18) follows again by (3.15) combined with (3.16).

Setting $\tilde{b} := (\tilde{b}_1, \dots, \tilde{b}_m)^T \in \mathbb{R}^m$ and $b := g'(0)\Phi a \in \mathbb{R}^m$ we get

$$\begin{aligned}
\|\tilde{b} - b\|_{l_1^d} &= \sum_{j=1}^m |g'(\xi_{h,j})\langle a, \varphi_j\rangle - g'(0)\langle a, \varphi_j\rangle| = \sum_{j=1}^m |g'(\xi_{h,j}) - g'(0)||\langle a, \varphi_j\rangle| \\
&\le \sum_{j=1}^m c_1 h|\langle a, \varphi_j\rangle|^2 \le c_1 h \sum_{j=1}^m \|a\|_{l_1^d}^2 \|\varphi_j\|_{l_\infty^d}^2 \\
&= c_1 h.
\end{aligned}$$

Therefore we obtain

$$\tilde{b} = b + e = g'(0)\Phi a + e \tag{3.19}$$

for $e \in \mathbb{R}^m$ with $\|e\|_{l_1^m} \le c_1 h$ and, similarly, $\|e\|_{l_\infty^m} \le c_1 h/m$ and $\|e\|_{l_2^m} \le c_1 h/\sqrt{m}$. This allows us to estimate the J-norm of e as follows

$$\begin{aligned}
\|e\|_J &= \max\left\{\sqrt{m}\|e\|_{l_\infty^m}\ ;\ \sqrt{\frac{m}{\log(d/m)}}\|e\|_{l_2^m}\right\}\\
&\le \max\left\{\frac{c_1 h}{\sqrt{m}}\ ;\ \frac{c_1 h}{\sqrt{\log(d/m)}}\right\}\\
&\le \max\left\{c_1 h;\ \frac{c_1 h}{\sqrt{2\log\log 6}}\right\}\\
&= c_1 h,
\end{aligned}$$

where we used $d \ge (\log 6)^2 m$ for the second inequality. Hence, by using Theorem 2.5 for $\tilde{e} = e/(c_1 h)$ there exists some vector $u \in \mathbb{R}^d$ with $\Phi u = e$ and $\|u\|_{\ell_1^d} \le C_3 c_1 h$.

Take some $1/3 > \delta > 0$ fixed, e.g. $\delta = 1/6$, and apply Theorem 2.4 to $g'(0)a + u$. This gives us

$$\begin{aligned}
\|\Delta_{l_1^d}(\tilde{b}) - g'(0)a\|_{l_1^d} &= \|\Delta_{l_1^d}\left(\Phi(g'(0)a + u)\right) - g'(0)a\|_{l_1^d}\\
&\le \|\Delta_{l_1^d}\left(\Phi(g'(0)a + u)\right) - g'(0)a - u\|_{l_1^d} + \|u\|_{l_1^d}\\
&\le C_0 \cdot \sigma_s^d\left(g'(0)a + u\right)_1 + \|u\|_{l_1^d}\\
&\le C_0 g'(0) \cdot \sigma_s^d(a)_1 + (1 + C_0)\|u\|_{l_1^d}\\
&\le (1 + C_0)\left(g'(0) \cdot \sigma_s^d(a)_1 + \|u\|_{l_1^d}\right).
\end{aligned}$$

Finally, by setting $\tilde{a} := \Delta_{l_1^d}(\tilde{b})$ and $\hat{a} := \tilde{a}/\|\tilde{a}\|_{l_1^d}$, Lemma 3.1 provides

$$\begin{aligned}
\|a - \hat{a}\|_{l_1^d} &\le 2(1 + C_0) \cdot \frac{g'(0) \cdot \sigma_s^d(a)_1 + \|u\|_{l_1^d}}{\|\tilde{a}\|_{l_1^d}}\\
&\le 2(1 + C_0) \cdot \frac{g'(0) \cdot \sigma_s^d(a)_1 + C_3 c_1 h}{\|\tilde{a}\|_{l_1^d}}.
\end{aligned} \tag{3.20}$$

From this point on we can proceed as in the proof of Theorem 3.3. We can again estimate the l_1^d-norm of $\tilde{a}$ from below. We get

$$\begin{aligned}
\|\tilde{a}\|_{l_1^d} &= \|\Delta_{l_1^d}(\Phi(g'(0)a + u))\|_{l_1^d}\\
&\ge \|g'(0)a + u\|_{l_1^d} - \|\Delta_{l_1^d}(\Phi(g'(0)a + u)) - g'(0)a - u\|_{l_1^d}\\
&\ge g'(0)\|a\|_{l_1^d} - \|u\|_{l_1^d} - \sigma_s^d(g'(0)a + u)_1\\
&\ge g'(0) - 2\|u\|_{l_1^d} - g'(0)\sigma_s^d(a)_1\\
&\ge g'(0) - 2C_3 c_1 h - g'(0)\sigma_s^d(a)_1.
\end{aligned} \tag{3.21}$$

Putting $c = C_3 c_1$ and $C_0' = 2(1 + C_0)$, we get the second inequality in (3.18). The first inequality in (3.18) is then again provided by (3.16). □

Remark 3.6. (i) In particular, if a is s-sparse, we get $\sigma_s^d(a)_1 = 0$ and, therefore,

$$\|f - \hat{f}\|_\infty \le C \frac{ch}{g'(0) - 2ch}.$$

(ii) If the sparsity level of a is $s \in \mathbb{N}$, the condition $2s \le (C_2 m)/\log(d/m)$ implies $m \ge 2s \log(d)/C_2$. Thus, in this case we need $m = O(s \log d)$ measurements to reconstruct the vector a.

4. Approximation of ridge functions with noisy measurements

In this section we study another aspect of recovery of ridge functions. We consider ridge functions defined on the unit ball as in [18] but we assume that the measurements are affected by random noise. This topic already appeared in a short remark in Section 4 of [20], but in contrary to that work, we additionally assume that the vector a satisfies a compressibility condition.

To be more precise, we consider ridge functions

$$f : B^d = \{x \in \mathbb{R}^d \mid \|x\|_{l_2^d} < 1\} \to \mathbb{R}, \ x \mapsto f(x) = g(\langle a, x \rangle).$$

We assume, that the ridge vector $a \in \mathbb{R}^d$ is l_2^d-normalized

$$\|a\|_{l_2^d} = 1 \tag{4.1}$$

and compressible in the following sense,

$$\|a\|_{l_1^d} \le R, \quad R > 0. \tag{4.2}$$

Furthermore, we assume that the ridge profile is a differentiable function $g : [-1, 1] \to \mathbb{R}$ with (3.3)–(3.5).

We shall use again the setting of Remark 2.7. Let $d \ge (\log 6)^2 m$ and let $\Phi \in \mathbb{R}^{m \times d}$ be a normalized Bernoulli matrix (2.2) with rows $\varphi_1, \ldots, \varphi_m$. By Theorem 2.3 it is ensured that Φ satisfies the RIP of order $2s$ with $0 < \delta_{2s} \le \delta := 1/6$ with high probability for every positive integer s with $3s \le (C_2 m)/\log(d/m)$, where the constant C_2 is the constant from Theorem 2.3.

But in contrary to (3.7), we now assume that the evaluation of f is perturbed by noise. To make the presentation technically simpler, we shall assume that the value $f(0)$ is given precisely (i.e. without noise). This can be achieved (with high precision) by re-sampling the value $f(0)$ several times, and applying Hoeffding's inequality.

Hence, we set for $j = 1, \ldots, m$ and a small constant $h > 0$

$$b_j := \frac{(f(h\varphi_j) + \tilde{z}_j) - f(0)}{h} = \frac{f(h\varphi_j) - f(0)}{h} + \frac{\tilde{z}_j}{h}.$$

We assume that the random noise $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_m)^T \in \mathbb{R}^m$ has independent components $\tilde{z}_j \sim \mathcal{N}(0, \sigma^2)$. Since $\tilde{z}_j$ are independent, it is well known that

$$z_j := \frac{\tilde{z}_j}{h} \sim \mathcal{N}\left(0, \frac{\sigma^2}{h^2}\right) \tag{4.3}$$

are also independent. As in the case with exact measurements the mean value theorem gives us

$$\frac{f(h\varphi_j)-f(0)}{h}=\frac{g(\langle a,h\varphi_j\rangle)-g(0)}{h}=g'(\xi_{h,j})\langle a,\varphi_j\rangle$$

for some real $\xi_{h,j}$ with $|\xi_{h,j}|\le|\langle a,h\varphi_j\rangle|$, hence

$$b_j=g'(\xi_{h,j})\langle a,\varphi_j\rangle+z_j.$$

To recover the vector a from these measurements let us first define the deterministic noise $e\in\mathbb{R}^m$ by

$$e_j=\langle a,\varphi_j\rangle(g'(\xi_{h,j})-g'(0)),\quad j=1,\dots,m, \tag{4.4}$$

i.e.

$$b=g'(0)\Phi a+e+z. \tag{4.5}$$

We then recover $\hat{a}$ with the help of the Dantzig selector (2.4) instead of l_1-minimization. Finally, for the construction of $\hat{g}$ and $\hat{f}$, we can use the direct approach of [18], which is given by

$$\hat{g}:\mathbb{R}\to\mathbb{R},\ t\mapsto f(t\hat{a})\quad\text{and}\quad\hat{f}:B^d\to\mathbb{R},\ x\mapsto\hat{g}(\langle\hat{a},x\rangle).$$

Let us summarize this procedure as the following algorithm.

Algorithm C

- *Input:* Ridge function $f(x)=g(\langle a,x\rangle)$ with (4.1), (4.2), (3.3)–(3.5), $h,\sigma>0$ and $m\le d/(\log 6)^2$
- Construct the $m\times d$ normalized Bernoulli matrix Φ, c.f. (2.2), with rows denoted by $\varphi_1,\dots,\varphi_m\in\mathbb{R}^d$
- Put $b_j=\dfrac{(f(h\varphi_j)+\tilde{z}_j)-f(0)}{h}$, $j=1,\dots,m$
- Put $\hat{a}=\dfrac{\Delta_{DS}(b)}{\|\Delta_{DS}(b)\|_{l_2^d}}$ for $\lambda_d=\sqrt{2\log d}$
- Put $\hat{g}:\mathbb{R}\to\mathbb{R},\ t\mapsto f(t\hat{a})$
- Put $\hat{f}:B^d\to\mathbb{R},\ x\mapsto\hat{g}(\langle\hat{a},x\rangle)$
- *Output:* $\hat{f}$

Theorem 4.1. *Let $f:B^d\to\mathbb{R}$ be a ridge function $f(x)=g(\langle a,x\rangle)$ with (4.1), (4.2), (3.3)–(3.5). Furthermore, let $h,\sigma>0$ and let $s\le m\le d$ be positive integers with (2.5). Let $\tilde{z}_j\sim\mathcal{N}(0,\sigma^2)$ be independent. Then there is a constant $C_2>0$, such that the function $\hat{f}$ defined by Algorithm* C *satisfies with high probability*

$$\|f-\hat{f}\|_\infty\le 2c_0\|a-\hat{a}\|_{l_2^d}\le\frac{4c_0\mathrm{err}(a,\hat{a})}{g'(0)-\mathrm{err}(a,\hat{a})}, \tag{4.6}$$

where

$$\mathrm{err}(a,\hat{a}):=\left(\min_{1\le s_*\le s}2C_5\log d\left(s_*\frac{\sigma^2}{h^2}+\tilde{R}^2s_*^{-1}\right)\right)^{\frac{1}{2}}+C_7\frac{h}{\sqrt{s}}, \tag{4.7}$$

$$\tilde{R}:=2(R+2C_6h)$$

for some constants C_5, C_6, C_7. The second inequality in (4.6) only holds if the denominator is positive.

Proof. To prove this theorem, we apply Theorem 2.9 to (4.5). Therefore we need to estimate the norm of $e \in \mathbb{R}^m$, defined by (4.4). We obtain

$$\begin{aligned}\|e\|_{l_2^m}^2 &= \sum_{j=1}^m \left[\langle a, \varphi_j\rangle (g'(\xi_{h,j}) - g'(0))\right]^2 \le \sum_{j=1}^m \left(c_1 h \langle a, \varphi_j\rangle^2\right)^2 \\ &\le c_1^2 h^2 \sum_{j=1}^m \left(\|a\|_{l_1^d} \|\varphi_j\|_{l_\infty^d}\right)^4 \le \frac{c_1^2 h^2 R^4}{m}\end{aligned}$$

and similarly we can show

$$\|e\|_{l_\infty^m} \le \frac{c_1 h R^2}{m}.$$

We can now apply Theorem 2.9 with $\varepsilon = hR^2/\sqrt{m}$ to get

$$\begin{aligned}\|\varDelta_{DS}(b) - g'(0)a\|_{l_2^d} &\le \left(\min_{1\le s_*\le s} 2C_5 \log d \left(s_* \frac{2\sigma^2}{h^2} + \tilde{R}^2 s_*^{-1}\right)\right)^{\frac{1}{2}} + C_7 \frac{h}{\sqrt{s}} \\ &=: \operatorname{err}(a, \hat{a})\end{aligned}$$

with $\tilde{R} = 2(R + 2C_6 h)$ and some constants C_5, C_6, C_7. And since we know that a is l_2^d-normalized we set

$$\hat{a} := \frac{\varDelta_{DS}(b)}{\|\varDelta_{DS}(b)\|_{l_2^d}}.$$

Applying Lemma 3.1 we get

$$\|a - \hat{a}\|_{l_2^d} \le \frac{2\operatorname{err}(a,\hat{a})}{\|\varDelta_{DS}(b)\|_{l_2^d}} \le \frac{2\operatorname{err}(a,\hat{a})}{g'(0)\|a\|_{l_2^m} - \operatorname{err}(a,\hat{a})} = \frac{2\operatorname{err}(a,\hat{a})}{g'(0) - \operatorname{err}(a,\hat{a})},$$

where the last inequality only holds if the denominator is positive. This proves the second inequality in (4.6).

The proof of the first part of (4.6) proceeds as in [18]. First we define an approximation $\hat{g}$ to g

$$\hat{g}: \mathbb{R} \to \mathbb{R},\ t \mapsto f(t\hat{a}). \tag{4.8}$$

This is indeed a good approximation to g as for any $t \in [-1, 1]$ we get

$$\begin{aligned}|\hat{g}(t) - g(t)| &= |g(\langle a, t\cdot\hat{a}\rangle) - g(t)| \le c_0 \left|t\left(1 - \langle a, \hat{a}\rangle\right)\right| = c_0|t| \cdot \left|\langle a, a - \hat{a}\rangle\right| \\ &\le c_0 \cdot \|a - \hat{a}\|_{l_2^d}.\end{aligned} \tag{4.9}$$

With this approximation $\hat{g}$ to g we define the function $\hat{f}$ by

$$\hat{f}: B^d \to \mathbb{R}, \qquad x \mapsto \hat{g}(\langle \hat{a}, x\rangle).$$

It remains to show that $\hat{f}$ is a good approximation to f. With the help of (4.8) and (4.9) we obtain

$$\begin{aligned}|f(x) - \hat{f}(x)| &= |g(\langle a, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| \\ &\le |g(\langle \hat{a}, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| + |g(\langle a, x\rangle) - g(\langle \hat{a}, x\rangle)|\end{aligned}$$

$$\begin{aligned}&\le c_0 \cdot \|a-\hat{a}\|_{l_2^d} + c_0|\langle a-\hat{a}, x\rangle|\\ &\le 2c_0\|a-\hat{a}\|_{l_2^d}\end{aligned}$$

for all $x \in B^d$. □

Remark 4.2. (i) We assumed that the step size $h > 0$ is an input of Algorithm C. From (4.7), we see that choosing h too small would amplify the noise, where choosing h large would lead to a worse approximation of derivatives by differences. We refer also to [20] for a brief discussion of this phenomenon. There is therefore an $h > 0$, which leads to an optimal compromise between the two parts of the right-hand side of (4.7). Unfortunately, this optimal h depends on the (a-priori unknown) function g and can be therefore hardly identified in the beginning. We exhibit in Section 6.2 also numerical evidence of this effect.

(ii) The main difference between our discussion of noisy sampling and the approach applied in [20] is that we assumed the ridge vector a to be compressible, cf. (4.2). Roughly speaking, this is reflected in the logarithmic dependence on d in (4.7) compared to (8) in [20].

(iii) We studied ridge functions on the unit ball in this section to be able to make use of the estimates on the Dantzig selector available in the literature. To adapt this approach to functions defined on the unit cube, it would be necessary to introduce the ℓ_1-version of Theorem 2.8 first.

5. Approximation of translated radial functions

The methods we presented so far, as well as the methods of [18], were developed in the (quite restrictive) frame of ridge functions. The aim of this section is to demonstrate that the same tools can be useful also in approximation of functions of different type. We consider the class of translated radial functions, i.e. functions of the form

$$f : B^d \to \mathbb{R},\ x \mapsto f(x) = g(\|a-x\|_{l_2^d}^2)$$

for some fixed l_2^d-normalized vector $a \in \mathbb{R}^d$

$$\|a\|_{l_2^d} = 1 \tag{5.1}$$

and a function $g : [0,4] \to \mathbb{R}$. Hence, f is constant on the spheres centered in a or, equivalently, it is a radial function translated by a. Typically, we shall again assume that g and g' are Lipschitz continuous with constants c_0 and c_1, respectively.

The idea to recover those functions is similar to the case of ridge functions. First we recover the center a and then we define approximations $\hat{g}$ to g and $\hat{f}$ to f.

For a small constant $h > 0$ and fixed vectors $x_i \in \mathbb{R}^d$, $i = 1, \dots, d$ we set

$$\tilde{a}_i := \frac{f(he_i + x_i) - f(x_i)}{h},$$

where $e_1, \dots, e_d$ are again the canonical basis vectors of $\mathbb{R}^d$. With help of the mean value theorem we can express this as

$$\begin{aligned}\tilde{a}_i &= \frac{f(he_i + x_i) - f(x_i)}{h} = \frac{g(\|a-he_i-x_i\|_{l_2^d}^2) - g(\|a-x_i\|_{l_2^d}^2)}{h}\\ &= g'(\xi_{h,i})\frac{\|a-he_i-x_i\|_{l_2^d}^2 - \|a-x_i\|_{l_2^d}^2}{h}\end{aligned}$$

for some real $\xi_{h,i}$ between $\|a-he_i-x_i\|_{l_2^d}^2$ and $\|a-x_i\|_{l_2^d}^2$. The numerator can be simplified by

$$\begin{aligned}\|a-he_i-x_i\|_{l_2^d}^2-\|a-x_i\|_{l_2^d}^2&=\langle a-he_i-x_i,a-he_i-x_i\rangle-\langle a-x_i,a-x_i\rangle\\&=-2h\langle a,e_i\rangle+h^2\langle e_i,e_i\rangle+2h\langle e_i,x_i\rangle\\&=-2ha_i+h^2+2hx_{i,i}.\end{aligned}$$

Let us choose $x_i=-(h/2)e_i$ to get

$$\begin{aligned}\tilde{a}_i&=\frac{f((h/2)e_i)-f(-(h/2)e_i)}{h}\\&=-2g'(\xi_{h,i})a_i\end{aligned}$$

for some $\xi_{h,i}$ between $\|a-(h/2)e_i\|_{l_2^d}^2$ and $\|a+(h/2)e_i\|_{l_2^d}^2$. Next let us note that $\xi_{h,i}$ is very close to $1=\|a\|_{l_2^d}^2$:

$$\begin{aligned}\left|\xi_{h,i}-1\right|&\le\max\left\{\left|1-\|a-(h/2)e_i\|_{l_2^d}^2\right|,\ \left|1-\|a+(h/2)e_i\|_{l_2^d}^2\right|\right\}\\&=\max\left\{\left|-ha_i-h^2/4\right|,\ \left|ha_i-h^2/4\right|\right\}\\&\le h+h^2/4.\end{aligned}\tag{5.2}$$

Finally we obtain that $\hat{a}$ is a good approximation to $-2g'(\|a\|_{l_2^d}^2)a=-2g'(1)a$, since

$$\begin{aligned}\|\tilde{a}+2g'(1)a\|_{l_2^d}^2&=\sum_{i=1}^d\left(-2g'(\xi_{h,i})a_i+2g'(1)a_i\right)^2\\&=4\sum_{i=1}^d\left(\left(g'(\xi_{h,i})-g'(1)\right)a_i\right)^2\le4\sum_{i=1}^d\left(c_1\left|\xi_{h,i}-1\right|a_i\right)^2\\&\le4c_1^2\sum_{i=1}^d\left(\left(h+h^2/4\right)a_i\right)^2=4c_1^2\left(h+h^2/4\right)^2\sum_{i=1}^da_i^2\\&=4c_1^2\left(h+h^2/4\right)^2.\end{aligned}$$

Thus $\tilde{a}$ is almost a multiple of a. Again, we need to assume that the derivative of g' is non-trivial in some sense. Due to the construction, we replace (3.5) by the condition

$$g'(1)\neq0.$$

Then the l_2^d-normalized vector

$$\hat{a}:=\frac{\tilde{a}}{\|\tilde{a}\|_{l_2^d}}$$

approximates a, possibly up to a sign. Choosing any vector $\hat{a}^\perp\in\mathbb{R}^d$ orthogonal to $\hat{a}$, we can identify the sign by sampling along $\hat{a}^\perp$. Afterwards, the correct sign might be assigned to $\hat{a}$. We will therefore restrict ourselves to the case

$$g'(1)>0.\tag{5.3}$$

Once an approximation of a was recovered, it is again easy to define an approximation of g, and finally of f. We summarize this procedure as the following algorithm.

Algorithm D

- *Input:* Translated radial function $f\colon B^d \to \mathbb{R}$ with $f(x) = g(\|a-x\|_{l_2^d}^2)$, a and g with (5.1), (3.3), (3.4) and (5.3), $h > 0$
- Put $\tilde{a}_i := \frac{f(he_i/2)-f(-he_i/2)}{h}$, $i = 1, \dots, d$
- Put $\hat{a} := \frac{\tilde{a}}{\|\tilde{a}\|_{l_2^d}}$
- Put $\hat{g}\colon [0,4] \to \mathbb{R},\ t \mapsto f(\hat{a}(1-\sqrt{t}))$
- Put $\hat{f}\colon B^d \to \mathbb{R},\ x \mapsto \hat{g}(\|\hat{a}-x\|_{l_2^d}^2)$
- *Output:* $\hat{f}$

The performance of Algorithm D is estimated by the following theorem.

Theorem 5.1. *Let $f\colon B^d \to \mathbb{R}$, $g\colon [0,4] \to \mathbb{R}$ and $a \in \mathbb{R}^d$ be such that $f(x) = g(\|a-x\|_{l_2^d}^2)$ and a and g satisfy (5.1), (3.3), (3.4) and (5.3). Then*

$$\|f-\hat{f}\|_\infty \le c_0\left(2\|\hat{a}-a\|_{l_2^d} + \|\hat{a}-a\|_{l_2^d}^2\right) \tag{5.4}$$

and

$$\|\hat{a}-a\|_{l_2^d} \le \frac{2c_1\left(h+h^2/4\right)}{g'(1)-c_1(h+h^2/4)} \tag{5.5}$$

if $g'(1)-c_1(h+h^2/4)$ is positive.

Proof. First, we estimate the difference between a and $\hat{a}$. By (5.2) and (3.4)

$$g'(1) - |g'(\xi_{h,i})| \le |g'(1)-g'(\xi_{h,i})| \le c_1|1-\xi_{h,i}| \le c_1(h+h^2/4),$$

hence

$$|g'(\xi_{h,i})| \ge g'(1) - c_1(h+h^2/4). \tag{5.6}$$

Therefore, if the right hand side of (5.6) is positive, we get

$$\begin{aligned}\|\tilde{a}\|_{l_2^d}^2 &= \sum_{i=1}^d |\tilde{a}_i|^2 = 4\sum_{i=1}^d |g'(\xi_{h,i})a_i|^2\\ &\ge 4\sum_{i=1}^d \left(g'(1)-c_1(h+h^2/4)\right)^2 a_i^2\\ &= 4\left(g'(1)-c_1(h+h^2/4)\right)^2.\end{aligned}$$

Now we apply Lemma 3.1 to obtain

$$\|\hat{a}-a\|_{l_2^d} \le \frac{4c_1(h+h^2/4)}{\|\tilde{a}\|_{l_2^d}} \le \frac{2c_1(h+h^2/4)}{g'(1)-c_1(h+h^2/4)}. \tag{5.7}$$

Given the approximation $\hat{a}$ to a we define an approximation $\hat{g}$ to g by

$$\hat{g}\colon [0,4] \to \mathbb{R}, \qquad t \mapsto f\left(\hat{a}\left(1-\sqrt{t}\right)\right).$$

Essentially, $\hat{g}$ is the restriction of f onto $\{\lambda\hat{a} : \lambda \in \mathbb{R}\} \cap B^d$. Using (3.3) we obtain the estimate

$$\begin{aligned} |g(t)-\hat{g}(t)| &= \left|g(t)-g(\|a-\hat{a}+\sqrt{t}\hat{a}\|_{l_2^d}^2)\right| \le c_0 \left|t-\|a-\hat{a}+\sqrt{t}\hat{a}\|_{l_2^d}^2\right| \\ &= c_0 \left|2\sqrt{t}\langle a-\hat{a},\hat{a}\rangle + \|a-\hat{a}\|_{l_2^d}^2\right| = c_0 \left|2\left(1-\sqrt{t}\right)(1-\langle a,\hat{a}\rangle)\right| \\ &= c_0 \left|1-\sqrt{t}\right| \cdot \|a-\hat{a}\|_{l_2^d}^2 \le c_0 \|a-\hat{a}\|_{l_2^d}^2 \end{aligned} \tag{5.8}$$

for all $t \in [0,4]$. Next we define

$$\hat{f}\colon B^d \to \mathbb{R},\ x \mapsto \hat{g}(\|\hat{a}-x\|_{l_2^d}^2).$$

With (5.7) and (5.8) we get the final estimate

$$\begin{aligned} |f(x)-\hat{f}(x)| &= \left|g(\|a-x\|_{l_2^d}^2) - \hat{g}(\|\hat{a}-x\|_{l_2^d}^2)\right| \\ &\le \left|g(\|a-x\|_{l_2^d}^2) - g(\|\hat{a}-x\|_{l_2^d}^2)\right| + \left|g(\|\hat{a}-x\|_{l_2^d}^2) - \hat{g}(\|\hat{a}-x\|_{l_2^d}^2)\right| \\ &\le c_0 \left|\|a-x\|_{l_2^d}^2 - \|\hat{a}-x\|_{l_2^d}^2\right| + c_0 \|a-\hat{a}\|_{l_2^d}^2 \\ &= 2c_0 \left|\langle a-\hat{a},x\rangle\right| + c_0 \|a-\hat{a}\|_{l_2^d}^2 \\ &\le c_0 \left(2\|a-\hat{a}\|_{l_2^d} + \|a-\hat{a}\|_{l_2^d}^2\right). \quad \square \end{aligned}$$

Remark 5.2. We assumed in Theorem 5.1, that the function g and its derivative g' are both Lipschitz. If we assume this property only on the interval $(1-(h+h^2/4), 1+(h+h^2/4))$, we can still recover at least (5.7). This applies even to the case, when g (and also its derivative) are unbounded near the origin. In that case, we can still approximate the position of the singularity, although uniform approximation of f is out of reach.

5.1. Extensions of Theorem 5.1

As in the approximation scheme for ridge functions we can use techniques from compressed sensing to recover f if a is compressible. To be more precise, if a satisfies

$$\|a\|_{l_1^d} \le R \tag{5.9}$$

and $\Phi \in \mathbb{R}^{m\times d}$ is a normalized Bernoulli matrix with rows $\varphi_1,\dots,\varphi_m \in \mathbb{R}^d$, we define

$$\tilde{b}_j := \frac{f((h/2)\varphi_j) - f(-(h/2)\varphi_j)}{h}, \qquad j = 1,\dots,m.$$

As f is defined only on the unit ball B^d and $\|\varphi_j\|_{l_2^d} = \sqrt{d/m}$, we must always have at least $h \le 2\sqrt{m/d}$ to ensure that $(h/2)\varphi_j \in B^d$. To allow for comparison with the non-compressible case just discussed in Theorem 5.1, we denote

$$\tilde{h} = h/2 \cdot \sqrt{d/m},$$

which leads to

$$\tilde{b}_j = \frac{f\left(\tilde{h}\frac{\varphi_j}{\|\varphi_j\|_{l_2^d}}\right) - f\left(-\tilde{h}\frac{\varphi_j}{\|\varphi_j\|_{l_2^d}}\right)}{h}. \tag{5.10}$$

By defining the deterministic noise $e \in \mathbb{R}^m$

$$\tilde{b} = -2g'(1)\Phi a + e \tag{5.11}$$

we can show with similar calculations as before that

$$\|e\|_{l_2^m} \leq \eta := 2c_1 R\left(\frac{2R\tilde{h}}{\sqrt{d}} + \tilde{h}^2\right). \tag{5.12}$$

Using the $(P_{1,\eta})$ minimizer of [5] we put

$$\tilde{a} = \arg\min_{z\in\mathbb{R}^d} \|z\|_{l_1^d} \quad \text{s.t. } \|\Phi z - \tilde{b}\|_{l_2^m} \leq \eta$$

with η given by the right hand side of (5.12). We then get the estimate, cf. [19, Theorem 4.22] or [2, Theorem 1.6],

$$\|\tilde{a} - 2g'(1)a\|_{l_2^d} \leq \varrho := \frac{C\sigma_s^d(2g'(1)a)_1}{\sqrt{s}} + D\eta$$

with two universal constants $C, D > 0$. Here again $s \leq C_2 m/\log(d/m)$. Lemma 3.1 (with l_2^d instead of l_1^d) gives for $\hat{a} = \tilde{a}/\|\tilde{a}\|_{l_2^d}$

$$\|\hat{a} - a\|_{l_2^d} \leq 2\varrho/\|\tilde{a}\|_{l_2^d}.$$

Finally, using

$$\|\tilde{a}\|_{l_2^d} \geq 2g'(1)\|a\|_{l_2^d} - \|\tilde{a} - 2g'(1)a\|_{l_2^d} \geq 2g'(1) - \varrho,$$

we get

$$\|\hat{a} - a\|_{l_2^d} \leq \frac{2\varrho}{2g'(1) - \varrho}$$

if $2g'(1) > \varrho$.

This gives a replacement of (5.7), the rest of the proof of Theorem 5.1 then applies without further modifications. We summarize this procedure as Algorithm E below and describe its performance.

Algorithm E

- *Input:* Translated radial function $f\colon B^d \to \mathbb{R}$ with $f(x) = g(\|a - x\|^2_{l_2^d})$, a and g with (5.1), (3.3), (3.4), (5.3) and (5.9), $\tilde{h} > 0$, $m \le d/(\log 6)^2$
- Take $\Phi \in \mathbb{R}^{m\times d}$ a normalized Bernoulli matrix, cf. (2.2)
- Put $\tilde{b}_j = \dfrac{\|\varphi_j\|_{\ell_2^d}}{2\tilde{h}} \cdot \left[f\left(\tilde{h}\dfrac{\varphi_j}{\|\varphi_j\|_{l_2^d}}\right) - f\left(-\tilde{h}\dfrac{\varphi_j}{\|\varphi_j\|_{l_2^d}}\right)\right]$
- Put $\eta := 2c_1 R\left(\frac{2R\tilde{h}}{\sqrt{d}} + \tilde{h}^2\right)$
- Put $\tilde{a} := \arg\min_{z\in\mathbb{R}^d} \|z\|_{l_1^d} \quad \text{s.t.} \quad \|\Phi z - \tilde{b}\|_{l_2^m} \le \eta$
- Put $\hat{a} := \dfrac{\tilde{a}}{\|\tilde{a}\|_{l_2^d}}$
- Put $\hat{g}\colon [0,4] \to \mathbb{R},\ t \mapsto f(\hat{a}(1-\sqrt{t}))$
- Put $\hat{f}\colon B^d \to \mathbb{R},\ x \mapsto \hat{g}(\|\hat{a} - x\|^2_{l_2^d})$
- *Output:* $\hat{f}$

Theorem 5.3. *Let $f\colon B^d \to \mathbb{R}$, $g\colon [0,4] \to \mathbb{R}$ and $a \in \mathbb{R}^d$ be such that $f(x) = g(\|a - x\|^2_{l_2^d})$ and a and g satisfy (5.1), (3.3), (3.4), (5.3) and (5.9). Let $s \le m \le d$ be positive integers satisfying (2.5). Then*

$$\|f - \hat{f}\|_\infty \le c_0\left(2\|\hat{a} - a\|_{l_2^d} + \|\hat{a} - a\|^2_{l_2^d}\right) \tag{5.13}$$

and

$$\|\hat{a} - a\|_{l_2^d} \le \frac{2\varrho}{2g'(1) - \varrho} \tag{5.14}$$

if $2g'(1) > \varrho$. Here

$$\varrho := \frac{C\sigma_s^d(2g'(1)a)_1}{\sqrt{s}} + D\eta$$

for two universal constants $C, D > 0$.

Remark 5.4. Once we have this approximation scheme using techniques from compressed sensing, we can easily extend it to an approximation scheme with noisy measurements. We assume again that $\tilde{b}$ from (5.10) is corrupted by noise z/h, where the components of $z = (z_1, \ldots, z_m)^T$ are again independent $\mathcal{N}(0, \sigma^2)$ distributed random variables. Formula (5.11) is then replaced by $\tilde{b} = -2g'(1)\Phi a + e + z/h$ and the Dantzig selector can be applied.

6. Numerical results

In this section we investigate the performance of the algorithms presented so far in several model situations. The results shed a new light on some of the aspects, which we did not discuss in detail, especially on the size of the constants used in the previous theorems. All the approximation schemes started by looking for a good approximation $\hat{a}$ of the unknown direction a and, consequently, the quality of the uniform approximation of f by $\hat{f}$ was then bounded by the

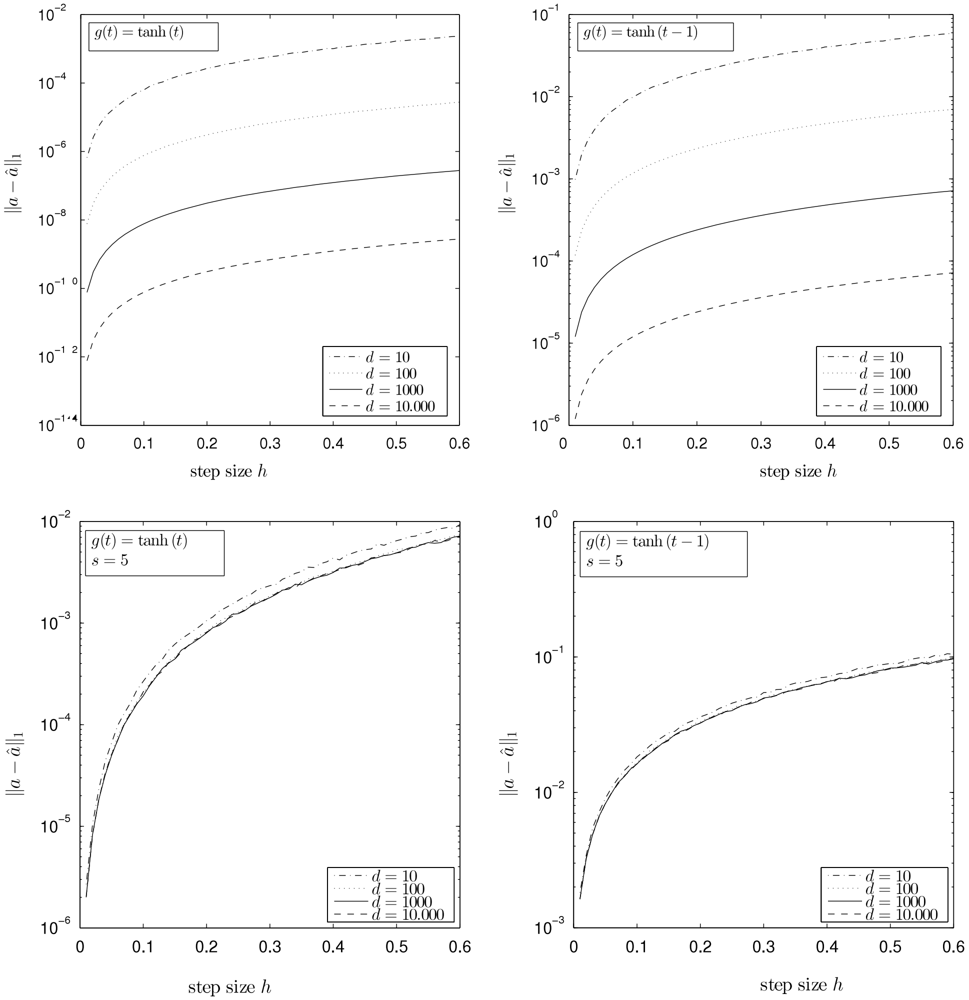


Fig. 1. Approximation of a non-sparse profile a according to Algorithm A with $g(t) = \tanh(t)$ (left top) and $g(t) = \tanh(t - 1)$ (right top). Approximation of sparse profiles a with $s = 5$ by Algorithm A: left bottom with $g(t) = \tanh(t)$ and right bottom with $g(t) = \tanh(t - 1)$.

corresponding distance between $\hat{a}$ and a. In what follows, we will therefore discuss only the approximation error between a and $\hat{a}$.

6.1. *Ridge functions on cubes*

We start with Algorithm A, i.e. with approximation of a ridge function $f(x) = g(\langle a, x\rangle)$ defined on the cube $[-1, 1]^d$ with $\|a\|_{\ell_1^d} = 1$. We have considered different dimensions ($d \in \{10, 100, 1000, 10.000\}$). As Algorithm A does not make any use of sparsity of a, it is reasonable to assume, that all its coordinates are equally likely to be non-zero. The entries of a were therefore always independently normally distributed (i.e. $a_i \sim \mathcal{N}(0, 1)$), afterwards a got l_1^d-normalized according to (3.2). For comparison, the bottom line of Fig. 1 displays also the performance of Algorithm A on sparse vectors.

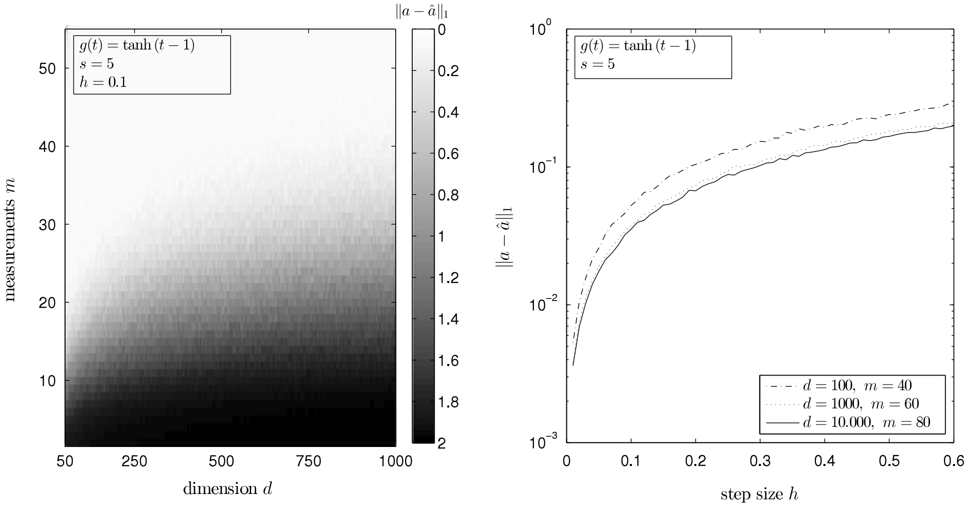


Fig. 2. Phase transition for the approximation of a and the average error of $\|a-\hat{a}\|_{\ell_1^d}$ according to Algorithm B.

Fig. 1 shows the approximation error $\|a-\hat{a}\|_{\ell_1^d}$ in dependence of the step size $h > 0$ for two different profiles $g(t) = \tanh(t)$ and $g(t) = \tanh(t-1)$. Note that the y-axis scales logarithmically. For each step size, we repeated the calculation thousand times and took the average afterwards. Let us give some remarks on Fig. 1.

- The approximation improves rapidly with growing dimension. This is a consequence of considering random non-sparse ridge vectors a and of (3.8), cf. also Remark 3.4. By the concentration of measure phenomenon [21], the quantity $\|a\|_{\ell_2^d}/\|a\|_{\ell_1^d}$ behaves as $d^{-1/2}$ for typical random vectors $a \in \mathbb{R}^d$. When considering sparse vectors as inputs in Algorithm A (Fig. 1, bottom line), this effect disappears and there is no improvement with increasing dimension.
- Smaller step size h implies also better quality of approximation, but already reasonable sizes of h (i.e. $h = 0.2$) imply surprisingly small errors.
- Finally, the second derivative of the first profile at zero vanishes, were it is non-zero for the second profile. Therefore, the first order differences approximate the first order derivative less accurately in that case, leading to larger (but still surprisingly small) approximation errors.

The left part of Fig. 2 shows the dependence of the number of the sampling points m on the dimension d and sparsity s, cf. (2.5), when using Algorithm B. We fixed the ridge profile $g(t) = \tanh(t-1)$, the sparsity $s = 5$ and the step size $h = 0.1$ and constructed an s-sparse random vector a by MATLAB command `sprandn`, followed by the ℓ_1^d-normalization. For each integer d between 50 and 1000 and for each integer m between 1 and 55, we then run Algorithm B 120 times and the average approximation error $\|a-\hat{a}\|_{\ell_1^d}$ corresponds afterwards to the shade of gray of the point with coordinates d and m. In accordance with the theory of compressed sensing (and with Remark 3.6), we observe that for a random matrix Φ and a random sparse vector a the number of measurements needs to grow only logarithmically in the dimension d to guarantee good approximation with high probability. The right part of Fig. 2 then shows the average value of $\|a-\hat{a}\|_{\ell_1^d}$ over one hundred repetitions for the same profile and sparsity for three different pairs of (d, m). We observe, that especially for large dimensions even extremely small number of measurements guarantees already reasonable approximation errors.

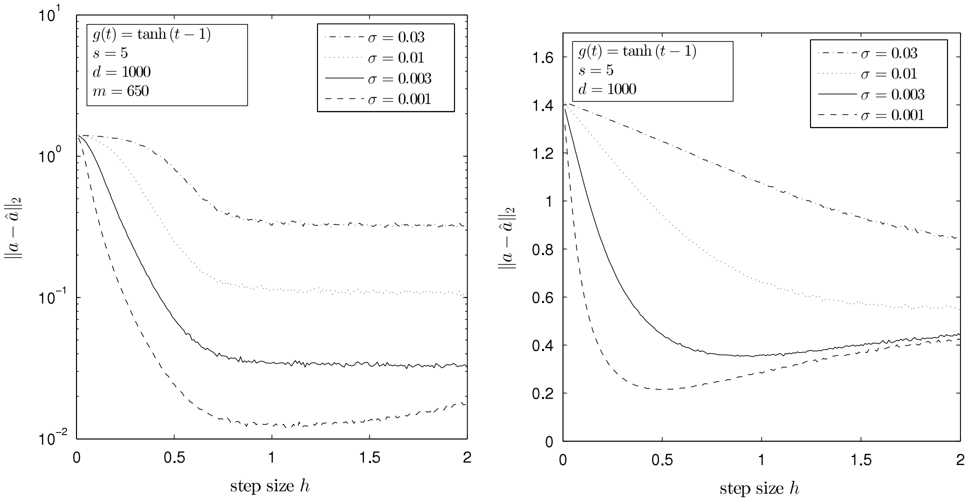


Fig. 3. Approximation of a with noisy measurements according to Algorithm C (left) and a modification of Algorithm A (right). Note, that only the y-axis of the left plot is logarithmic.

6.2. Noisy measurements

Fig. 3 studies the performance of the recovery of the ridge vector a from noisy measurements as described in Algorithm C. We fixed the parameters $d = 1000, m = 400$, and $s = 5$, the ridge profile $g(t) = \tanh(t-1)$ and four different noise levels $\sigma \in \{0.03, 0.01, 0.003, 0.001\}$. The number of repetitions for each step size was set to 1000. We have used the ℓ_1-MAGIC implementation of Dantzig selector, available at the web page of Justin Romberg at http://users.ece.gatech.edu/~justin/l1magic/. As the noise level gets amplified by the factor $1/h$, when taking the first order differences, cf. (4.3), it is not surprising that the recovery fails completely for small values of h. On the other hand, for large values of h, the correspondence between first order differences and first order derivatives gets weaker and the quality of approximation deteriorates as well. This effect is clearly visible from (4.7) and, numerically, in the left part of Fig. 3, where there is an optimal h for the recovery of a. Strictly speaking, the functions considered in Section 4 were defined only on the unit ball B^d, so that the value of h in Fig. 3 should be limited to be smaller than $\sqrt{m/d}$. We have decided to include also larger values h to exhibit the optimal h, although for our profile and our parameters it lies outside of this interval.

Although not discussed before, it is quite straightforward to modify the non-probabilistic Algorithm A also to the case of noisy measurements. Essentially, the gradient $\nabla f(0)$ is then approximated by the first-order differences, this time corrupted by noise. We applied this approach to the profile $g(t) = \tanh(t-1)$ and parameters just described with the results plotted in the right part of Fig. 3. We observe that the approximation errors get much larger, demonstrating once again the success of the Dantzig selector.

6.3. Shifted radial functions

In Fig. 4 we considered the approximation of the pole a of a shifted radial function f with $f(x) = g(\|a - x\|^2_{l_2^d})$ and $g(t) = -1/t$. On the left plot, we fixed the sparsity $s = 5$ and

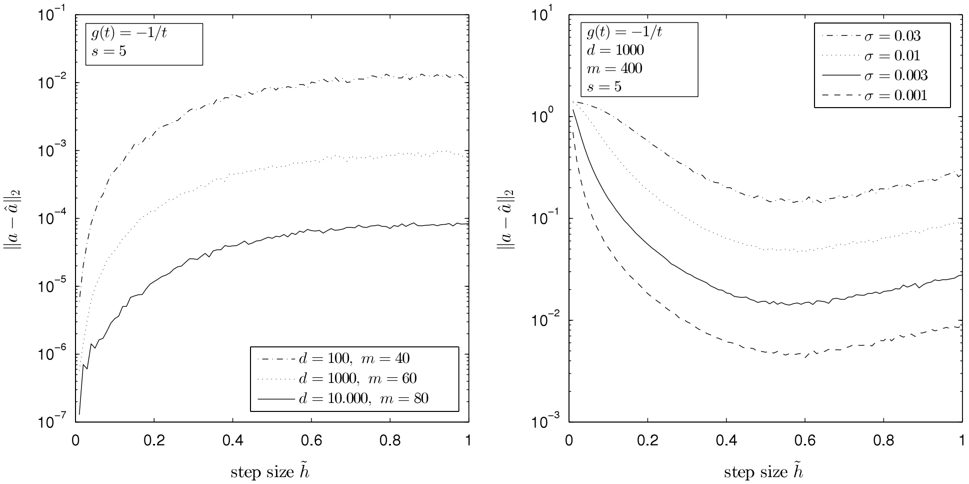


Fig. 4. Approximation of a according to Algorithm E with sparsity (left) and with noisy measurements (right).

considered three values of $d = 100, d = 1000$ and $d = 10.000$. The number of measurements was then $m = 40, m = 60$, or $m = 80$, respectively. The number of repetitions for each step size was set to 40 in the left part of Fig. 4 and to one hundred in the right part. Finally, we run Algorithm E and plot the average approximation error of $\|a - \hat{a}\|_{\ell_2^d}$ against the step size h. The right hand plot of Fig. 4 shows the noise-aware modification of Algorithm E described in Remark 5.4.

Acknowledgments

The first author acknowledges the support by the DFG Research Center MATHEON “Mathematics for key technologies” in Berlin. The last author was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices, Constr. Approx. 28 (2008) 253–263.
[2] H. Boche, R. Calderbank, G. Kutyniok, J. Vybíral, A survey of compressed sensing, in: Applied and Numerical Harmonic Analysis, Birkhäuser, Boston (in press).
[3] M.D. Buhmann, A. Pinkus, Identifying linear combinations of ridge functions, Adv. Appl. Math. 22 (1999) 103–118.
[4] E. Candès, Harmonic analysis of neural networks, Appl. Comput. Harmon. Anal. 6 (1999) 197–218.
[5] E. Candès, The restricted isometry property and its implications for compressed sensing, C. R. Acad. Sci., Paris I 346 (2008) 589–592.
[6] E. Candès, D.L. Donoho, Ridgelets: a key to higher-dimensional intermittency? Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 357 (1999) 2495–2509.
[7] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2006) 489–509.
[8] E. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Inform. Theory 51 (2005) 4203–4215.
[9] E. Candès, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Ann. Stat. 35 (2007) 2313–2351.
[10] A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best k-term approximation, J. Amer. Math. Soc. 22 (2009) 211–231.

[11] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, D. Picard, Capturing ridge functions in high dimensions from point queries, Constr. Approx. 35 (2012) 225–243.

[12] M.A. Davenport, M.F. Duarte, Y.C. Eldar, G. Kutyniok, Introduction to Compressed Sensing. Compressed Sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 1–64.

[13] R. DeVore, G.G. Lorentz, Constructive Approximation, in: Grundlehren der Mathematischen Wissenschaften, vol. 303, Springer, Berlin, 1993.

[14] R. DeVore, G. Petrova, P. Wojtaszczyk, Instance-optimality in probability with an ℓ_1-minimization decoder, Appl. Comput. Harmon. Anal. 27 (2009) 275–288.

[15] R. DeVore, G. Petrova, P. Wojtaszczyk, Approximation of functions of few variables in high dimensions, Constr. Approx. 33 (2011) 125–143.

[16] D.L. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (2006) 1289–1306.

[17] M. Fornasier, H. Rauhut, Compressive sensing, in: Otmar Scherzer (Ed.), Handbook of Mathematical Methods in Imaging, Springer, 2011, pp. 187–228.

[18] M. Fornasier, K. Schnass, J. Vybíral, Learning functions of few arbitrary linear parameters in high dimensions, Found. Comput. Math. 12 (2012) 229–262.

[19] S. Foucart, H. Rauhut, A mathematical introduction to compressive sensing, in: Applied and Numerical Harmonic Analysis, Birkhäuser, Boston, 2013.

[20] T. Hemant, V. Cevher, Active learning of multi-index function models, in: Advances in Neural Information Processing Systems 25, 2012, pp. 1475–1483. available at http://books.nips.cc/papers/files/nips25/NIPS2012_0701.pdf.

[21] M. Ledoux, The Concentration of Measure Phenomenon, in: Mathematical Surveys and Monographs, vol. 89, American Mathematical Society, Providence, 2001.

[22] V. Ya. Lin, A. Pinkus, Fundamentality of ridge functions, J. Approx. Theory 75 (1993) 295–311.

[23] A.E. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, Smallest singular value of random matrices and geometry of random polytopes, Adv. Math. 195 (2005) 491–523.

[24] B.P. Logan, L.A. Shepp, Optimal reconstruction of a function from its projections, Duke Math. J. 42 (1975) 645–659.

[25] S. Mayer, T. Ullrich, J. Vybíral, Entropy and sampling numbers of classes of ridge functions, Constr. Approx. (in press).

[26] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems, Volume I: Linear Information, in: EMS Tracts in Mathematics, vol. 6, Eur. Math. Soc. Publ. House, Zürich, 2008.

[27] E. Novak, H. Woźniakowski, Approximation of infinitely differentiable multivariate functions is intractable, J. Complexity 25 (2009) 398–404.

[28] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems, Volume II: Standard Information for Functionals, in: EMS Tracts in Mathematics, vol. 12, Eur. Math. Soc. Publ. House, Zürich, 2010.

[29] E. Novak, H. Woźniakowski, Tractability of Multivariate Problems, Volume III: Standard Information for Operators, in: EMS Tracts in Mathematics, vol. 18, Eur. Math. Soc. Publ. House, Zürich, 2012.

[30] A. Pinkus, Approximating by ridge functions, in: Surface Fitting and Multiresolution Methods, 1997, pp. 279–292.

[31] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.

[32] K. Schnass, J. Vybíral, Compressed learning of high-dimensional sparse functions, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 3924–3927.

[33] P. Wojtaszczyk, Complexity of approximation of functions of few variables in high dimensions, J. Complexity 27 (2011) 141–150.

Constr Approx (2015) 42:231–264
DOI 10.1007/s00365-014-9267-x

CONSTRUCTIVE
APPROXIMATION

Entropy and Sampling Numbers of Classes of Ridge Functions

Sebastian Mayer · Tino Ullrich · Jan Vybíral

Received: 12 November 2013 / Revised: 16 July 2014 / Accepted: 14 August 2014 /
Published online: 27 September 2014

Abstract We study the properties of ridge functions $f(x) = g(a \cdot x)$ in high dimensions d from the viewpoint of approximation theory. The function classes considered consist of ridge functions such that the profile g is a member of a univariate Lipschitz class with smoothness $\alpha > 0$ (including infinite smoothness) and the ridge direction a has p-norm $\|a\|_p \leq 1$. First, we investigate entropy numbers in order to quantify the compactness of these ridge function classes in L_∞. We show that they are essentially as compact as the class of univariate Lipschitz functions. Second, we examine sampling numbers and consider two extreme cases. In the case $p = 2$, sampling ridge functions on the Euclidean unit ball suffers from the curse of dimensionality. Moreover, it is as difficult as sampling general multivariate Lipschitz functions, which is in sharp contrast to the result on entropy numbers. When we additionally assume that all feasible profiles have a first derivative uniformly bounded away from zero at the origin, the complexity of sampling ridge functions reduces drastically to the complexity of sampling univariate Lipschitz functions. In between, the sampling problem's degree of difficulty varies, depending on the values of α and p. Surprisingly, we see almost the entire hierarchy of tractability levels as introduced in the recent monographs by Novak and Woźniakowski.

Communicated by Allan Pinkus.

S. Mayer · T. Ullrich (✉)
Hausdorff-Center for Mathematics, Endenicher Allee 62, 53115 Bonn, Germany
e-mail: tino.ullrich@hcm.uni-bonn.de

J. Vybíral
Department of Mathematical Analysis, Charles University, Sokolovska 83,
186 00 Prague 8, Czech Republic

Keywords Ridge functions · Sampling numbers · Entropy numbers · Rate of convergence · Information-based complexity · Curse of dimensionality

Mathematics Subject Classification 41A10 · 41A25 · 41A50 · 41A63 · 46E35 · 65D05 · 65D15

1 Introduction

Functions depending on a large number of variables (or even infinitely many variables) naturally appear in many real-world applications. Since analytical representations are rarely available, there is a need to compute approximations to such functions or at least functionals thereof. Examples include parametric and stochastic PDEs [7,34], data analysis and learning theory [1,8,17], quantum chemistry [11], and mathematical finance [29].

It is a very well-known fact that approximation of smooth multivariate functions suffers from the so-called *curse of dimensionality* in many cases. In particular, for fixed smoothness, the order of approximation decays rapidly with increasing dimension [9,23]. A recent result [27] from the area of *information-based complexity* states that on the unit cube, even uniform approximation of infinitely differentiable functions is intractable in high dimensions. These results naturally lead to the search for assumptions other than smoothness, which would allow for tractable approximation but would still be broad enough to include real-world applications. There are many different conditions of this kind. Usually, they require additional structure; for example, that the functions under consideration are tensor products or belong to some sort of weighted function space. We refer to [35] for an introduction to information-based complexity and [26,28] for a detailed discussion of (in)tractability of high-dimensional problems.

In this work, we are interested in functions, which take the form of a *ridge*. This means that for each function f, there is direction a along which f may vary; along lines perpendicular to a the function is constant. In other words, the function is of the form $f(x) = g(a \cdot x)$, where g is a univariate function called the profile. Ridge functions provide a simple, coordinate-independent model that describes inherently one-dimensional structures hidden in a high-dimensional ambient space.

That the unknown functions take the form of a ridge is a frequent assumption in statistics, for instance, in the context of *single index models*. For several of such statistical problems, minimax bounds have been studied on the basis of algorithms that exploit the ridge structure [15,20,32]. Another approach with ridge functions, which has attracted attention for more than 30 years, is to approximate by ridge functions. An early work in this direction is [22], which was motivated by computerized tomography, and in which the term “ridge function” was actually coined. Another seminal paper is [14], which introduced *projection pursuit regression* for data analysis. More recent works include the mathematical analysis of neural networks [3,31] and wavelet-type analysis [4]. For a survey on further approximation theoretical results, we refer the reader to [30].

For classical setups in statistics and data analysis, it is typical that we have no influence on the choice of sampling points. In contrast, problems of *active learning* allow one to *freely* choose a limited number of samples from which to recover the function. Such a situation occurs, for instance, if sampling the unknown function at a point is realized by a (costly) PDE solver. In this context, ridge functions have appeared only recently as function models. The papers [6] and [12,37] provide several algorithms and upper bounds for the approximation error, the latter two even for the more general situation that $f(x) = g(Ax)$ with A a $(k \times d)$ matrix.

In the present paper, the central objective is to determine the complexity of approximating ridge functions when the only available information is a limited amount of function values. We make the following assumptions: The ridge functions' domain is the d-dimensional Euclidean unit ball; the profiles are Lipschitz of order $\alpha > 0$ (including infinite smoothness $\alpha = \infty$); the ridge vectors are contained in a ℓ_p^d-ball with $0 < p \leq 2$. Additionally, we study the situation when one additionally knows that $|g'(0)| \geq \kappa$ for all admissible profiles g and some prescribed $0 < \kappa \leq 1$ (of course, this only makes sense in the case of $\alpha > 1$). For the function classes given by these a priori assumptions, we prove lower and upper bounds for the deterministic worst-case error with regard to standard information. Following [25], we use the term *sampling numbers* for this worst-case error.

For given Lipschitz smoothness α, the ridge function classes are contained in the unit ball of the space of general multivariate Lipschitz functions of order α. The latter, in turn, is related to isotropic d-variate Besov spaces. For those spaces, it is known that their *entropy numbers*, which quantify the compactness in L_∞, provide a fair indicator for the behavior of sampling numbers, see [25]. We investigate whether or not this is still the case for the ridge function classes. It turns out that they are essentially as compact as the class of univariate Lipschitz functions of the same order for all possible parameter values. For the sampling problem, however, we find a much more diverse picture. At first glance, the simple structure of ridge functions suggests that approximating them should not be too much harder than approximating a univariate function. But this is far from true in general. In fact, the sampling problem's degree of difficulty crucially depends on the constraint $|g'(0)| \geq \kappa$ in our setting. If $\kappa > 0$, then it becomes possible to first recover the ridge direction efficiently. What remains then is only the one-dimensional problem of sampling the profile. Thus, the ridge structure has a sweeping impact in this scenario and leads to a *polynomially tractable* problem. Moreover, the behavior of entropy and sampling numbers is similar. But without the constraint on first derivatives, the picture is completely different. Sampling ridge functions is now essentially as hard as sampling general Lipschitz functions over the same domain, given that all vectors in the domain may occur as ridge direction ($p = 2$). It even suffers from the *curse of dimensionality* as long as we only have finite smoothness of profiles. Supposing that $p < 2$, which can be interpreted as imposing a sparsity constraint on the ridge vectors, mitigates the situation to some extent. To our surprise, we see almost the entire spectrum of degrees of tractability as introduced in the recent monographs by Novak and Woźniakowski. In any case, however, entropy and sampling numbers behave totally differently.

The paper is organized as follows. In Sect. 2, we define the setting in a precise way and introduce central concepts. Section 3 is dedicated to the study of entropy numbers for the ridge function classes. Lower and upper bounds on sampling numbers are found in Sect. 4. Finally, in Sect. 5, we interpret our findings on sampling numbers in the language of information-based complexity.

2 Preliminaries

Notation For $x \in \mathbb{R}^d$, recall the (quasi-)norms $\|x\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$ for $0 < p < \infty$, and $\|x\|_\infty = \max\{|x_1|, \ldots, |x_d|\}$. When X denotes a (quasi-)Banach space, equipped with the (quasi-)norm $\|\cdot\|_X$, we write $B_X = \{f \in X : \|f\|_X < 1\}$ for the open unit ball and $\bar{B}_X$ for its closure. In the special case that $X = \ell_p^d(\mathbb{R}) = (\mathbb{R}^d, \|\cdot\|_p)$, we additionally use the notation B_p^d for the open unit ball and $\mathbb{S}_p^{d-1}$ for the unit sphere in ℓ_p^d.

The notation $f \lesssim g$ means that $f \leq Cg$ for some constant $C > 0$. Likewise, we write $f \gtrsim g$ if $f \geq cg$ for some constant $c > 0$, and $f \asymp g$ if both $f \lesssim g$ and $f \gtrsim g$.

2.1 Ridge Function Classes

The specific form of ridge functions suggests that one describe a class of such functions in terms of two parameters: one to determine the smoothness of profiles and the other to restrict the norm of ridge directions.

Regarding smoothness, we require that ridge profiles are Lipschitz of some order. For the reader's convenience, let us briefly recall this notion. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and s be a natural number. The function space $C^s(\Omega)$ consists of those functions over the domain Ω, which have partial derivatives up to order s in the interior $\mathring{\Omega}$ of Ω, and these derivatives are moreover bounded and continuous in Ω. Formally,

$$C^s(\Omega) = \left\{ f : \Omega \to \mathbb{R} : \quad \|f\|_{C^s} := \max_{|\gamma| \leq s} \|D^\gamma f\|_\infty < \infty \right\},$$

where, for any multi-index $\gamma = (\gamma_1, \ldots, \gamma_d) \in \mathbb{N}_0^d$, the partial differential operator D^γ is given by

$$D^\gamma f := \frac{\partial^{|\gamma|} f}{\partial x_1^{\gamma_1} \cdots \partial x_d^{\gamma_d}}.$$

Here, we have written $|\gamma| = \sum_{i=1}^d \gamma_i$ for the order of D^γ. For the vector of first derivatives, we use the usual notation $\nabla f = (\partial f/\partial x_1, \ldots, \partial f/\partial x_d)$. Besides $C^s(\Omega)$, we further need the space of infinitely differentiable functions $C^\infty(\Omega)$ defined by

$$C^\infty(\Omega) = \left\{ f : \Omega \to \mathbb{R} : \quad \|f\|_{C^\infty} := \sup_{\gamma \in \mathbb{N}_0^d} \|D^\gamma f\|_\infty < \infty \right\}. \tag{2.1}$$

For a function $f : \Omega \to \mathbb{R}$ and any positive number $0 < \beta \leq 1$, the *Hölder constant* of order β is given by

$$|f|_\beta := \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|f(x) - f(y)|}{2\min\{1, \|x - y\|_1\}^\beta} .$$

This definition immediately implies the relation

$$|f|_\beta \leq |f|_{\beta'} \text{ if } 0 < \beta < \beta' \leq 1. \tag{2.2}$$

Now, for any $\alpha > 0$, we can define the *Lipschitz space* $\mathrm{Lip}_\alpha(\Omega)$. If we let $s = \llbracket\alpha\rrbracket$ be the largest integer *strictly less* than α, it contains those functions in $C^s(\Omega)$ which have partial derivatives of order s which are moreover Hölder-continuous of order $\beta = \alpha - s > 0$. Formally,

$$\mathrm{Lip}_\alpha(\Omega) = \Big\{f \in C^s(\Omega) : \quad \|f\|_{\mathrm{Lip}_\alpha(\Omega)} := \max\{\|f\|_{C^s}, \max_{|\gamma|=s} |D^\gamma f|_\beta\} < \infty\Big\}.$$

For $s \in \mathbb{N}_0$ and $1 \geq \beta_2 > \beta_1 > 0$, the following embeddings hold true:

$$C^\infty(\Omega) \subset \mathrm{Lip}_{s+\beta_2}(\Omega) \subset \mathrm{Lip}_{s+\beta_1}(\Omega) \subset C^s(\Omega) \subset \mathrm{Lip}_s(\Omega), \tag{2.3}$$

where the respective identity operators are of norm one. In other words, the respective unit balls satisfy the same relation. Note that the fourth inclusion only makes sense if $s \geq 1$. The third embedding is a trivial consequence of the definition. The second embedding follows from the third and (2.2). The fourth embedding and the second imply the first. So it remains to establish the fourth embedding. We have to show that for every $\gamma \in \mathbb{N}_0^d$ with $|\gamma| = s - 1$, it holds that $|D^\gamma f|_1 \leq \|f\|_{C^s}$. On the one hand, Taylor's formula in $\mathbb{R}^d$ gives for some $0 < \theta < 1$,

$$\begin{aligned} |D^\gamma f(x) - D^\gamma f(y)| &= |\nabla(D^\gamma f)(x + \theta(y - x)) \cdot (x - y)| \\ &\leq \max_{|\beta|=s} \|D^\beta f\|_\infty \cdot \|x - y\|_1 \\ &\leq \|f\|_{C^s}\|x - y\|_1. \end{aligned}$$

On the other hand, we have $|D^\gamma f(x) - D^\gamma f(y)| \leq 2\|f\|_{C^s}$. Both estimates together yield $|D^\gamma f|_1 \leq \|f\|_{C^s}$.

Having introduced Lipschitz spaces, we can give a formal definition of our classes of ridge functions. For the rest of the paper, we fix as function domain the closed unit ball

$$\Omega = \bar{B}_2^d = \{x \in \mathbb{R}^d \ : \ \|x\|_2 \leq 1\}.$$

As before, let $\alpha > 0$ denote the order of Lipschitz smoothness. Further, let $0 < p \leq 2$. We define the class of ridge functions with Lipschitz profiles as

$$\mathcal{R}_d^{\alpha,p} = \left\{ f : \Omega \to \mathbb{R} \ : \ f(x) = g(a \cdot x), \ \|g\|_{\mathrm{Lip}_\alpha[-1,1]} \leq 1, \ \|a\|_p \leq 1 \right\}. \quad (2.4)$$

In addition, we define the class of ridge functions with infinitely differentiable profiles by

$$\mathcal{R}_d^{\infty,p} = \left\{ f : \Omega \to \mathbb{R} \ : \ f(x) = g(a \cdot x), \ \|g\|_{C^\infty[-1,1]} \leq 1, \ \|a\|_p \leq 1 \right\}.$$

Let us collect basic properties of these classes.

Lemma 2.1 *For any $\alpha > 0$ and $0 < p \leq 2$, the class $\mathcal{R}_d^{\alpha,p}$ is contained in $\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}$ and $\mathcal{R}_d^{\infty,p}$ is contained in $\bar{B}_{C^\infty(\Omega)}$.*

Proof Let $f \in \mathcal{R}_d^{\alpha,p}$ and $s = \llbracket \alpha \rrbracket$. Furthermore, let $\gamma \in \mathbb{N}_0^d$ be such that $|\gamma| \leq s$. Then, there exists $g \in \mathrm{Lip}_\alpha([-1,1])$ with

$$D^\gamma f(x) = D^{|\gamma|} g(a \cdot x) a^\gamma, \quad x \in \Omega,$$

where we used the convention $a^\gamma = \prod_{i=1}^d a_i^{\gamma_i}$. Therefore, we have

$$\|D^\gamma f\|_\infty \leq \|D^{|\gamma|} g\|_\infty \|a\|_\infty^{|\gamma|} \leq \|a\|_p^{|\gamma|} \leq 1.$$

If we let $s \to \infty$, this immediately implies $\mathcal{R}_d^{\infty,p} \subset \bar{B}_{C^\infty(\Omega)}$. Moreover, if $|\gamma| = s$ and $\beta = \alpha - s$, we obtain by Hölder's inequality for $x, y \in \Omega$,

$$\begin{aligned} |D^\gamma f(x) - D^\gamma f(y)| &= |a^\gamma| \cdot |D^s g(a \cdot x) - D^s g(a \cdot y)| \\ &\leq \|a\|_p^s \cdot |D^s g|_\beta \cdot 2 \min\{1, \|a\|_p \cdot \|x - y\|_1\}^\beta \\ &\leq 2 \min\{1, \|x - y\|_1\}^\beta. \end{aligned}$$

Consequently, we have $\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \leq 1$ and hence $\mathcal{R}_d^{\alpha,p} \subset \bar{B}_{\mathrm{Lip}_\alpha(\Omega)}$. □

Note that in the special case $\alpha = 1$, we have Lipschitz-continuous profiles. Whenever $0 < \alpha_1 < \alpha_2 \leq \infty$, we have $\mathcal{R}_d^{\alpha_2,p} \subset \mathcal{R}_d^{\alpha_1,p}$, which is an immediate consequence of (2.3). Likewise, for $p < q$, we have the relation $\mathcal{R}_d^{\alpha,p} \subset \mathcal{R}_d^{\alpha,q}$.

Finally, for Lipschitz smoothness $\alpha > 1$, we want to introduce a restricted version of $\mathcal{R}_d^{\alpha,p}$, where profiles obey the additional constraint $|g'(0)| \geq \kappa > 0$. See Sect. 4.2 and in particular Remark 4.9 for an explanation of why we study this additional constraint. We define

$$\mathcal{R}_d^{\alpha,p,\kappa} = \{g(a\cdot) \in \mathcal{R}_d^{\alpha,p} : \quad |g'(0)| \geq \kappa\}. \quad (2.5)$$

Hereafter, whenever we say that we consider ridge functions with first derivatives bounded away from zero in the origin, we mean that they are contained in the class $\mathcal{R}_d^{\alpha,p,\kappa}$ for some $0 < \kappa \leq 1$.

Taylor expansion. We introduce a straightforward, multivariate extension of Taylor's expansion on intervals to ridge functions in $\mathcal{R}_d^{\alpha,p}$ and functions in $\mathrm{Lip}_\alpha(\Omega)$. For $x, x^0 \in \mathring{\Omega}$, we define the function $\Phi_x(\cdot)$ by

$$\Phi_x(t) := f(x^0 + t(x - x^0)), \quad t \in [0, 1].$$

Lemma 2.2 *Let* $\alpha > 1$ *and* $\alpha = s + \beta$, $s \in \mathbb{N}$, $0 < \beta \leq 1$. *Let* $f \in \mathrm{Lip}_\alpha(\Omega)$ *and* $x, x^0 \in \mathring{\Omega}$. *Then, there is a real number* $\theta \in (0, 1)$ *such that*

$$f(x) = T_{s,x^0} f(x) + R_{s,x^0} f(x),$$

where the Taylor polynomial $T_{s,x^0} f(x)$ *is given by*

$$T_{s,x^0} f(x) = \sum_{j=0}^{s} \frac{\Phi_x^{(j)}(0)}{j!} = \sum_{|\gamma| \leq s} \frac{D^\gamma f(x^0)}{\gamma!}(x - x^0)^\gamma$$

and the remainder by

$$R_{s,x^0} f(x) = \frac{1}{s!}\Big(\Phi_x^{(s)}(\theta) - \Phi_x^{(s)}(0)\Big) \tag{2.6}$$

$$= \sum_{|\gamma|=s} \frac{D^\gamma f(x^0 + \theta(x - x^0)) - D^\gamma f(x^0)}{\gamma!}(x - x^0)^\gamma. \tag{2.7}$$

The previous lemma has a nice consequence for the approximation of functions from $\mathcal{R}_d^{\alpha,p}$ in the case $\alpha > 1$ and $0 < p \leq 2$. Let p' denote the dual index of p given by $1/\max\{p, 1\} + 1/p' = 1$.

Lemma 2.3 *Let* $\alpha = s + \beta > 1$ *and* $\Omega = \bar{B}_2^d$.

(i) *For* $f \in \mathrm{Lip}_\alpha(\Omega)$ *and* $x, x^0 \in \mathring{\Omega}$, *we have*

$$|f(x) - T_{s,x^0} f(x)| \leq 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \frac{\left\|x - x^0\right\|_1^\alpha}{s!}.$$

(ii) *Let* $0 < p \leq 2$. *Then, for* $f \in \mathcal{R}_d^{\alpha,p}$, *we have the slightly better estimate*

$$|f(x) - T_{s,x^0} f(x)| \leq \frac{2}{s!}\|x - x^0\|_{p'}^\alpha.$$

Proof To prove (i) we use (2.7) and the definition of $\mathrm{Lip}_\alpha(\Omega)$ and estimate as follows:

$$\begin{aligned} |f(x) - T_{s,x^0} f(x)| &\le \sum_{|\gamma|=s} \frac{|D^\gamma f(x^0 + \theta(x - x^0)) - D^\gamma f(x^0)|}{\gamma!} |(x - x^0)^\gamma| \\ &\le 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \min\{1, \|x - x^0\|_1\}^\beta \cdot \sum_{|\gamma|=s} \frac{\prod_{i=1}^d |x_i - x_i^0|^{\gamma_i}}{\gamma!}. \end{aligned}$$

Using mathematical induction, it is straightforward to verify the multinomial identity

$$(a_1 + \cdots + a_d)^s = \sum_{|\gamma|=s} \frac{s!}{\gamma!} a_1^{\gamma_1} \ldots a_d^{\gamma_d}.$$

Hence, choosing $a_i = |x_i - x_i^0|$, we can continue estimating

$$|f(x) - T_{s,x^0} f(x)| \le 2\|f\|_{\mathrm{Lip}_\alpha(\Omega)} \min\left\{1, \left\|x - x^0\right\|_1\right\}^\beta \frac{\|x - x^0\|_1^s}{s!}$$

and obtain the assertion in (i).
For showing the improved version (ii) for functions of type $f(x) = g(a \cdot x)$, we use formula (2.6) of the Taylor remainder. We easily see that for $t \in (0, 1)$, it holds that

$$\Phi_x^{(s)}(t) = g^{(s)}\Big(a \cdot (x^0 + t(x - x^0))\Big) \cdot [a \cdot (x - x^0)]^s.$$

Using Hölder continuity of $g^{(s)}$ of order β and Hölder's inequality, we see that

$$\begin{aligned} &|f(x) - T_{s,x^0} f(x)| \\ &\quad \le \frac{1}{s!}\Big|[a \cdot \Big(x - x^0\Big)]^s \cdot \Big\{g^{(s)}\Big(a \cdot (x^0 + \theta(x - x^0))\Big) - g^{(s)}(a \cdot x^0)\Big\}\Big| \\ &\quad \le \frac{1}{s!}\|a\|_p^s \cdot \left\|x - x^0\right\|_{p'}^s \cdot 2\min\{1, |\theta a \cdot \Big(x - x^0\Big)|^\beta\} \\ &\quad \le \frac{2}{s!}\left\|x - x^0\right\|_{p'}^\alpha. \end{aligned}$$

The proof is complete. □

2.2 Information Complexity and Tractability

In this work, we want to approximate ridge functions from $\mathcal{F} = \mathcal{R}_d^{\alpha,p}$ or $\mathcal{F} = \mathcal{R}_d^{\alpha,p,\kappa}$ by means of deterministic sampling algorithms, using a limited amount of function values. Any allowed algorithm S consists of an *information map* $N_S^{\mathrm{ada}} : \mathcal{F} \to \mathbb{R}^n$ and a *reconstruction map* $\varphi_S : \mathbb{R}^n \to L_\infty(\Omega)$. For given $f \in \mathcal{F}$, the former provides function values $f(x_1), \ldots, f(x_n)$ at points $x_1, \ldots, x_n \in \Omega$, which are allowed to be chosen *adaptively*. Adaptivity here means that x_i may depend on the preceding values $f(x_1), \dots, f(x_{i-1})$. According to [26], we speak of *standard information*. The reconstruction map then builds an approximation to f based on those function values provided by the information map.

Formally, we consider the class of deterministic, adaptive sampling algorithms $\mathcal{S}^{\text{ada}} = \bigcup_{n\in\mathbb{N}} \mathcal{S}_n^{\text{ada}}$, where

$$\mathcal{S}_n^{\text{ada}} = \Big\{ S : \mathcal{F} \to L_\infty(\bar{B}_2^d) \ : \\ S = \varphi \circ N^{\text{ada}}, \varphi : \mathbb{R}^m \to L_\infty \ N^{\text{ada}} : \mathcal{F} \to \mathbb{R}^m, \ m \le n \Big\}.$$

The *nth minimal worst-case error*

$$g_{n,d}^{\text{ada}}(\mathcal{F}, L_\infty) := \text{err}_{n,d}(\mathcal{F}, \mathcal{S}^{\text{ada}}, L_\infty) = \inf \left\{ \sup_{f\in\mathcal{F}} \|f - S(f)\|_\infty : S \in \mathcal{S}_n^{\text{ada}} \right\}$$

describes the approximation error of the best possible algorithm. Stressing that function values are the only available information, we refer to $g_{n,d}^{\text{ada}}(\mathcal{F}, L_\infty)$ as the nth *(adaptive) sampling number*. To reveal the effect of adaption, it is useful to compare adaptive algorithms with the subclass $\mathcal{S} \subset \mathcal{S}^{\text{ada}}$ of *nonadaptive*, deterministic algorithms, that is, for each algorithm $S \in \mathcal{S}$, the information map is now of the form $N_S = (\delta_{x_1}, \dots, \delta_{x_n})$, with $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \bar{B}_2^d$. This corresponds to *nonadaptive standard information* in [26]. The associated nth worst-case error

$$g_{n,d}(\mathcal{F}, L_\infty) := \inf_{S\in\mathcal{S}_n} \sup_{f\in\mathcal{F}} \big\| f - S(f) \big\|_\infty = \text{err}_{n,d}(\mathcal{F}, \mathcal{S}_n, L_\infty)$$

coincides with the standard nth *sampling number* as known in approximation theory [25]. As a third restriction, let us introduce the nth *linear* sampling number $g_{n,d}^{\text{lin}}(\mathcal{F}, L_\infty)$; here, only algorithms from $\mathcal{S}$ with linear reconstruction maps are allowed. Clearly,

$$g_{n,d}^{\text{ada}}(\mathcal{F}, L_\infty) \le g_{n,d}(\mathcal{F}, L_\infty) \le g_{n,d}^{\text{lin}}(\mathcal{F}, L_\infty).$$

Remark 2.4 Studying adaptive algorithms makes sense since the considered classes of ridge functions are not *convex*. Hence, the general results on linear problems [26, Section 4.2] do not apply here. Nevertheless, the analysis in Sect. 4 will reveal that neither adaptivity nor nonlinearity lead to any substantial improvement in the approximation of ridge functions defined on a Euclidean ball.

Whenever we speak of sampling of ridge functions, we refer to the problem of approximating ridge functions in $\mathcal{F}$ by sampling algorithms from $\mathcal{S}^{\text{ada}}$, the L_∞-approximation error measured in the worst case. Its *information complexity* $n(\varepsilon, d)$ is given for $0 < \varepsilon \le 1$ and $d \in \mathbb{N}$ by

$$n(\varepsilon, d) := \min \left\{ n \in \mathbb{N} : g_{n,d}^{\text{ada}}(\mathcal{F}, L_\infty) \le \varepsilon \right\}.$$

2.3 Entropy Numbers

The concept of entropy numbers is central to this work. An entropy number can be understood as a measure to quantify the compactness of a set with respect to some reference space. For a detailed discussion and historical remarks, we refer to the monographs [5,10]. The kth entropy number $e_k(K, X)$ of a subset K of a (quasi-)Banach space X is defined as

$$e_k(K, X) = \inf \left\{ \varepsilon > 0 : K \subset \bigcup_{j=1}^{2^{k-1}} (x_j + \varepsilon \bar{B}_X) \text{ for some } x_1, \ldots, x_{2^{k-1}} \in X \right\}.$$

Note that $e_k(K, X) = \inf\{\varepsilon > 0 : N_\varepsilon(K, X) \leq 2^{k-1}\}$ holds true, where

$$N_\varepsilon(K, X) := \min \left\{ n \in \mathbb{N} : \quad \exists x_1, \ldots, x_n \in X : \ K \subset \bigcup_{j=1}^{n} (x_j + \varepsilon \bar{B}_X) \right\}$$

denotes the *covering number* of the set K in the space X, which is the minimal natural number n such that there is an ε-net of K in X of n elements. We can introduce entropy numbers for operators, as well. The kth entropy number $e_k(T)$ of an operator $T : X \to Y$ between two quasi-Banach spaces X and Y is defined by

$$e_k(T) = e_k\left(T(\bar{B}_X), Y\right).$$

The results in Sects. 3 and 4 rely to a great degree on entropy numbers of the identity operator between the two finite dimensional spaces $X = \ell_p^d(\mathbb{R})$ and $Y = \ell_q^d(\mathbb{R})$. Their behavior is understood very well, see [10,21,33,36]. For the reader's convenience, we restate the result.

Lemma 2.5 *Let $0 < p \leq q \leq \infty$, and let k and d be natural numbers. Then,*

$$e_k(\bar{B}_p^d, \ell_q^d) \asymp \begin{cases} 1, & 1 \leq k \leq \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q}, & \log(d) \leq k \leq d, \\ 2^{-k/d} d^{1/q-1/p}, & k \geq d. \end{cases}$$

The constants behind "$\asymp$" depend neither on k nor on d. They only depend on the parameters p and q.

If we consider entropy numbers of ℓ_p^d-spheres instead of ℓ_p^d-balls in ℓ_q^d, the situation is quite similar. We are not aware of a reference where this has already been formulated thoroughly.

Lemma 2.6 *Let $d \in \mathbb{N}$, $d \geq 2$, $0 < p \leq q \leq \infty$, and $\bar{p} = \min\{1, p\}$. Then,*

(i) $2^{-k/(d-1)} d^{1/q-1/p} \lesssim e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \lesssim 2^{-k/(d-\bar{p})} d^{1/q-1/p}, \quad k \geq d.$

(ii)

$$e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \asymp \begin{cases} 1, \; 1 \le k \le \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q}, \; \log(d) \le k \le d. \end{cases}$$

The constants behind "$\asymp$" only depend on p and q.

Proof For given $\varepsilon > 0$, an ε-covering $\{y_1, \dots, y_N\}$ of $\mathbb{S}_p^{d-1}$ in ℓ_p^d fulfills

$$(1+\varepsilon)\bar{B}_p^d \setminus (1-\varepsilon)\bar{B}_p^d \subseteq \bigcup_{i=1}^N (y_i + 2^{1/\bar{p}}\varepsilon\bar{B}_p^d). \tag{2.8}$$

Let $\bar{q} = \min\{1, q\}$. For given $\varepsilon > 0$, a maximal set $\{x_1, \dots, x_M\} \subset \mathbb{S}_p^{d-1}$ of vectors with mutual distance greater than ε obeys

$$\bigcup_{i=1}^M (x_i + 2^{-1/\bar{q}}\,\varepsilon\bar{B}_q^d) \subseteq (1+\varepsilon_d^{\bar{p}})^{1/\bar{p}}\bar{B}_p^d \setminus (1-\varepsilon_d^{\bar{p}})^{1/\bar{p}}\bar{B}_p^d, \tag{2.9}$$

where $\varepsilon_d = 2^{-1/\bar{q}}\,\varepsilon\, d^{1/p-1/q}$.

(i). A standard volume argument applied to (2.8) yields $h(\varepsilon) \le N\varepsilon^d 2^{d/\bar{p}}$, where $h(\varepsilon) = (1+\varepsilon)^d - (1-\varepsilon)^d$. First-order Taylor expansion in ε allows one to estimate $h(\varepsilon) \ge d\varepsilon$. Solving for N yields a lower bound for covering numbers in the case $p = q$. The lower bound in the case $p \ne q$ follows from the trivial estimate $e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \ge d^{1/q-1/p}\, e_k(\mathbb{S}_p^{d-1}, \ell_p^d)$.
For the upper bound in the case $p = q$, a standard volume argument applied to (2.9) yields $M\varepsilon^d 2^{-d/\bar{p}} \le h_p(\varepsilon^{\bar{p}}/2)$ with $h_p(x) = (1+x)^{d/\bar{p}} - (1-x)^{d/\bar{p}}$. The mean value theorem gives $h_p(x) \le d/\bar{p}\, 2^{d/\bar{p}} x$ if $0 < x \le 1$. Hence, we get $h_p(\varepsilon^{\bar{p}}/2) \le d/\bar{p}\, 2^{d/\bar{p}}\varepsilon^{\bar{p}}/2$. Solving for M gives an upper bound for packing numbers and hence also for covering numbers. In the case $p \ne q$, we again use (2.9) and pass to volumes. This time, the quotient $\mathrm{vol}(B_p^d)/\mathrm{vol}(B_q^d)$ remains in the upper bound for M. The given bounds now easily translate to the stated bounds on entropy numbers. In the case $p \ne q$, one has to take

$$\left[\frac{\mathrm{vol}(B_p^d)}{\mathrm{vol}(B_q^d)}\right]^{1/(d-\bar{p})} \asymp d^{1/q-1/p}$$

into account to get the additional factor in d.

(ii). The proof by Kühn [21] immediately gives the lower bound. The upper bound follows trivially from $\mathbb{S}_p^{d-1} \subset \bar{B}_p^d$. □

Remark 2.7 Note that in the case $p \geq 1$, we have the sharp bounds

$$e_k(\mathbb{S}_p^{d-1}, \ell_q^d) \asymp \begin{cases} 1, & 1 \leq k \leq \log(d), \\ \left(\frac{\log(1+d/k)}{k}\right)^{1/p-1/q}, & \log(d) \leq k \leq d, \\ 2^{-\frac{k}{d-1}} d^{1/q-1/p}, & k \geq d. \end{cases}$$

In the case $p < 1$, there remains a gap between the upper and lower estimate for $e_k(\mathbb{S}_p^{d-1}, \ell_q^d)$ if $k \geq d$. However, this gap can be closed by using a different proof technique, see [18].

3 Entropy Numbers of Ridge Functions

This section is devoted to the study of entropy numbers of the classes $\mathcal{R}_d^{\alpha,p}$ and $\mathcal{R}_d^{\alpha,p,\kappa}$. Specifically, we want to relate their behavior to that of entropy numbers of uni- and multivariate Lipschitz functions. This will give us an understanding how "large" the ridge function classes are. Let us stress that we are interested in the dependence of the entropy numbers on the underlying dimension d, as it is usually done in the area of information-based complexity.

To begin with, we examine uni- and multivariate Lipschitz functions from $\text{Lip}_\alpha[-1,1]$ and $\text{Lip}_\alpha(\Omega)$. Recall the notation $B_\alpha := B_{\text{Lip}_\alpha[-1,1]}$ and $B_{\text{Lip}_\alpha(\Omega)}$ for the respective open unit balls. The behavior of entropy numbers of univariate Lipschitz functions is well known, see for instance [23, Chap. 15, §2, Thm. 2.6].

Lemma 3.1 *For $\alpha > 0$, there exist two constants $0 < c_\alpha < C_\alpha$ such that*

$$c_\alpha k^{-\alpha} \leq e_k(\bar{B}_\alpha, L_\infty([-1,1])) \leq C_\alpha k^{-\alpha}, \quad k \in \mathbb{N}.$$

This behavior does not change if we consider only functions with first derivative in the origin bounded away from zero, as we do with the profiles in the class $\mathcal{R}_d^{\alpha,p,\kappa}$.

Proposition 3.2 *Let $\alpha > 1$ and $0 < \kappa \leq 1$. Consider the class*

$$\text{Lip}_\alpha^\kappa(\big[-1,1\big]) = \{f \in \text{Lip}_\alpha(\big[-1,1\big]) : \|f\|_{\text{Lip}_\alpha[-1,1]} \leq 1\ ,\ \big|f'(0)\big| \geq \kappa\}.$$

For the entropy numbers of this class, we have two constants $0 < c_\alpha < C_\alpha$ such that

$$c_\alpha k^{-\alpha} \leq e_k(\text{Lip}_\alpha^\kappa([-1,1]), L_\infty([-1,1])) \leq C_\alpha k^{-\alpha}, \quad k \in \mathbb{N}.$$

Proof The upper bound is immediate by Lemma 3.1. The lower bound is proved in the same way as for general univariate Lipschitz functions of order α except that we have to adapt the "bad" functions such that they meet the constraint on the first derivative in the origin. Again, set $s = \llbracket \alpha \rrbracket$ and $\beta = \alpha - s > 0$. Consider the standard smooth bump function

$$\varphi(x) = \begin{cases} e^{-\frac{1}{1-x^2}}, & |x| < 1, \\ 0, & |x| \geq 1. \end{cases}$$

Let

$$\psi_{k,b}(x) = \frac{c_\alpha \cdot \varphi(5k(x-b))}{k^\alpha}, \quad k \in \mathbb{N},\ b \in \mathbb{R},$$

where $c_\alpha = 1/(5^\alpha \|\varphi\|_{\mathrm{Lip}_\alpha})$. The scaling factor $c_\alpha k^{-\alpha}$ assures $\psi_{k,b} \in \mathrm{Lip}_\alpha([-1,1])$. Let $a = \pi/4 - 1/5$ and $I = [a, a+2/5] \subset (0,1)$. We set $h(x) = \sin(x)$ and

$$\gamma = \sup_{j \in \mathbb{N}_0} \max_{x \in I} |h^{(j)}(x)| = \max_{x \in I} \max\{\cos(x), \sin(x)\} < 1. \tag{3.1}$$

For any multi-index $\theta = (\theta_1, \ldots, \theta_k) \in \{0,1\}^k$, let

$$g_\theta = (1-\gamma) \sum_{j=1}^{k} \theta_j \psi_{k,b_j}, \qquad b_j = a + \frac{2j-1}{5k}.$$

Observe that $\operatorname{supp} g_\theta \subset I$.

There are 2^k such multi-indices, and for two different multi-indices $\hat{\theta}$ and $\tilde{\theta}$, we have

$$\left\| g_{\hat{\theta}} - g_{\tilde{\theta}} \right\|_\infty = (1-\gamma) \left\| \psi_{k,0} \right\|_\infty = c_\alpha (1-\gamma) e^{-1} k^{-\alpha}.$$

Set $f_\theta = h + g_\theta$. Because of the scaling factors, it is assured that $f_\theta \in \mathrm{Lip}_\alpha^\kappa([-1,1])$. On the other hand, $f'_\theta(0) = \cos(0) = 1$. Obviously, $\left\| f_{\tilde{\theta}} - f_{\hat{\theta}} \right\|_\infty = \left\| g_{\tilde{\theta}} - g_{\hat{\theta}} \right\|_\infty$. We conclude

$$e_k(\mathrm{Lip}_\alpha^\kappa([-1,1]), L_\infty) \geq c'_\alpha k^{-\alpha}$$

for $c'_\alpha = (1-\gamma)e^{-1}c_\alpha$. □

Considering multivariate Lipschitz functions, decay rates of entropy numbers change dramatically compared to those of univariate Lipschitz functions; they depend exponentially on $1/d$. This is known if the domain is a cube $\Omega = I^d$, see [23, Chap. 15, §2]. We provide an extension to our situation where the domain is $\Omega = \bar{B}_2^d$.

Proposition 3.3 *Let $\alpha > 0$. For natural numbers n and k such that $2^{k-1} < n \leq 2^k$, we have*

$$e_n(\bar{B}_{\mathrm{Lip}_\alpha(\bar{B}_2^d)}, L_\infty(\bar{B}_2^d)) \geq c_\alpha e_{k+1}(id : \ell_2^d \to \ell_2^d)^\alpha.$$

In particular, we have $e_n(id : \mathrm{Lip}_\alpha(\bar{B}_2^d) \to L_\infty(\bar{B}_2^d)) \gtrsim n^{-\alpha/d}$.

Proof Consider the radial bump function $\varphi(x)$ given by

$$\varphi(x) = \begin{cases} e^{-\frac{1}{1-\|x\|_2^2}}, & \|x\|_2 < 1, \\ 0, & \|x\|_2 \geq 1. \end{cases}$$

Let $s = \llbracket \alpha \rrbracket$. With $c_\alpha := (\|\varphi\|_{\mathrm{Lip}_\alpha})^{-1}$, the rescaling

$$\varphi_\varepsilon^\alpha(x) := c_\alpha \varepsilon^\alpha \varphi(x/\varepsilon)$$

is contained in the closed unit ball of $\mathrm{Lip}_\alpha(\Omega)$.

For $0 < \varepsilon < e_{k+1}(\bar{B}_2^d, \ell_2^d)$, let $\{x_1, \dots, x_m\}$ be a maximal set of 2ε-separated points in the Euclidean ball $\bar{B}_2^d$, the distance measured in ℓ_2^d. Clearly, every closed ball of radius ε contains at most one x_i, and consequently every covering of $\bar{B}_2^d$ by balls of radius ε contains at least m elements. The choice of ε implies $m > 2^k \geq n$. For every multi-index $\theta \in \{0, 1\}^m$, we define

$$f_\theta(x) := \sum_{j=1}^{m} \theta_j \varphi_\varepsilon^\alpha(x - x_j).$$

By construction of $\varphi_\varepsilon^\alpha$, it is assured that $f_\theta \in \mathrm{Lip}_\alpha(\Omega)$ and $\| f_\theta \|_{\mathrm{Lip}_\alpha} \leq 1$. Moreover, we see immediately that $\| f_\theta \|_\infty = c_\alpha e^{-1} \varepsilon^\alpha$, and

$$\| f_\theta - f_{\theta'} \|_\infty \geq c_\alpha e^{-1} \varepsilon^\alpha =: \varepsilon_1$$

for $\theta \neq \theta'$. Therefore, the set $\{f_\theta : \theta \in \{0, 1\}^m\}$ consists of 2^m functions with mutual distances greater than or equal to ε_1. This implies

$$2^n < 2^m < N_{\varepsilon_1/2}(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty).$$

Hence, $e_n(id : \mathrm{Lip}_\alpha(\Omega) \to L_\infty(\Omega)) > \varepsilon_1/2$, and by the choice of ε, also

$$e_n(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty(\Omega)) > c'_\alpha e_k(id : \ell_2^d \to \ell_2^d)^\alpha$$

for $c'_\alpha = c_\alpha/(4e)$. Now, it follows immediately from the estimate above and Lemma 2.5 that

$$e_n(\bar{B}_{\mathrm{Lip}_\alpha(\Omega)}, L_\infty(\Omega)) \gtrsim 2^{-\alpha k/d} \gtrsim n^{-\alpha/d}.$$

□

Now consider ridge functions with Lipschitz profile as given by the class $\mathcal{R}_d^{\alpha,p}$.

Theorem 3.4 *Let d be a natural number, $\alpha > 0$, and $0 < p \le 2$. Then, for any $k \in \mathbb{N}$,*

$$\frac{1}{2}\max\{e_{2k}(\bar{B}_p^d, \ell_2^d), e_{2k}(\bar{B}_\alpha, L_\infty)\} \le e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty)$$
$$\le e_k(\bar{B}_p^d, \ell_2^d)^{\min\{\alpha,1\}} + e_k(\bar{B}_\alpha, L_\infty).$$

Proof Lower bounds: For $\varepsilon > 0$, let $g_1, \dots, g_n$ be a maximal set of univariate Lipschitz functions in $\bar{B}_\alpha$ with mutual distances $\left\| g_i - g_j \right\|_\infty > \varepsilon$ for $i \neq j$. Now, let $a = (1, 0, \dots, 0)$, and set $f_i(x) = g_i(a \cdot x)$ for $i = 1, \dots, n$. Then, of course, we have $f_i \in \mathcal{R}_d^{\alpha,p}$, and

$$\left\| f_i - f_j \right\|_\infty = \|g_i - g_j\|_\infty > \varepsilon.$$

Consequently, the functions $f_1, \dots, f_n$ are ε-separated, as well. This implies

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge \frac{1}{2} e_{2k}(\bar{B}_\alpha, L_\infty).$$

On the other hand, for $\varepsilon > 0$, let $a_1, \dots, a_n$ be a maximal set of vectors in $\bar{B}_p^d$ with pairwise distances $\left\| a_i - a_j \right\|_2 > \varepsilon$. Furthermore, let $g(t) = t$, and set $\tilde{f}_i(x) = g(a_i \cdot x)$ for $i = 1, \dots, n$. Then, $\tilde{f}_i \in \mathcal{R}_d^{\alpha,p}$, and

$$\begin{aligned}\|\tilde{f}_i - \tilde{f}_j\|_\infty &= \sup_{x \in \bar{B}_2^d} |\tilde{f}_i(x) - \tilde{f}_j(x)| = \sup_{x \in \bar{B}_2^d} |g(a_i \cdot x) - g(a_j \cdot x)| \\ &= \sup_{x \in \bar{B}_2^d} |(a_i - a_j) \cdot x| = \|a_i - a_j\|_2 > \varepsilon.\end{aligned}$$

Thus, the functions $\tilde{f}_1, \dots, \tilde{f}_n$ are ε-separated w.r.t. the L_∞-norm. This implies

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \ge \frac{1}{2} e_{2k}(\bar{B}_p^d, \ell_2^d).$$

Upper bound: We use the shorthand $\bar{\alpha} = \min\{\alpha, 1\}$. Let $1/2 > \varepsilon_1, \varepsilon_2 > 0$ be fixed, and set $\varepsilon := \varepsilon_1^{\bar{\alpha}} + \varepsilon_2$. Let $\mathcal{N} = \{g_1, \dots, g_n\}$ be a minimal ε_1-net of $\bar{B}_\alpha$ in the L_∞-norm. Further, let $\mathcal{M} = \{a_1, \dots, a_m\}$ be a minimal ε_2-net of $\bar{B}_p^d$ in the ℓ_2^d-norm.

Now, fix some ridge function $f : x \mapsto g(a \cdot x)$ in $\mathcal{R}_d^{\alpha,p}$, i.e., $\|g\|_{\mathrm{Lip}_\alpha} \le 1$ and $\|a\|_p \le 1$. Then, there is a function $g_i \in \mathcal{N}$ with $\|g - g_i\|_\infty \le \varepsilon_1$ and a vector $a_j \in \mathcal{M}$ with $\|a - a_j\|_2 \le \varepsilon_2$. We obtain

$$\begin{aligned}\|g(a \cdot x) - g_i(a_j \cdot x)\|_\infty &\le \sup_{x \in \bar{B}_2^d} |g(a \cdot x) - g(a_j \cdot x)| + |g(a_j \cdot x) - g_i(a_j \cdot x)| \\ &\le \sup_{x \in \bar{B}_2^d} |g|_{\bar{\alpha}} \cdot |a \cdot x - a_j \cdot x|^{\bar{\alpha}} + \|g - g_i\|_\infty \\ &\le \|a - a_j\|_2^{\bar{\alpha}} + \|g - g_i\|_\infty \le \varepsilon_1^{\bar{\alpha}} + \varepsilon_2 = \varepsilon.\end{aligned}$$

Hence, the set $\{x \to g(a \cdot x) : g \in \mathcal{N}, a \in \mathcal{M}\}$ is an ε-net of $\mathcal{R}_d^{\alpha,p}$ in $L_\infty(\Omega)$ with cardinality

$$\#\mathcal{N} \cdot \#\mathcal{M} = N_{\varepsilon_1}(\bar{B}_\alpha, L_\infty) \cdot N_{\varepsilon_2}(\bar{B}_p^d, \ell_2^d).$$

Consequently, $N_\varepsilon(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq \#\mathcal{N} \cdot \#\mathcal{M}$, and we conclude that

$$e_{2k}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_k(\bar{B}_p^d, \ell_2^d)^{\bar{\alpha}} + e_k(\bar{B}_\alpha, L_\infty).$$

□

Remark 3.5 In view of Proposition 3.2, it is easy to see that Theorem 3.4 remains valid if we replace the class $\mathcal{R}_d^{\alpha,p}$ by $\mathcal{R}_d^{\alpha,p,\kappa}$.

We exemplify the consequences of Theorem 3.4 by considering the case $p = 2$; for $0 < p < 2$, estimates would be similar. As the corollary below shows, entropy numbers of ridge functions asymptotically decay as fast as those of their profiles. In contrast to multivariate Lipschitz functions on Ω, the dimension d does not appear in the decay rate's exponent. It only affects how long we have to wait until the asymptotic decay becomes visible.

Corollary 3.6 *Let d be a natural number and $\alpha > 0$. For the entropy numbers of $\mathcal{R}_d^{\alpha,2}$ in $L_\infty(\Omega)$, we have*

$$\max(k^{-\alpha}, 2^{-k/d}) \lesssim e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \lesssim \begin{cases} 1, & k \leq c_\alpha d \log d, \\ k^{-\alpha}, & k \geq c_\alpha d \log d, \end{cases} \tag{3.2}$$

for some universal constant $c_\alpha > 0$ which does not depend on d.

Before we turn to the proof, let us note that (3.2) implies that

$$e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \asymp 1 \quad \text{if} \quad k \leq d,$$

and

$$e_k(\mathcal{R}_d^{\alpha,2}, L_\infty) \asymp k^{-\alpha} \quad \text{if} \quad k \geq c_\alpha d \ln d.$$

Hence, entropy numbers of ridge functions are guaranteed to decay like those of their profiles for $k \geq c_\alpha d \log d$—and surely behave differently for $k \leq d$.

Proof of Corollary 3.6 The lower bound in (3.2) follows from Theorem 3.4 combined with Lemma 2.5 and Lemma 3.1. The upper bounds are proven in the same manner, using the simple fact that for every $\alpha > 0$ there are two constants $c_\alpha, c'_\alpha > 0$, such that $k \geq c_\alpha d \log d$ implies that $2^{-\min\{\alpha,1\}k/d} \leq c'_\alpha k^{-\alpha}$. □

Summarizing this section, the classes of ridge functions with Lipschitz profiles of order α are essentially as compact as the class of univariate Lipschitz functions of order α. Consequently, when speaking in terms of metric entropy, these classes of functions must be much smaller than the class of multivariate Lipschitz functions of order α.

Remark 3.7 The reader who is interested in results on entropy numbers of other classes of ridge functions is referred to the recent work [24]. There, classes of sums of ridge functions are studied such that each sum of ridge functions forms a multivariate polynomial of some maximal degree.

4 Sampling Numbers of Ridge Functions

In light of Sect. 3, one is led to think that efficient sampling of ridge functions should be feasible. Moreover, their simple, two-component structure naturally suggests a two-step procedure: first, use a portion of the available function samples to identify either the profile or the direction; then, use the remaining samples to unveil the other component.

However, in Sect. 4.1, we learn that for ridge functions in the class $\mathcal{R}_d^{\alpha,p}$, sampling is almost as hard as sampling of general multivariate Lipschitz functions on the Euclidean unit ball. In particular, such two-step procedures as sketched above cannot work in an efficient manner. It needs additional assumptions on the ridge profiles or directions. We discuss this in Sect. 4.2.

4.1 Sampling of Functions in $\mathcal{R}_d^{\alpha,p}$

As usual, throughout the section, let $\alpha > 0$ be the Lipschitz smoothness of profiles, $s = \llbracket \alpha \rrbracket$ the order up to which derivatives exist, and let $0 < p \leq 2$ indicate the p-norm such that ridge directions are contained in the closed ℓ_p^d-ball.

The algorithms we use to derive upper bounds are essentially the same as those which are known to be optimal for general multivariate Lipschitz functions, albeit the ridge structure allows a slightly improved analysis, at least in the case $p < 2$.

Proposition 4.1 *Let $\alpha > 0$ and $0 < p \leq 2$. For $n \geq \binom{d+s}{s}$ sampling points, the nth sampling number is bounded from above by*

$$g_{n,d}^{lin}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq e_{k-\Delta}(\bar{B}_2^d, \ell_{p'}^d)^\alpha,$$

where $k = \lfloor \log n \rfloor + 2$, $\Delta = 1 + \lceil \log \binom{d+s}{s} \rceil$, and p' is the dual index of p.

Proof *Case $\alpha \leq 1$*: In this case, $s = 0$ and $\Delta = 1$. We choose sampling points $x_1, \ldots, x_{2^{k-2}}$ such that they form an ε-covering of $\bar{B}_2^d$ in $\ell_{p'}^d$. Given this covering, we construct (measurable) sets $U_1, \ldots, U_{2^{k-2}}$ such that $U_i \subseteq x_i + \varepsilon \bar{B}_{p'}^d$ for $i = 1, \ldots, 2^{k-2}$ and

$$\bigcup_{i=1}^{2^{k-2}} \left(x_i + \varepsilon \bar{B}^d_{p'}\right) = \bigcup_{i=1}^{2^{k-2}} U_i, \quad U_i \cap U_j = \emptyset \text{ for } i \neq j.$$

Now, we use piecewise constant interpolation: we approximate $f = g(a\cdot) \in \mathcal{R}^{\alpha,p}_d$ by $Sf := \sum_{i=1}^{2^{k-2}} f(x_i)\mathbb{1}_{U_i}$. Then,

$$\begin{aligned}\left\| f - Sf \right\|_\infty &= \sup_{i=1,\dots,2^{k-2}} \sup_{x\in U_i} \left| f(x) - f(x_i)\right| \\ &\leq \sup_{i=1,\dots,2^{k-2}} \sup_{x\in U_i} \left\| g \right\|_{\mathrm{Lip}_\alpha} \|a\|^\alpha_p \|x - x_i\|^\alpha_{p'} \leq \varepsilon^\alpha.\end{aligned}$$

The smallest ε is determined by the $(k-1)$st entropy number $e_{k-1}(\bar{B}^d_2, \ell^d_{p'})$. Consequently,

$$g^{\mathrm{lin}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \leq g^{\mathrm{lin}}_{2^{k-2},d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \leq e_{k-1}(\bar{B}^d_2, \ell^d_{p'})^\alpha.$$

Case $\alpha > 1$: Let the sampling points $x_1, \dots, x_{2^{k-\Delta-1}}$ and the sets $U_1, \dots, U_{2^{k-\Delta-1}}$ be as above. However, instead of piecewise constant interpolation, we apply on each of the sets $U_i \subseteq x_i + \varepsilon\bar{B}^d_{p'}$ a Taylor formula of order s around the center x_i.

That is, to approximate a given $f = g(a\cdot) \in \mathcal{R}^{\alpha,p}_d$, we set $Sf := \sum_{i=1}^{2^{k-\Delta-1}} T_{x_i,s} f \mathbb{1}_{U_i}$. Then, by Lemma 2.3 (ii), we have

$$\left\| f - Sf \right\|_\infty = \sup_{i=1,\dots,2^{k-\Delta-1}} \sup_{x\in U_i} \left| f(x) - T_{x_i,s} f(x)\right| \leq \frac{1}{s!}\|x - x_i\|^\alpha_{p'} \leq \varepsilon^\alpha.$$

It takes $2^{k-\Delta-1}\binom{d+s}{s} \leq n$ function values to approximate all the $T_{x_i,s}$ above up to arbitrary precision by finite-order differences, cf. [38].

The smallest ε is now determined by the $(k-\Delta)$th entropy number $e_{k-\Delta}(\bar{B}^d_2, \ell^d_{p'})$. We conclude

$$g^{\mathrm{lin}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \leq g^{\mathrm{lin}}_{2^{k-\Delta-1},d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \leq e_{k-\Delta}(\bar{B}^d_2, \ell^d_{p'})^\alpha.$$

□

We turn to an analysis of lower bounds for the classes $\mathcal{R}^{\alpha,p}_d$. Our strategy is to find "bad" directions which map, for a given budget $n \in \mathbb{N}$, all possible choices of n sampling points to a small range of $[-1, 1]$. There, we let the "fooling" profiles be zero; outside of that range, we let the profiles climb as steep as possible. Proposition 4.2 below states the lower bound that results from this strategy provided that the "bad" directions are given by some $\mathcal{M} \subseteq \bar{B}^d_p \setminus \{0\}$. We discuss appropriate choices of $\mathcal{M}$ later. Hereafter, we use the mapping $\Psi : \mathbb{R}^d \setminus \{0\} \to \mathbb{S}^{d-1}_2$ defined by $x \mapsto x/\|x\|_2$.

Proposition 4.2 *Let $\alpha > 0$, $0 < p \leq 2$, and $\mathcal{M} \subseteq \bar{B}_p^d \setminus \{0\}$. Then, for all natural numbers k and n with $n \leq 2^{k-1}$, we have*

$$g_{n,d}^{ada}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq c_\alpha \inf_{a \in \mathcal{M}} \|a\|_2^\alpha \cdot e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}.$$

The constant c_α depends only on α.

Proof Let us first describe the "fooling" profiles in detail. For each $a \in \mathcal{M}$ and $\varepsilon < 1$, we define a function

$$g_{a,\varepsilon}(t) = \vartheta_\alpha \big[(t - \|a\|_2(1 - \varepsilon^2/2))_+\big]^\alpha \tag{4.1}$$

on the interval $[-1, 1]$. The factor ϑ_α assures that $\|g_{a,\varepsilon}\|_{\mathrm{Lip}_\alpha[-1,1]} = 1$. Set $f_{a,\varepsilon}(x) = g_{a,\varepsilon}(a \cdot x)$. By construction, we have that $f_{a,\varepsilon} \in \mathcal{R}_d^{\alpha,p}$. Moreover, whenever $x \in \bar{B}_2^d$ and $a \in \mathcal{M}$ is such that

$$\varepsilon^2 < \big\|x - \Psi(a)\big\|_2^2, \tag{4.2}$$

then $\varepsilon^2 \leq 2 - 2(x \cdot \Psi(a))$ and hence

$$x \cdot a = \|a\|_2(x \cdot \Psi(a)) < \|a\|_2(1 - \varepsilon^2/2).$$

Therefore, (4.2) implies $f_{a,\varepsilon}(x) = 0$.

Now, let $n \leq 2^{k-1}$ and $S \in \mathcal{S}_n^{\mathrm{ada}}$ be an adaptive algorithm with a budget of n sampling points. Clearly, the first sampling point x_1 must have been fixed by S in advance. Then, let $x_2, \ldots, x_n$ be the sampling points, which S would choose when applied to the zero function. Furthermore, let $F(x_1, \ldots, x_n) \subseteq \mathcal{R}_d^{\alpha,p}$ denote the set of functions that make S choose the very points $x_1, \ldots, x_n$. Obviously, we have $f_{a,\varepsilon} \in F(x_1, \ldots, x_n)$ if (4.2) holds for every x_i, $i = 1, \ldots, n$. This is true for some $a \in \mathcal{M}$ if we choose $\varepsilon < e_k(\Psi(\mathcal{M}), \ell_2^d)$. For the respective function $f_{a,\varepsilon}$, we have in particular $N_S^{\mathrm{ada}}(f_{a,\varepsilon}) = 0$, and hence, $S[f_{a,\varepsilon}] = S[-f_{a,\varepsilon}]$. Consequently,

$$\begin{aligned} \max\big\{\|f_{a,\varepsilon} - S[f_{a,\varepsilon}]\|_\infty, \|-f_{a,\varepsilon} - S[-f_{a,\varepsilon}]\|_\infty\big\} &\geq \|f_{a,\varepsilon}\|_\infty \\ = g_{a,\varepsilon}(\|a\|_2) = c_\alpha \|a\|_2^\alpha \varepsilon^{2\alpha}, \end{aligned} \tag{4.3}$$

where $c_\alpha := 2^{-\alpha}\vartheta_\alpha$. Since ε has been chosen arbitrarily but less than $e_k(\Psi(\mathcal{M}), \ell_2^d)$, we are allowed to replace ε by $e_k(\Psi(\mathcal{M}), \ell_2^d)$ in (4.3) and get

$$\sup_{f \in \mathcal{R}_d^{\alpha,p}} \|f - S(f)\|_\infty \geq c_\alpha \inf_{a \in \mathcal{M}} \|a\|_2^\alpha \cdot e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}.$$

Taking the infimum over all algorithms $S \in \mathcal{S}_n^{\mathrm{ada}}$ yields

$$g_{n,d}^{\mathrm{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq c_\alpha \inf_{a \in \mathcal{M}} \|a\|_2^\alpha \, e_k(\Psi(\mathcal{M}), \ell_2^d)^{2\alpha}.$$

□

Theorem 4.3 *Let $\alpha > 0$, $s = \llbracket \alpha \rrbracket$, and $0 < p \leq 2$. For the classes $\mathcal{R}_d^{\alpha,p}$, we have the following bounds:*

(i) *The nth (linear) sampling number is bounded from above by*

$$g_{n,d}^{\text{lin}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq C_{p,\alpha} \begin{cases} 1, & n \leq 2d\binom{d+s}{s}, \\ \left[\frac{\log(1+d/\log n_1)}{\log n_1}\right]^{\alpha(1/\max\{1,p\}-1/2)}, & 2d\binom{d+s}{s} < n \leq 2^{d+1}\binom{d+s}{s}, \\ n^{-\alpha/d}\ d^{-\alpha(1/\max\{p,1\}-1/2)}, & n > 2^{d+1}\binom{d+s}{s}, \end{cases}$$

where $n_1 = n/[2\binom{d+s}{s}]$ and the constant $C_{p,\alpha}$ depends only on α and p.

(ii) *The nth (adaptive) sampling number is bounded from below by*

$$g_{n,d}^{\text{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq c_{p,\alpha} \begin{cases} 1, & n < d, \\ \left[\frac{\log_2(1+d/(2+\log_2 n))}{2+\log_2 n}\right]^{\alpha(1/p-1/2)}, & d \leq n < 2^{d-1}, \\ n^{-2\alpha/(d-1)}\ d^{-\alpha(1/p-1/2)}, & n \geq 2^{d-1}. \end{cases}$$

The constant $c_{p,\alpha}$ depends only on α and p.

Proof (i) The upper bound is a direct consequence of Proposition 4.1 and Lemma 2.5. Note that, for k and Δ as in Proposition 4.1, it holds true that $k - \Delta - 2 \leq \log n_1 \leq k - \Delta$. Note also that

$$\binom{d+s}{s}^{\alpha/d} \leq (1+s)^{s\alpha/d} d^{s\alpha/d} \leq ((1+s)e)^{s\alpha}$$

ensures that the constant $C_{p,\alpha}$ can be chosen independently of d and n.

(ii) *Case $n < d$.* Let $\mathcal{M} = \{\pm e_1, \ldots, \pm e_d\}$ be the set of positive and negative canonical unit vectors. Clearly, we have $\sharp\mathcal{M} = 2d$, and every two distinct vectors in $\mathcal{M}$ have mutual ℓ_2^d-distance equal to or greater than $\sqrt{2}$. Let k be the smallest integer such that $n \leq 2^{k-1}$; this implies $2^{k-1} < 2d$. Hence, whenever 2^{k-1} balls of radius ε cover the set $\mathcal{M}$, there is at least one ε-ball, which contains two elements from $\mathcal{M}$. In consequence, we have $2\varepsilon \geq \sqrt{2}$ and hence $e_k(\mathcal{M}, \ell_2^d) \geq \sqrt{2}/2$. By Proposition 4.2 and the fact that $\mathcal{M} = \Psi(\mathcal{M})$, we obtain

$$g_{n,d}^{\text{ada}}(\mathcal{R}_d^{\alpha,p}, L_\infty) \geq c_\alpha e_k(\mathcal{M}, \ell_2^d)^{2\alpha} \geq c_\alpha 2^{-\alpha}.$$

Case $d \leq n < 2^{d-1}$. For $m \leq d$, consider the subset of m-sparse vectors of the p-sphere,

$$\mathfrak{S}_{m,p}^{d-1} = \left\{x \in \mathbb{S}_p^{d-1} : \sharp\,\mathrm{supp}\,(x) = m\right\}.$$

Using the combinatorial construction of [16], cf. also [13], we know that there exist at least $(d/(4m))^{m/2}$ vectors in $\Psi(\mathfrak{S}^{d-1}_{m,p}) = \mathfrak{S}^{d-1}_{m,2}$ having mutual ℓ_2^d-distance greater than $1/\sqrt{2}$. Therefore, we have

$$\ell \le m/2\log(d/(4m)) \quad \Longrightarrow \quad e_\ell(\Psi(\mathfrak{S}^{d-1}_{m,p}), \ell_2^d) \ge \sqrt{2}/4. \tag{4.4}$$

Let k again be the smallest integer such that $n \le 2^{k-1}$. Hence, $k \le d$. Choose

$$m^* := \big\lfloor \min\{4k/\log(d/(4k)), k\}\big\rfloor \le k.$$

Because of $k > \log d$, we have $\min\{\log d, 4\} \le m^* \le d$. Write $\mathcal{M} = \mathfrak{S}^{d-1}_{m^*,p}$. If $k \le d/64$, then $\log(d/(4k)) \ge 4$ and $k \le m^*\log(d/(4k))/2 \le m^*\log(d/(4m^*))/2$. Hence, by (4.4), one has $e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d) \ge \sqrt{2}/4$. Consequently, by Proposition 4.2, it follows that

$$\begin{aligned} g^{\text{ada}}_{n,d}(\mathcal{R}^{\alpha,d}_d, L_\infty) &\ge c_\alpha (m^*)^{\alpha(1/2-1/p)} e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d)^{2\alpha} \\ &\ge c_\alpha 8^{-\alpha}4^{-\alpha(1/p-1/2)}\left[\frac{\log(d/(4k))}{k}\right]^{\alpha(1/p-1/2)} \\ &\ge c_\alpha 8^{-\alpha}8^{-\alpha(1/p-1/2)}\left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)} \\ &\ge c_{p,\alpha}\left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)}. \end{aligned}$$

On the other hand, if $d/64 < k \le d$, then $m^* = k$. By $\mathbb{S}_2^{k-1} \subset \Psi(\mathfrak{S}^{d-1}_{m^*,p}) \subset \mathbb{S}_2^{d-1}$ and Lemma 2.6, we have $e_k(\Psi(\mathfrak{S}^{d-1}_{m^*,p}), \ell_2^d) \asymp 1$. Proposition 4.2, together with $\log(1+d/k) < 8$ for $k > d/64$, implies

$$\begin{aligned} g^{\text{ada}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) \ge c'_\alpha k^{-\alpha(1/p-1/2)} &\ge c'_\alpha 8^{-\alpha(1/p-1/2)}\left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)} \\ &= c'_{p,\alpha}\left(\frac{\log(1+d/k)}{k}\right)^{\alpha(1/p-1/2)}. \end{aligned}$$

Case $n \ge 2^{d-1}$. Again, k is chosen such that $2^{k-2} < n \le 2^{k-1}$, which implies $k \ge d$. In this case, we choose $\mathcal{M} = \mathbb{S}_p^{d-1}$. By Lemma 2.6 and Proposition 4.2, we obtain

$$\begin{aligned} g^{\text{ada}}_{n,d}(\mathcal{R}^{\alpha,p}_d, L_\infty) &\ge c_\alpha\ d^{-\alpha(1/p-1/2)} e_k(\mathbb{S}_2^{d-1}, \ell_2^d)^{2\alpha} \\ &\ge c_\alpha d^{-\alpha(1/p-1/2)}\ (4n)^{-2\alpha/(d-1)} \\ &\ge c_\alpha 4^{-2\alpha} d^{-\alpha(1/p-1/2)} n^{-2\alpha/(d-1)}. \end{aligned}$$

This completes the proof. □

Remark 4.4 Consider the situation $p = 2$. For sampling numbers with $n \le 2^{d-1}$, we have

$$g_{n,d}^{\text{ada}}(\mathcal{R}_d^{\alpha,2}, L_\infty) \asymp 1.$$

For sampling numbers with $n \ge 2^{d+1}\binom{d+s}{s}$, we have

$$n^{-2\alpha/(d-1)} \lesssim g_{n,d}^{\text{ada}}(\mathcal{R}_d^{\alpha,2}, L_\infty) \lesssim n^{-\alpha/d}. \tag{4.5}$$

The upper estimate on sampling numbers is exactly the same as for multivariate Lipschitz functions from $\text{Lip}_\alpha(\Omega)$. Although there is a gap between lower and upper bound in (4.5), the factor $1/(d-1)$ in the exponent of the lower bound allows us to conclude that sampling of ridge functions in $\mathcal{R}_d^{\alpha,2}$ is nearly as hard as sampling of general Lipschitz functions from $\text{Lip}_\alpha(\Omega)$. Hence, we have the opposite situation to Sect. 3, where ridge functions in $\mathcal{R}_d^{\alpha,2}$ behave similarly to univariate Lipschitz functions.

Remark 4.5 Let us consider the modified ridge function classes $\tilde{\mathcal{R}}_d^{\alpha,p}$ and $\bar{\mathcal{R}}_d^{\alpha,p}$ defined by

$$\tilde{\mathcal{R}}_d^{\alpha,p} := \left\{f : [0,1]^d \to \mathbb{R} : f(x) = g(a\cdot x),\ \|g\|_{\text{Lip}_\alpha[0,1]} \le 1,\ \|a\|_p \le 1,\ a \ge 0\right\} \tag{4.6}$$

for $0 < p \le 1$, and

$$\begin{aligned}\bar{\mathcal{R}}_d^{\alpha,p} := \big\{f : \bar{B}_2^d \cap [0,1]^d &\to \mathbb{R} : f(x) = g(a\cdot x),\\ \|g\|_{\text{Lip}_\alpha[0,1]} &\le 1,\ \|a\|_p \le 1,\ a \ge 0\big\}\end{aligned} \tag{4.7}$$

for $0 < p \le 2$. Here, $a \ge 0$ means that all coordinates of a are nonnegative.

(i) In the recent paper [6], it has been shown that there is an adaptive algorithm, which attains a decay rate of $n^{-\alpha}$, for the worst-case L_∞-approximation error with respect to the class $\tilde{\mathcal{R}}_d^{\alpha,1}$, provided that $n \ge d$. In terms of adaptive sampling numbers (such that the feasible algorithms are adjusted to the domain $[0,1]^d$), this reads as

$$g_{n,d}^{\text{ada}}(\tilde{\mathcal{R}}_d^{\alpha,1}, L_\infty) \le C_\alpha n^{-\alpha}, \quad n \ge d. \tag{4.8}$$

At the same time, a careful inspection of the proofs of Propositions 4.1, 4.2, and Theorem 4.3 shows that the results can be carried over to the classes $\bar{\mathcal{R}}_d^{\alpha,p}$ for all $0 < p \le 2$. In particular, for $0 < p \le 1$, we have the lower bound

$$g_{n,d}^{\text{ada}}(\bar{\mathcal{R}}_d^{\alpha,p}, L_\infty) \ge c_{p,\alpha} n^{-2\alpha/(d-1)} d^{\alpha(1/2-1/p)}, \quad n \in \mathbb{N}. \tag{4.9}$$

The estimates (4.8) and (4.9) appear to be conflicting at first glance. We encounter the rather surprising phenomenon that enlarging the domain of the class of functions under consideration leads to better approximation rates. To understand this, let us briefly sketch the adaptive algorithm of [6]. For $f = g(a\cdot) \in \tilde{\mathcal{R}}_d^{\alpha,p}$ not the

zero function, the idea is to first sample along the diagonal of the first orthant, that is, at points $x = t(1, \dots, 1)$ with $t \in [0, 1]$. Importantly, it is guaranteed that we can take samples from the whole relevant range $[0, \|a\|_1]$ of the profile g of f. This in turn assures that, by sampling adaptively along the diagonal, we find a small range in $[0, \|a\|_1]$ where the absolute value of g' is strictly larger than 0. Then, the ridge direction a can be recovered in a similar way as in Sect. 4.2.
On the other hand, for the classes $\bar{\mathcal{R}}_d^{\alpha,p}$, this adaptive algorithm will not work. Assume we sample again along the (rescaled) diagonal. This time, we can be sure that we are able to reach every point in the intervall $[0, \|a\|_1/\sqrt{d}]$. But this interval is in most cases strictly included in the relevant interval $[0, \|a\|_2]$ for g. Hence, it is not guaranteed anymore that we sample the whole relevant range of g and find an interval on which g' is not zero.

(ii) Admittedly, the domain $\Omega = [0, 1]^d \cap B_2^d$ in (4.7) is a somewhat artificial choice in case of $p \leq 1$, whereas the cube $\Omega = [0, 1]^d$ seems natural. Conversely, the definition in (4.6) is not reasonable in the case $p > 1$, since then $a \cdot x$ might exceed the domain interval for g. However, $\Omega = [0, 1]^d \cap B_2^d$ is the natural choice for $p = 2$ in (4.7). In this situation, we suffer from the curse of dimensionality for adaptive algorithms using standard information, see Remark 4.4 and Theorem 5.1,(1) below. This shows that the condition $p \leq 1$ is essential in the setting of [6] and that (4.8) cannot be true for the class $\bar{\mathcal{R}}_d^{\alpha,2}$.

Remark 4.6 We are not aware of any results on the approximation of ridge functions when arbitrary *bounded, linear functionals* are admitted in the information map, see Sect. 2.2. It seems to be an open problem whether or not such *linear information* would lead to substantially better bounds for the worst-case error.

4.2 Recovery of Ridge Directions

At the beginning of Sect. 4, we have sketched two-step procedures for the recovery of ridge functions. In this section, we discuss under which conditions these two-step procedures are feasible within our setting. The adaptive algorithm of [6], which we have already discussed in Remark 4.5, first approximates the profile g. Unfortunately, we could already argue that this algorithm cannot work in our setting. There is an opposite approach in Fornasier et al. [12], which first tries to recover the ridge direction and conforms to our setting. Following the ideas of [2], the authors developed an efficient scheme using Taylor's formula to approximate ridge functions with C^s profile obeying certain integral condition on the modulus of its derivative. This condition was satisfied, for example, if $\left|g'(0)\right| \geq \kappa > 0$. In their approach, the smoothness parameter s had to be at least 2. Using a slightly different analysis, this scheme turns out to work for Lipschitz profiles of order $\alpha > 1$.

Before we turn to the analysis, let us sketch the Taylor-based scheme in more detail. As transposes of matrices and vectors appear frequently, for reasons of convenience, we write $a \cdot x = a^T x$ for the remainder of this subsection. Now, Taylor's formula in direction e_i yields

$$\begin{aligned} f(he_i) &= f(0) + h\nabla f(\xi_h^{(i)} e_i)^T e_i \\ &= g(0) + hg'(\xi_h^{(i)} a_i) a_i. \end{aligned}$$

Hence, we can expose the vector a, distorted by a diagonal matrix with components

$$\xi_h = \left(g'(\xi_h^{(1)} a_1), \ldots, g'(\xi_h^{(d)} a_d) \right)$$

on the diagonal. In total, we have to spend only $d+1$ function evaluations for that. Moreover, each of ξ_h's components can be pushed arbitrarily close to $g'(0)$. This gives an estimate $\hat{a}$ of $a/\|a\|_2$, along which we can now conduct classical univariate approximation. Effectively, one samples a distorted version of g given by

$$\tilde{g} : \left[-1, 1\right] \to \mathbb{R},\ t \mapsto f(t\hat{a}) = g\left(ta^T \hat{a} \right).$$

The approximation $\hat{g}$ obtained in this way, together with $\hat{a}$, forms the sampling approximation to f,

$$\hat{f}(x) = \hat{g}(\hat{a}^T x).$$

Observe that $\tilde{g}(\hat{a}^T x) = g(a^T \hat{a}\hat{a}^T x)$, so it is crucial that $\hat{a}\hat{a}^T$ spans a subspace, which is close to the one-dimensional subspace spanned by aa^T, in the sense that

$$\left\| a^T (I_d - \hat{a}\hat{a}^T) \right\|_2$$

has to be small. Importantly, this provides the freedom to approximate a only up to a sign. Finally, let us note that if the factor $g'(0)$ can become arbitrary small, the information we get through Taylor's scheme about a also becomes arbitrarily bad. Hence, for this approach to work, it is necessary to require $\left|g'(0)\right| \geq \kappa$.

Lemma 4.7 *Let* $0 < \beta \leq 1$, $0 < \kappa \leq 1$, *and* $\varepsilon > 0$. *Further, let* $\delta = \frac{\varepsilon \cdot \kappa}{2+\varepsilon}$ *and* $h = (\delta/2)^{1/\beta}$. *For any* $g \in \mathrm{Lip}_{1+\beta}^{\kappa}(\left[-1, 1\right])$ *and* $a \in \bar{B}_2^d$ *with* $a \neq 0$, *let* $f = g(a\cdot)$. *Set*

$$\tilde{a}_i = \frac{f(he_i) - f(0)}{h}, \quad i = 1, \ldots, d \tag{4.10}$$

and $\hat{a} = \tilde{a}/\|\tilde{a}\|_2$. *Then*

$$\left\| \operatorname{sign}(g'(0))\hat{a} - a/\|a\|_2 \right\|_2 \leq \varepsilon.$$

Proof By the mean value theorem of calculus, there exist $\xi_h^{(i)} \in \left[0, h\right]$ such that

$$\tilde{a}_i = g'(\xi_h^{(i)} a_i) a_i.$$

By Hölder continuity, we get

$$|g'(\xi_h^{(i)}a_i)-g'(0)|<2|g'|_\beta|a_i|^\beta|h|^\beta\le\delta$$

for all $i=1,\dots,d$. Let us observe that $\delta<\kappa$, and therefore, $\tilde{a}\neq 0$ and $\hat{a}$ is well defined. Set $\xi=(g'(\xi_h^{(i)}a_i))_{i=1}^d$. Then, we can write $\tilde{a}=\operatorname{diag}(\xi)a$. For the norm of $\tilde{a}$, we get

$$\begin{aligned}\|\tilde{a}\|_2&\le\|\operatorname{diag}(\xi)a-g'(0)a\|_2+|g'(0)|\|a\|_2\\&\le\max_{i=1,\dots,d}|g'(\xi_h^{(i)}a_i)-g'(0)|\|a\|_2+|g'(0)|\|a\|_2\\&\le(\delta+|g'(0)|)\|a\|_2.\end{aligned}$$

Analogously, by the inverse triangle inequality, $\|\tilde{a}\|_2\ge(|g'(0)|-\delta)\|a\|_2$. In particular,

$$\big|\|\tilde{a}\|_2/\|a\|_2-|g'(0)|\big|\le\delta.$$

Now, writing $\gamma=\operatorname{sign}(g'(0))$, we observe

$$\begin{aligned}\big\|\gamma\hat{a}-a/\|a\|_2\big\|_2&\le\big\|\gamma\hat{a}-|g'(0)|a/\|\tilde{a}\|_2\big\|_2+\big\||g'(0)|a/\|\tilde{a}\|_2-a/\|a\|_2\big\|_2\\&=\|\tilde{a}\|_2^{-1}\big(\|(\operatorname{diag}(\xi)-g'(0)I_d)\,a\|_2+\big||g'(0)|-\|\tilde{a}\|_2/\|a\|_2\big|\,\|a\|_2\big)\\&\le2\delta\|a\|_2/\|\tilde{a}\|_2\le2\delta/(|g'(0)|-\delta)\le2\delta/(\kappa-\delta)=\varepsilon.\end{aligned}$$

□

Having recovered the ridge direction, we manage to unveil the one-dimensional structure from the high-dimensional ambient space. In other words, recovery of the ridge direction is a *dimensionality reduction* step. What remains is the problem of sampling the profile, which can be done using standard techniques. In combination, this leads to the following result:

Theorem 4.8 *Let $\alpha>1$ and $0<\kappa\le1$.*

(i) *Let $n\le d-1$. Then $g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa},L_\infty)=g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,2,\kappa},L_\infty)=1$.*
(ii) *Let $n\ge d+1$. Then,*

$$c_\alpha\cdot n^{-\alpha}\le g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa},L_\infty)\le g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{\alpha,2,\kappa},L_\infty)\le C_\alpha(n-d)^{-\alpha}$$

with constants c_α and C_α, which depend on α only.

Proof (i) It is enough to show that $g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa},L_\infty)\ge1$ for $n\le d-1$. Let us assume that a given (adaptive) approximation method samples at $x_1,\dots,x_n$, and let us denote by L their linear span. Then, $\dim L\le n<d$, and we may find $a\in\mathbb{R}^d$ with $\|a\|_2=1$ orthogonal to all $x_1,\dots,x_n$. Finally, if we define $g(t)=t$, we

obtain

$$\begin{aligned} 1 &= \|g(a^T\cdot)\|_\infty \\ &\le \frac{1}{2}\cdot\left\{\|g(a^T\cdot)-S_n(g(a^T\cdot))\|_\infty + \|-g(a^T\cdot)-S_n(-g(a^T\cdot))\|_\infty\right\} \\ &\le g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty). \end{aligned}$$

(ii) Fix some $0<\varepsilon<1$. Let $\hat{a}$ denote the reconstruction of a obtained by Lemma 4.7, which uses $d+1$ sampling points of f. We estimate g by sampling the distorted version

$$\tilde{g}:[-1,1]\to\mathbb{R},\ t\mapsto f(t\hat{a})=g\left(ta^T\hat{a}\right).$$

Re-using the value $g(0)$ which we have already employed for the recovery of a, we spend $k=n-d\ge 1$ sampling points and obtain a function $\hat{g}$ with $\|\hat{g}-\tilde{g}\|_\infty\le\varepsilon := C'_\alpha k^{-\alpha}\|\tilde{g}\|_{\mathrm{Lip}_\alpha}$.

Now write $\hat{f}(x)=\hat{g}(\hat{a}^Tx)$ as our approximation to f. To control the total approximation error, observe that

$$|\hat{f}(x)-f(x)|\le\left|\hat{g}(\hat{a}^Tx)-\tilde{g}(\hat{a}^Tx)\right|+\left|\tilde{g}(\hat{a}^Tx)-g(a^Tx)\right| =: E_1+E_2.$$

For the first error term E_1, we immediately get

$$E_1\le\|\hat{g}-\tilde{g}\|_\infty\le\varepsilon=C'_\alpha\|\tilde{g}\|_{\mathrm{Lip}_\alpha}k^{-\alpha}\le C'_\alpha k^{-\alpha}$$

as $\|\tilde{g}\|_{\mathrm{Lip}_\alpha}\le\|a\|_2\,\|g\|_{\mathrm{Lip}_\alpha}\le 1$.

For the second error term, note that

$$\begin{aligned} E_2 &= \left|g(a^T\hat{a}\hat{a}^Tx)-g(a^Tx)\right|\le\|g\|_{\mathrm{Lip}_\alpha}\left\|a^T(I_d-\hat{a}\hat{a}^T)\right\|_2\,\|x\|_2 \\ &\le\|g\|_{\mathrm{Lip}_\alpha}\,\|x\|_2\,\|a\|_2\,\left\|a^T/\|a\|_2\,(I_d-\hat{a}\hat{a}^T)\right\|_2. \end{aligned}$$

We do not know the exact value of the subspace stability term $\|a^T/\|a\|_2\,(I_d-\hat{a}\hat{a}^T)\|_2$. But because $\hat{a}\hat{a}^T$ is the identity in the direction of $\hat{a}$, we have the estimate

$$\begin{aligned} \left\|a^T/\|a\|_2\,(I_d-\hat{a}\hat{a}^T)\right\|_2 &= \left\|\left(a/\|a\|_2-\operatorname{sign}(g'(0))\hat{a}\right)^T\,(I_d-\hat{a}\hat{a}^T)\right\|_2 \\ &\le\|I_d-\hat{a}\hat{a}^T\|_{2\to2}\,\left\|a/\|a\|_2-\operatorname{sign}(g'(0))\hat{a}\right\|_2 \\ &\le\varepsilon. \end{aligned}$$

For the last inequality, we have used Lemma 4.7 and the fact that $\|I_d-\hat{a}\hat{a}^T\|_{2\to2}\le 1$. As a consequence,

$$E_2\le\|x\|_2\,\|a\|_2\,\|g\|_{\mathrm{Lip}_\alpha}\,\varepsilon\le\varepsilon.$$

Putting everything together, we conclude

$$\|\hat{f} - f\|_\infty \le 2\varepsilon \le 2C'_\alpha k^{-\alpha}.$$

Let us turn to the lower bound. Assume we are given a feasible approximation method S_n that samples at points $\{x_1, \dots, x_n\} \subset \Omega$. Let $\psi_{k,b}$ be as in the proof of Proposition 3.2. There is an interval $I' \subset I = [\pi/4 - 1/5, \pi/4 + 1/5]$ of length $|I'| = 1/(5n)$ such that I' does not contain any of the first coordinates of $x_1, \dots, x_n$; in other words, it is disjoint with $\{x_1 \cdot e_1, \dots, x_n \cdot e_1\}$, where $e_1 = (1, 0, \dots, 0)$ is the first canonical unit vector. Furthermore, let b be the center of I', put $\psi = \psi_{2n,b}$, and $a = e_1$. Finally, with γ as in (3.1), we write

$$\begin{aligned} f(x) &= \sin(x \cdot e_1),\\ f_+(x) &= \sin(x \cdot e_1) + (1-\gamma)\psi(x \cdot e_1),\\ f_-(x) &= \sin(x \cdot e_1) - (1-\gamma)\psi(x \cdot e_1). \end{aligned}$$

As $S_n(f) = S_n(f_+) = S_n(f_-)$ and all the three functions are in $\mathcal{R}_d^{\alpha,2,\kappa}$, we may use the triangle inequality

$$\begin{aligned} \|(1-\gamma)\psi\|_\infty &= \|(1-\gamma)\psi(e_1\cdot)\|_\infty\\ &\le \frac{1}{2}\Big\{\|(1-\gamma)\psi(e_1\cdot) + f - S_n(f)\|_\infty\\ &\quad + \|(1-\gamma)\psi(e_1\cdot) - [f - S_n(f)]\|_\infty\Big\}\\ &= \frac{1}{2}\Big\{\|f_+ - S_n(f_+)\|_\infty + \|f_- - S_n(f_-)\|_\infty\Big\} \end{aligned}$$

to conclude that

$$g_{n,d}(\mathcal{R}_d^{\alpha,2,\kappa}, L_\infty) \gtrsim n^{-\alpha},$$

with a constant depending only on α. □

Remark 4.9 Let us briefly comment why we assume $g'(0) \ge \kappa$ and not $g'(t_0) \ge \kappa$ for some arbitrary but known $t_0 \in [-1, 1]$. Let $x \in \bar{B}_2^d$ be some arbitrary sampling point (taken, e.g., uniformly at random in $\bar{B}_2^d$). Since the only a-priori information is $\|a\|_2 \le 1$, by the concentration of measure phenomenon, the inner product $a \cdot x$ will most likely be close to zero when d is large. Hence, to exploit $g'(t_0) \ge \kappa$ for some $t_0 \ne 0$, we effectively have to know the vector a beforehand.

Remark 4.10 Once we have control of the derivative at the origin, recovery of the ridge direction and approximation of the ridge profile can be addressed independently. Formula (4.10) is based on the simple observation that

$$\frac{\partial f}{\partial x_i}(0) = g'(0)a_i = g'(0)\langle a, e_i\rangle$$

might be well approximated by first-order differences. Furthermore, this holds also for every other direction $\varphi \in \mathbb{S}_2^{d-1}$, i.e.,

$$\frac{\partial f}{\partial \varphi}(0) = g'(0)\langle a, \varphi\rangle$$

can be approximated by differences

$$\frac{f(h\varphi) - f(0)}{h}.$$

Taking the directions $\varphi_1, \dots, \varphi_{m_\Phi}$ at random (and appropriately normalized), one can approximate the scalar products $\{\langle a, \varphi_i\rangle\}_{i=1}^{m_\Phi}$. Finally, if one assumes that $a \in \bar{B}_p^d$ for $0 < p \leq 1$, one can recover a good approximation to a by the *sparse recovery* methods of the modern area of *compressed sensing*. This approach has been investigated in [12].

Although the algorithms of compressed sensing involve random matrices, once a random matrix with good sensing properties (typically with small constants of their restricted isometry property) is fixed, the algorithms become fully deterministic. This allows one to transfer the estimates of [12] into an upper bound for the deterministic worst-case error $g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty)$.

Let $0 < p \leq 1$ and

$$c\kappa^{-\frac{2p}{2-p}} \log d \leq m_\Phi \leq Cd$$

for two universal positive constants c, C. It follows from the results of [12] that drawing the directions $\varphi_1, \dots, \varphi_{m_\Phi}$ *once* yields with high probability a deterministic algorithm that needs $n > m_\Phi$ sampling points to recover any function $f \in \mathcal{R}_d^{2,p,\kappa}$ up to precision

$$\left[\frac{m_\Phi}{\log(d/m_\Phi)}\right]^{1/2-1/p} + (n - m_\Phi)^{-2}.$$

If $1/p \leq 5/2$ and $c'\kappa^{-\frac{2p}{2-p}} \log d \leq n \leq C'd$, this implies that

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty) \lesssim \left[\frac{n}{\log(d/n)}\right]^{1/2-1/p},$$

and the same estimate holds if $1/p > 5/2$ and $c'\kappa^{-\frac{2p}{2-p}} \log d \leq n \leq c''(\log d)^{\frac{1/p-1/2}{1/p-5/2}}$. Finally, if $c''(\log d)^{\frac{1/p-1/2}{1/p-5/2}} \leq n \leq C'd$, we obtain

$$g_{n,d}^{\mathrm{lin}}(\mathcal{R}_d^{2,p,\kappa}, L_\infty) \lesssim n^{-2}.$$

5 Tractability Results

The field of information-based complexity [26] deals with a family of properties of so-called *tractability*, which allow one to classify ridge function sampling by degrees of difficulty. With regard to these properties, the studied ridge function classes are surprisingly rich. We run across almost the whole hierarchy of degrees of tractability if we vary the problem parameters α and p, or add the constraint on the profiles' first derivative in the origin.

Let us briefly introduce the standard properties of tractability. We say that a problem is *polynomially tractable* if its information complexity $n(\varepsilon, d)$ is bounded polynomially in ε^{-1} and d; i.e., there exist numbers $C, r, q > 0$ such that

$$n(\varepsilon, d) \leq C\, \varepsilon^{-r}\, d^q \quad \text{for all } \ 0 < \varepsilon < 1 \text{ and all } d \in \mathbb{N}.$$

A problem is called *quasi-polynomially tractable* if there exist two constants $C, t > 0$ such that

$$n(\varepsilon, d) \leq C \exp(t(1 + \ln(1/\varepsilon))(1 + \ln d)). \tag{5.1}$$

It is called *weakly tractable* if

$$\lim_{1/\varepsilon + d \to \infty} \frac{\log n(\varepsilon, d)}{1/\varepsilon + d} = 0; \tag{5.2}$$

i.e., the information complexity $n(\varepsilon, d)$ depends exponentially neither on $1/\varepsilon$ nor on d.

We say that a problem is *intractable* if (5.2) does not hold. If for some fixed $0 < \varepsilon < 1$, the number $n(\varepsilon, d)$ is an exponential function in d, then a problem is, of course, intractable. In that case, we say that the problem suffers from *the curse of dimensionality*. To make it precise, we face the curse if there exist positive numbers c, ε_0, γ such that

$$n(\varepsilon, d) \geq c(1 + \gamma)^d \qquad \text{for all } \ 0 < \varepsilon \leq \varepsilon_0 \text{ and infinitely many } d \in \mathbb{N}.$$

In the language of IBC, Theorems 4.3 and 4.8 now read as follows:

Theorem 5.1 *Consider the problem of ridge function sampling as defined in Sect. 2.2. Assume that ridge profiles have at least Lipschitz smoothness $\alpha > 0$; further, assume that ridge directions are contained in the closed ℓ_p^d-unit ball for $p \in (0, 2]$. Then, sampling of ridge functions in the class $\mathcal{R}_d^{\alpha,p}$*

(1) suffers from the curse of dimensionality if $p = 2$ and $\alpha < \infty$,
(2) never suffers from the curse of dimensionality if $p < 2$,
(3) is intractable if $p < 2$ and $\alpha \leq \frac{1}{1/p - 1/2}$,
(4) is weakly tractable if $p < 2$ and $\alpha > \frac{1}{1/\max\{1,p\} - 1/2}$,
(5) is quasi-polynomially tractable if $\alpha = \infty$,
(6) and with positive first derivatives of the profiles in the origin it is polynomially tractable, no matter what the values of α and p are.

To prove Theorem 5.1, we translate Theorem 4.3 into bounds on the information complexity

$$n(\varepsilon, d) = \min\{n \in \mathbb{N} : \ g_{n,d}(\mathcal{R}_d^{\alpha,p}, L_\infty) \leq \varepsilon\}.$$

Lemma 5.2 *Let $p < 2$ and $\alpha > 0$. Set $\eta = \alpha(1/2 - 1/p') = \alpha(1/\max\{1, p\} - 1/2)$ and define*

$$\varepsilon_1^U := C_{p,\alpha} \left[\frac{\log(1 + d/\log d)}{\log d}\right]^{\eta}, \qquad \varepsilon_2^U := C_{p,\alpha} \left(\frac{1}{d}\right)^{\eta}.$$

Then, there are positive constants C_0 and C_1 such that

$$\log n(\varepsilon, d) \le C_0 + C_1 \begin{cases} \log d, & \varepsilon_1^U \le \varepsilon \le 1, \\ \log d \cdot (1/\varepsilon)^{1/\eta}, & \varepsilon_2^U \le \varepsilon < \varepsilon_1^U, \\ \log(1/\varepsilon) \cdot (1/\varepsilon)^{1/\eta}, & \varepsilon < \varepsilon_2^U. \end{cases}$$

The constants depend only on p and α.

Lemma 5.3 *Let $p < 2$ and $\alpha > 0$. Set*

$$\varepsilon_1^L := c_{p,\alpha} \left[\frac{\log(1 + d/\log d)}{\log d}\right]^{\alpha(1/p-1/2)}, \quad \varepsilon_2^L := c_{p,\alpha} \left(\frac{1}{d}\right)^{\alpha(1/p-1/2)},$$
$$\varepsilon_3^L := 4^{-\alpha}\varepsilon_2^L.$$

Then, there are universal constants c_0, c_1, which depend only on p and α, such that

$$\log n(\varepsilon, d) \ge c_0 + c_1 (1/\varepsilon)^{\alpha^{-1}(1/p-1/2)^{-1}}$$

for $\varepsilon_3^L \le \varepsilon < \varepsilon_1^L$.

Proof of Theorem 5.1 (1). For $n \le 2^{d-2}$, the lower bound in Theorem 4.3 gives

$$g_{n,d}(\mathcal{R}_d^{\alpha,2}, L_\infty) \ge c_{p,\alpha} =: \varepsilon_0.$$

Hence, $n(\varepsilon, d) \ge 2^{d-2}$ for all $\varepsilon < \varepsilon_0$, and we have the curse of dimensionality.

(2). Since $\alpha_1 > \alpha_2$ implies $\mathcal{R}_d^{\alpha_1,p} \subseteq \mathcal{R}_d^{\alpha_2,p}$, we can w.l.o.g. assume $\alpha \le 1$. We choose an arbitrary $\varepsilon_2^U \le \varepsilon \le 1$. By Lemma 5.2,

$$n(\varepsilon, d) \le 2^{C_0} d^{C_1 \varepsilon^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}}.$$

By our assumption $\varepsilon \ge \varepsilon_2^U$, this is true for all natural $d > (C_{p,\alpha}/\varepsilon)^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}$. Hence, the curse of dimensionality does not occur.

(3). Set $\gamma = \alpha(1/p - 1/2)$. Assume $d \to \infty$ and $\varepsilon_3^L \le \varepsilon < \varepsilon_2^L$. The latter implies

$$\left(\frac{c_{p,\alpha}}{4^\alpha}\right)^{1/\gamma} (1/\varepsilon)^{1/\gamma} \le d < c_{p,\alpha}^{1/\gamma} (1/\varepsilon)^{1/\gamma}.$$

This yields

$$\frac{\log_2 n(\varepsilon,d)}{d+1/\varepsilon} \geq \frac{c_0}{d+1/\varepsilon} + c_1 \frac{(1/\varepsilon)^{1/\gamma}}{c_{p,\alpha}^{1/\gamma}(1/\varepsilon)^{1/\gamma}+1/\varepsilon}.$$

Assuming that $\alpha \leq 1/(1/p-1/2)$, we have $\gamma \leq 1$ and thus $1/\varepsilon \leq (1/\varepsilon)^{1/\gamma}$. We conclude that

$$\frac{\log n(\varepsilon,d)}{d+1/\varepsilon} \geq \frac{c_1}{c_{p,\alpha}^{1/\gamma}+1} > 0.$$

Consequently, the problem is not weakly tractable and thus is intractable.

(4). Set $x = 1/\varepsilon + d$. By Lemma 5.2 and $1/\varepsilon \leq x$, $d \leq x$, we have

$$\log n(\varepsilon,d) \leq C_0 + C_1 \log(x) x^{\alpha^{-1}(1/\max\{1,p\}-1/2)^{-1}}.$$

Now, if $\alpha > \frac{1}{1/\max\{1,p\}-1/2}$, then $\lim_{x\to\infty} x^{-1}\log n(\varepsilon,d) = 0$.

(5). By embedding arguments, it is enough to consider the class $\mathcal{R}_d^{\infty,2}$. We approximate the function $f \in \mathcal{R}_d^{\infty,2}$ via the Taylor polynomial $T_{s,0}f(x)$ in $x^0 = 0$. Lemma 2.3, (ii) gives for every $s \in \mathbb{N}$ the bound

$$\|f - T_{s,0}f\|_\infty \leq \frac{2}{s!}.$$

Let $\varepsilon > 0$ be given, and let $s \in \mathbb{N}$ be the smallest integer such that $2/s! \leq \varepsilon$. Then, $(s-1)! \leq 2/\varepsilon$, and therefore $[(s-1)/e]^{s-1} \leq (s-1)! \leq 2/\varepsilon$. This gives

$$(s-1)\ln((s-1)/e) \leq \ln(2/\varepsilon). \tag{5.3}$$

We know from [38] that it requires $\binom{s+d}{s}$ function values to approximate the Taylor polynomial up to arbitrary (but fixed) precision. Hence, using (5.3), we see that there is a constant $t > 0$ such that

$$\ln n(\varepsilon,d) \leq s\ln(e(d+1)) \leq t(1+\ln(1/\varepsilon))(1+\ln d),$$

which is (5.1).

(6). From Theorem 4.8, we can immediately conclude $\varepsilon^{-1/\alpha} \lesssim n(\varepsilon,d) \lesssim \varepsilon^{-1/\alpha}$, where the constants behind "$\lesssim$" behave polynomially in d. Consequently, sampling of ridge functions in $\mathcal{R}_d^{\alpha,2,\kappa}$ is polynomially tractable. □

By Lemma 2.1, we know that $\mathcal{R}_d^{\infty,2}$ is a subclass of the unit ball in $C^\infty(\Omega)$. Besides, we know that approximation using function values is quasi-polynomially tractable in $\mathcal{R}_d^{\infty,2}$, see Theorem 5.1. What is the respective tractability level in $C^\infty(\Omega)$? Or, to put it differently: how much do we gain by imposing a ridge structure in $C^\infty(\Omega)$? The seminal paper [27] tells us that approximation in $C^\infty([0,1]^d)$ suffers from the curse of dimensionality when norming the space in the way we did in (2.1). In contrast, we will show that sampling in $C^\infty(\Omega)$ is still weakly tractable. This is not too much of a surprise. Due to the concentration of measure phenomenon, the Euclidean unit ball's volume gets "very small" in high dimensions d; its measure scales like $(2\pi e/d)^{d/2}$. Anyhow, the result suggests that one still benefits from supposing a ridge structure; infinitely differentiable ridge functions from $\mathcal{R}_d^{\infty,2}$ probably can be approximated more easily than general functions from the unit ball of $C^\infty(\Omega)$. This is not guaranteed, however, because we do not show that one cannot get anything better than weak tractability for the sampling of functions in the unit ball of $C^\infty(\Omega)$.

Theorem 5.4 *The sampling problem for $C^\infty(\Omega)$, where the error is measured in $L_\infty(\Omega)$, is weakly tractable.*

Proof Applying Lemma 2.3, (i) together with (2.3), we obtain for any $f \in C^\infty(\Omega)$ with $\|f\|_{C^\infty(\Omega)} \leq 1$ and every $s \in \mathbb{N}$ the relation

$$\begin{aligned} |f(x) - T_{s,0}f(x)| &\leq \frac{2}{(s-1)!}\|x\|_1^s, \quad x \in \Omega, \\ &\leq \frac{2d^{s/2}}{(s-1)!}. \end{aligned}$$

Let $s \in \mathbb{N}$ be the smallest integer such that $2d^{s/2}/(s-1)! \leq \varepsilon$. This leads to

$$\frac{1}{\sqrt{d}}\left(\frac{s-2}{e\sqrt{d}}\right)^{s-2} \leq \frac{(s-2)!}{d^{\frac{s-1}{2}}} \leq \frac{2}{\varepsilon},$$

which implies

$$(s-2)\ln\left(\frac{s-2}{e\sqrt{d}}\right) \leq \ln(2/\varepsilon) + \frac{1}{2}\ln(d). \tag{5.4}$$

To approximate the Taylor polynomial $T_{s,0}f$ with arbitrary precision (uniformly in f), we need $\binom{d+s}{s}$ function values, see [38, p. 4]. Let us distinguish two cases. If $(s-2) \leq e^2\sqrt{d}$, we obtain

$$\ln n(\varepsilon, d) \leq s\ln(e(d+1)) \leq (e^2\sqrt{d}+2)\cdot\ln(e(d+1))$$

and hence (5.2). If $s-2 > e^2\sqrt{d}$, then (5.4) yields $s-2 \leq \ln(2/\varepsilon) + \ln(d)$. Thus,

$$\ln n(\varepsilon, d) \leq s\ln(e(d+1)) \leq (\ln(2/\varepsilon) + \ln(d) + 2)\cdot\ln(e(d+1)),$$

and again (5.2) holds true. This establishes weak tractability. □

Remark 5.5 (i) The result in Theorem 5.4 is also a consequence of the arguments in [19, Sections 5.2, 5.3, and Section 6] by setting $L_{j,d} = d^{j/2}$.

(ii) Recently, Vybíral [38] showed that there is quasi-polynomial tractability if one replaces the classical norm $\sup_{\gamma\in\mathbb{N}_0^d}\|D^\gamma f\|_\infty$ by $\sup_{k\in\mathbb{N}_0}\sum_{|\gamma|=k}\|D^\gamma f\|_\infty/\gamma!$ in $C^\infty([0,1]^d)$. In contrast to that, Theorem 5.4 shows weak tractability for the classical norm on the unit ball.

Acknowledgments The authors would like to thank Aicke Hinrichs, Erich Novak, and Mario Ullrich for pointing out relations to the paper [19], as well as Sjoerd Dirksen, Thomas Kühn, and Winfried Sickel for useful comments and discussions. The last author acknowledges the support by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin. The last author was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education.

References

1. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data. Springer, Heidelberg (2011)
2. Buhmann, M.D., Pinkus, A.: Identifying linear combinations of ridge functions. Adv. Appl. Math. **22**, 103–118 (1999)
3. Candés, E.J.: Harmonic analysis of neural networks. Appl. Comput. Harmon. Anal. **6**, 197–218 (1999)
4. Candés, E.J., Donoho, D.L.: Ridgelets: a key to higher-dimensional intermittency? Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **357**, 2495–2509 (1999)
5. Carl, B., Stefani, I.: Entropy, Compactness and the Approximation of Operators. Cambridge Tracts in Mathematics, vol. 98. Cambridge University Press, Cambridge (1990)
6. Cohen, A., Daubechies, I., DeVore, R.A., Kerkyacharian, G., Picard, D.: Capturing ridge functions in high dimensions from point queries. Constr. Approx. **35**, 225–243 (2012)
7. Creutzig, J., Dereich, S., Müller-Kronbach, T., Ritter, K.: Infinite-dimensional quadrature and approximation of distributions. Found. Comput. Math. **9**, 391–429 (2009)
8. Cucker, F., Zhou, D.-X.: Learning theory: an approximation theory viewpoint. Cambridge Monographs on Applied and Computational Mathematics, vol. 24. Cambridge University Press, Cambridge (2007)
9. DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Springer, Berlin (1993)
10. Edmunds, D.E., Triebel, H.: Function Spaces, Entropy Numbers, Differential Operators. Cambridge Tracts in Mathematics, vol. 120. Cambridge University Press, Cambridge (1996)
11. Flad, H.J., Hackbusch, W., Khoromskij, B.N., Schneider, R.: Concepts of data-sparse tensor-product approximation in many-particle modeling. In: Olshevsky, V., Tyrtyshnikov, E. (eds.) Matrix Methods: Theory, Algorithms and Applications. World Scientific, Singapore (2010)
12. Fornasier, M., Schnass, K., Vybíral, J.: Learning functions of few arbitrary linear parameters in high dimensions. Found. Comput. Math. **12**, 229–262 (2012)
13. Foucart, S., Pajor, A., Rauhut, H., Ullrich, T.: The Gelfand widths of lp-balls for $0 < p \leq 1$. J. Complexity **26**, 629–640 (2010)
14. Friedman, J.H., Stuetzle, W.: Projection pursuit regression. J. Am. Stat. Assoc. **76**, 817–823 (1981)
15. Golubev, G.K.: Asymptotically minimax estimation of a regression function in an additive model. Problemy Peredachi Informatsii **28**, 101–112 (1992)
16. Graham, R., Sloane, N.: Lower bounds for constant weight codes. IEEE Trans. Inform. Theory **26**, 37–43 (1980)
17. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, New York (2001)
18. Hinrichs, A., Mayer, S.: Entropy numbers of spheres in Banach and quasi-Banach spaces. University of Bonn, preprint
19. Hinrichs, A., Novak, E., Ullrich, M., Woźniakowski, H.: The curse of dimensionality for numerical integration of smooth functions II. J. Complex. **30**, 117–143 (2014)
20. Hristache, M., Juditsky, A., Spokoiny, V.: Direct estimation of the index coefficient in a single-index model. Ann. Stat. **29**, 595–623 (2001)
21. Kühn, T.: A lower estimate for entropy numbers. J. Approx. Theory **110**, 120–124 (2001)
22. Logan, B.P., Shepp, L.A.: Optimal reconstruction of a function from its projections. Duke Math. J. **42**, 645–659 (1975)
23. Lorentz, G., von Golitschek, M., Makovoz, Y.: Constructive Approximation: Advanced Problems. Volume 304 of Grundlehren der Mathematischen Wissenschaften, Springer, Berlin (1996)
24. Maiorov, V.: Geometric properties of the ridge manifold. Adv. Comput. Math. **32**, 239–253 (2010)
25. Novak, E., Triebel, H.: Function spaces in Lipschitz domains and optimal rates of convergence for sampling. Constr. Approx. **23**, 325–350 (2006)
26. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume I: Linear Information. EMS Tracts in Mathematics, vol. 6, Eur. Math. Soc. Publ. House, Zürich (2008)
27. Novak, E., Woźniakowski, H.: Approximation of infinitely differentiable multivariate functions is intractable. J. Complex. **25**, 398–404 (2009)

28. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12, Eur. Math. Soc. Publ. House, Zürich (2010)
29. Paskov, S., Traub, J.: Faster evaluation of financial derivatives. J. Portf. Manag. **22**, 113–120 (1995)
30. Pinkus, A.: Approximating by ridge functions. In: Le Méhauté, A., Rabut, C., Schumaker, L.L. (eds.) Surface Fitting and Multiresolution Methods, pp. 279–292. Vanderbilt University Press, Nashville (1997)
31. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica **8**, 143–195 (1999)
32. Raskutti, G., Wainwright, M.J., Yu, B.: Minimax-optimal rates for sparse additive models over kernel classes via convex programming. J. Mach. Learn. Res. **13**, 389–427 (2012)
33. Schütt, C.: Entropy numbers of diagonal operators between symmetric Banach spaces. J. Approx. Theory **40**, 121–128 (1984)
34. Schwab, C., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numerica **20**, 291–467 (2011)
35. Traub, J., Wasilkowski, G., Woźniakowski, H.: Information-Based Complexity. Academic Press, New York (1988)
36. Triebel, H.: Fractals and Spectra. Birkhäuser, Basel (1997)
37. Tyagi, H., Cevher, V.: Active learning of multi-index function models. In: Bartlett, P., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1475–1483. Curran Associates, Red Hook (2012)
38. Vybíral, J.: Weak and quasi-polynomial tractability of approximation of infinitely differentiable functions. J. Complex. **30**, 48–55 (2014)

Non-asymptotic Analysis of ℓ_1-norm Support Vector Machines

Anton Kolleck, Jan Vybíral

Abstract

Support Vector Machines (SVM) with ℓ_1 penalty became a standard tool in analysis of highdimensional classification problems with sparsity constraints in many applications including bioinformatics and signal processing. Although SVM have been studied intensively in the literature, this paper has to our knowledge first non-asymptotic results on the performance of ℓ_1-SVM in identification of sparse classifiers. We show that a d-dimensional s-sparse classification vector can be (with high probability) well approximated from only $O(s\log(d))$ Gaussian trials. The methods used in the proof include concentration of measure and probability in Banach spaces.

Index Terms

Support vector machines, compressed sensing, machine learning, regression analysis, signal reconstruction, classification algorithms, functional analysis, random variables

I. Introduction

A. Support Vector Machines

Support vector machines (SVM) are a group of popular classification methods in machine learning. Their input is a set of data points $x_1,\dots,x_m \in \mathbb{R}^d$, each equipped with a label $y_i \in \{-1,+1\}$, which assigns each of the data points to one of two groups. SVM aims for binary linear classification based on separating hyperplane between the two groups of training data, choosing a hyperplane with separating gap as large as possible.

Since their introduction by Vapnik and Chervonenkis [27], the subject of SVM was studied intensively. We will concentrate on the so-called soft margin SVM [8], which allow also for misclassification of the training data are the most used version of SVM nowadays.

In its most common form (and neglecting the bias term), the soft-margin SVM is a convex optimization program

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi\in\mathbb{R}^m}} \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^m \xi_i \quad \text{subject to} \quad y_i\langle x_i, w\rangle \geq 1-\xi_i \quad \text{and} \quad \xi_i \geq 0 \tag{I.1}$$

for some tradeoff parameter $\lambda > 0$ and so called slack variables ξ_i. It will be more convenient for us to work with the following equivalent reformulation of (I.1)

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^m [1-y_i\langle x_i, w\rangle]_+ \quad \text{subject to} \quad \|w\|_2 \leq R, \tag{I.2}$$

where $R > 0$ gives the restriction on the size of w. We refer to monographs [25], [28], [29] and references therein for more details on SVM and to [13, Chapter B.5] and [9, Chapter 9] for a detailed discussion on dual formulations.

B. ℓ_1-SVM

As the classical SVM (I.1) and (I.2) do not use any pre-knowledge about w, one typically needs to have more training data than the underlying dimension of the problem, i.e. $m \gg d$. Especially in analysis of high-dimensional data, this is usually not realistic and we typically deal with much less training data, i.e. with $m \ll d$. On the other hand, we can often assume some structural assumptions on w, in the most simple case that it is *sparse*, i.e. that most of its coordinates are zero. Motivated by the success of LASSO [26] in sparse linear regression, it was proposed in [6] that replacing the ℓ_2-norm $\|w\|_2$ in (I.2) by its ℓ_1-norm $\|w\|_1 = \sum_{j=1}^d |w_j|$ leads to sparse classifiers $w \in \mathbb{R}^d$. This method was further popularized in [34] by Zhu, Rosset, Hastie, and Tibshirani, who developed an algorithm that efficiently computes the whole solution path (i.e. the solutions of (I.2)

A. Kolleck is with the Department of Mathematics, Technical University Berlin, Street of 17. June 136, 10623 Berlin, Germany, (e-mail:kolleck@math.tu-berlin.de). A. Kolleck was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

J. Vybíral is with the Department of Mathematical Analysis, Charles University, Sokolovská 83, 186 00, Prague 8, Czech Republic, (e-mail: vybiral@karlin.mff.cuni.cz). J. Vybíral was supported by the ERC CZ grant LL1203 of the Czech Ministry of Education and by the Neuron Fund for Support of Science.

for a wide range of parameters $R > 0$). We refer also to [5], [2], [18] and [19] for other generalizations of the concept of SVM.

Using the ideas of concentration of measure [20] and random constructions in Banach spaces [21], the performance of LASSO was analyzed in the recent area of compressed sensing [11], [7], [3], [10], [12].

ℓ_1-SVM (and its variants) found numerous applications in high-dimensional data analysis, most notably in bioinformatics for gene selection and microarray classification [30], [31], [15]. Finally, ℓ_1-SVM's are closely related to other popular methods of data analysis, like elastic nets [32] or sparse principal components analysis [33].

C. Main results

The main aim of this paper is to analyze the performance of ℓ_1-SVM in the non-asymptotic regime. To be more specific, let us assume that the data points $x_1, \dots, x_m \in \mathbb{R}^d$ can be separated by a hyperplane according to the given labels $y_1, \dots, y_m \in \{-1, +1\}$, and that this hyperplane is normal to a s-sparse vector $a \in \mathbb{R}^d$. Hence, $\langle a, x_i \rangle > 0$ if $y_i = 1$ and $\langle a, x_i \rangle < 0$ if $y_i = -1$. We then obtain $\hat{a}$ as the minimizer of the ℓ_1-SVM. The first main result of this paper (Theorem II.3) then shows that $\hat{a}/\|\hat{a}\|_2$ is a good approximation of a, if the data points are i.i.d. Gaussian vectors and the number of measurements scales linearly in s and logarithmically in d.

Later on, we introduce a modification of ℓ_1-SVM by adding an additional ℓ_2-constraint. It will be shown in Theorem IV.1, that it still approximates the sparse classifiers with the number of measurements m growing linearly in s and logarithmically in d, but the dependence on other parameters improves. In this sense, this modification outperforms the classical ℓ_1-SVM.

D. Organization

The paper is organized as follows. Section II recalls the concept of ℓ_1-Support Vector Machines of [34]. It includes the main result, namely Theorem II.3. It shows that the ℓ_1-SVM allows to approximate sparse classifier a, where the number of measurements only increases logarithmically in the dimension d as it is typical for several reconstruction algorithms from the field of compressed sensing. The two most important ingredients of its proof, Theorems II.1 and II.2, are also discussed in this part. The proof techniques used are based on the recent work of Plan and Vershynin [24], which in turn makes heavy use of classical ideas from the areas of concentration of measure and probability estimates in Banach spaces [20], [21].

Section III gives the proofs of Theorems II.1 and II.2. In Section IV we discuss several extensions of our work, including a modification of ℓ_1-SVM, which combines the ℓ_1 and ℓ_2 penalty.

Finally, in Section V we show numerical tests to demonstrate the convergence results of Section II. In particular, we compare different versions of SVM and 1-Bit Compressed Sensing, which was first introduced by Boufounos and Baraniuk in [4] and then discussed and continued in [23], [24], [22], [1], [17] and others.

E. Notation

We denote by $[\lambda]_+ := \max(\lambda, 0)$ the positive part of a real number $\lambda \in \mathbb{R}$. By $\|w\|_1, \|w\|_2$ and $\|w\|_\infty$ we denote the ℓ_1, ℓ_2 and ℓ_∞ norm of $w \in \mathbb{R}^d$, respectively. We denote by $\mathcal{N}(\mu, \sigma^2)$ the normal (Gaussian) distribution with mean μ and variance σ^2. When ω_1 and ω_2 are random variables, we write $\omega_1 \sim \omega_2$ if they are equidistributed. Multivariate normal distribution is denoted by $\mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^d$ is its mean and $\Sigma \in \mathbb{R}^{d \times d}$ is its covariance matrix. By $\log(x)$ we denote the natural logarithm of $x \in (0, \infty)$ with basis e. Further notation will be fixed in Section II under the name of "Standing assumptions", once we fix the setting of our paper.

II. ℓ_1-NORM SUPPORT VECTOR MACHINES

In this section we give the setting of our study and the main results. Let us assume that the data points $x_1, \dots, x_m \in \mathbb{R}^d$ are equipped with labels $y_i \in \{-1, +1\}$ in such a way that the groups $\{x_i : y_i = 1\}$ and $\{x_i : y_i = -1\}$ can indeed be separated by a sparse classifier a, i.e. that

$$y_i = \operatorname{sign}(\langle x_i, a \rangle), \quad i = 1, \dots, m \tag{II.1}$$

and

$$\|a\|_0 = \#\{j : a_j \neq 0\} \leq s. \tag{II.2}$$

As the classifier is usually not unique, we cannot identify a exactly by any method whatsoever. Hence we are interested in a good approximation of a obtained by ℓ_1-norm SVM from a minimal number of training data. To achieve this goal, we will assume that the training points

$$x_i = r\tilde{x}_i, \quad \tilde{x}_i \sim \mathcal{N}(0, \mathrm{Id}) \tag{II.3}$$

are i.i.d. measurement vectors for some constant $r > 0$.

To allow for more generality, we replace (II.2) by

$$\|a\|_2 = 1, \quad \|a\|_1 \le R. \tag{II.4}$$

Let us observe, that $\|a\|_2 = 1$ and $\|a\|_0 \le s$ implies also $\|a\|_1 \le \sqrt{s}$, i.e. (II.4) with $R = \sqrt{s}$.

Furthermore, we denote by $\hat{a}$ the minimizer of

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^m [1 - y_i\langle x_i, w\rangle]_+ \quad \text{subject to} \quad \|w\|_1 \le R. \tag{II.5}$$

Let us summarize the setting of our work, which we will later on refer to as "Standing assumptions" and which we will keep for the rest of this paper.

Standing assumptions:

(i) $a \in \mathbb{R}^d$ is the true (nearly) sparse classifier with $\|a\|_2 = 1, \quad \|a\|_1 \le R$, $R \ge 1$, which we want to approximate;
(ii) $x_i = r\tilde{x}_i, \quad \tilde{x}_i \sim \mathcal{N}(0, \mathrm{Id}), i = 1, \dots, m$ are i.i.d. training data points for some constant $r > 0$;
(iii) $y_i = \operatorname{sign}(\langle x_i, a\rangle), \quad i = 1, \dots, m$ are the labels of the data points;
(iv) $\hat{a}$ is the minimizer of (II.5);
(v) Furthermore, we denote

$$K = \{w \in \mathbb{R}^d \mid \|w\|_1 \le R\}, \tag{II.6}$$

$$f_a(w) = \frac{1}{m}\sum_{i=1}^m [1 - y_i\langle x_i, w\rangle]_+, \tag{II.7}$$

where the subindex a denotes the dependency of f_a on a (via y_i).

In order to estimate the difference between a and $\hat{a}$ we adapt the ideas of [24]. First we observe

$$\begin{aligned} 0 &\le f_a(a) - f_a(\hat{a}) \\ &= \big(\mathbb{E}f_a(a) - \mathbb{E}f_a(\hat{a})\big) + \big(f_a(a) - \mathbb{E}f_a(a)\big) \\ &\quad - \big(f_a(\hat{a}) - \mathbb{E}f_a(\hat{a})\big) \\ &\le \mathbb{E}(f_a(a) - f_a(\hat{a})) + 2\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)|, \end{aligned}$$

i.e.

$$\mathbb{E}(f_a(\hat{a}) - f_a(a)) \le 2\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)|. \tag{II.8}$$

Hence, it remains

- to bound the right hand side of (II.8) from above and
- to estimate the left hand side in (II.8) by the distance between a and $\hat{a}$ from below.

We obtain the following two theorems, whose proofs are given in Section III.

Theorem II.1. *Let $u > 0$. Under the "Standing assumptions" it holds*

$$\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \le \frac{8\sqrt{8\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u$$

with probability at least

$$1 - 8\left(\exp\left(\frac{-mu^2}{32}\right) + \exp\left(\frac{-mu^2}{32r^2R^2}\right)\right).$$

Theorem II.2. *Let the "Standing assumptions" be fulfilled and let $w \in K$. Put*

$$c = \langle a, w\rangle, \quad c' = \sqrt{\|w\|_2^2 - \langle a, w\rangle^2}$$

and assume that $c' > 0$. If furthermore $c \le 0$, then $\pi\mathbb{E}(f_a(w) - f_a(a))$ can be estimated from below by

$$\frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r}.$$

If $c > 0$, then $\pi\mathbb{E}(f_a(w) - f_a(a))$ can be estimated from below by

$$\frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1/cr}(1 - crt)e^{\frac{-t^2}{2}}\,dt + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r}.$$

Combining Theorems II.1 and II.2 with (II.8) we obtain our main result.

Theorem II.3. *Let $d \geq 2$, $0 < \varepsilon < 0.18$, $r > \sqrt{2\pi}(0.57 - \pi\varepsilon)^{-1}$ and $m \geq C\varepsilon^{-2}r^2R^2\log(d)$ for some constant C. Under the "Standing assumptions" it holds*

$$\frac{\left\| a - \frac{\hat{a}}{\|\hat{a}\|_2} \right\|_2}{\langle a, \frac{\hat{a}}{\|\hat{a}\|_2} \rangle} \leq C' \left(\varepsilon + \frac{1}{r} \right) \tag{II.9}$$

with probability at least

$$1 - \gamma \exp\left(-C'' \log(d)\right) \tag{II.10}$$

for some positive constants γ, C', C''.

Remark II.4. 1) If the classifier $a \in \mathbb{R}^d$ with $\|a\|_2 = 1$ is s-sparse, we always have $\|a\|_1 \leq \sqrt{s}$ and we can choose $R = \sqrt{s}$ in Theorem II.3. The dependence of m, the number of samples needed, is then linear in s and logarithmic in d. Intuitively, this is the best what we can hope for. On the other hand, we leave it open, if the dependence on ε and r is optimal in Theorem II.3.

2) Theorem II.3 uses the constants C, C' and C'' only for simplicity. More explicitly we show that taking

$$m \geq 4\varepsilon^{-2}\left(8\sqrt{8\pi} + 19rR\sqrt{2\log(2d)}\right)^2,$$

we get the estimate

$$\frac{\|a - \hat{a}/\|\hat{a}\|_2\|_2}{\langle a, \hat{a}/\|\hat{a}\|_2 \rangle} \leq 2e^{1/2}\left(\pi\varepsilon + \frac{\sqrt{2\pi}}{r}\right)$$

with probability at least

$$1 - 8\left(\exp\left(\frac{-r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-\log(2d)}{16}\right)\right).$$

3) If we introduce an additional parameter $t > 0$ and choose $m \geq 4\varepsilon^{-2}(8\sqrt{8\pi} + (18+t)rR\sqrt{2\log(2d)})^2$, nothing but the probability changes to

$$1 - 8\left(\exp\left(\frac{-t^2r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-t^2\log(2d)}{16}\right)\right).$$

Hence, by fixing t large, we can increase the value of C'' and speed up the convergence of (II.10) to 1.

Proof of Theorem II.3: To apply Theorem II.1 we choose

$$u = \frac{rR\sqrt{2\log(2d)}}{\sqrt{m}}$$

and

$$m \geq 4\varepsilon^{-2}(8\sqrt{8\pi} + 19rR\sqrt{2\log(2d)})^2$$

and we obtain the estimate

$$\sup_{w \in K} |f_a(w) - \mathbb{E}f_a(w)| \leq \frac{8\sqrt{8\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u \leq \frac{\varepsilon}{2}$$

with probability at least

$$\begin{aligned} &1 - 8\left(\exp\left(\frac{-mu^2}{32}\right) + \exp\left(\frac{-mu^2}{32r^2R^2}\right)\right) \\ &= 1 - 8\left(\exp\left(\frac{-r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-\log(2d)}{16}\right)\right). \end{aligned}$$

Using (II.8) this already implies

$$\mathbb{E}\big(f_a(\hat{a}) - f_a(a)\big) \leq \varepsilon \tag{II.11}$$

with at least the same probability. Now we want to apply Theorem II.2 with $w = \hat{a}$ to estimate the left hand side of this inequality. Therefore we first have to deal with the case $c' = \sqrt{\|\hat{a}\|_2^2 - \langle a, \hat{a} \rangle^2} = 0$, which only holds if $\hat{a} = \lambda a$ for some

$\lambda \in \mathbb{R}$. If $\lambda > 0$, then $\hat{a}/\|\hat{a}\|_2 = a$ and the statement of the Theorem holds trivially. If $\lambda \leq 0$, then the condition $f(\hat{a}) \leq f(a)$ can be rewritten as

$$\sum_{i=1}^{m}[1 + |\lambda| \cdot |\langle x_i, a\rangle|]_+ \leq \sum_{i=1}^{m}[1 - |\langle x_i, a\rangle|]_+.$$

This inequality holds if, and only if, $\langle x_i, a\rangle = 0$ for all $i = 1, \ldots, m$ - and this in turn happens only with probability zero.

We may therefore assume that $c' \neq 0$ holds almost surely and we can apply Theorem II.2. Here we distinguish the three cases $c = \langle \hat{a}, a\rangle \leq 0$, $0 < c \leq 1/r$ and $1/r < c$. First, we will show that the two cases $c \leq 0$ and $0 < c < 1/r$ lead to a contradiction and then, for the case $c > 1/r$, we will prove our claim.

1. case $c \leq 0$*:* Using Theorem II.2 we get the estimate

$$\pi\mathbb{E}(f_a(\hat{a}) - f_a(a)) \geq \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r} \geq \frac{\pi}{2} - \frac{\sqrt{2\pi}}{r}$$

and (II.11) gives (with our choices for r and ε) the contradiction

$$\frac{1}{\pi}\left(\frac{\pi}{2} - \frac{\sqrt{2\pi}}{r}\right) \leq \mathbb{E}(f_a(\hat{a}) - f_a(a)) \leq \varepsilon.$$

2. case $0 < c \leq 1/r$*:* As in the first case we use Theorem II.2 in order to show a contradiction. First we get the estimate

$$\begin{aligned}
&\pi\mathbb{E}(f_a(\hat{a}) - f_a(a)) \\
&\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1/cr}(1 - crt)e^{\frac{-t^2}{2}}\,dt + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r} \\
&\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1/cr}(1 - crt)e^{\frac{-t^2}{2}}\,dt - \frac{\sqrt{2\pi}}{r}.
\end{aligned}$$

Now we consider the function

$$g\colon (0,\infty) \to \mathbb{R}, \quad z \mapsto \int_0^{1/z}(1 - zt)e^{\frac{-t^2}{2}}\,dt.$$

It holds $g(z) \geq 0$ and

$$g'(z) = -\int_0^{1/z} te^{\frac{-t^2}{2}}\,dt < 0,$$

so g is monotonic decreasing. With $cr < 1$ this yields

$$\begin{aligned}
\pi\mathbb{E}(f_a(\hat{a}) - f_a(a)) &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1/cr}(1 - crt)e^{\frac{-t^2}{2}}\,dt - \frac{\sqrt{2\pi}}{r} \\
&= \frac{\sqrt{\pi}}{\sqrt{2}}g(cr) - \frac{\sqrt{2\pi}}{r} \geq \frac{\sqrt{\pi}}{\sqrt{2}}g(1) - \frac{\sqrt{2\pi}}{r} \\
&= \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1}(1 - t)e^{\frac{-t^2}{2}}\,dt - \frac{\sqrt{2\pi}}{r} \\
&\geq 0.57 - \frac{\sqrt{2\pi}}{r}.
\end{aligned}$$

Again, (II.11) now gives the contradiction

$$\frac{1}{\pi}\left(0.57 - \frac{\sqrt{2\pi}}{r}\right) \leq \mathbb{E}(f_a(\hat{a}) - f_a(a)) \leq \varepsilon.$$

We conclude that it must hold $c' > 0$ and $c > 1/r$ almost surely.

3. case $1/r < c$*:* In this case we get the estimate

$$\begin{aligned}
\pi\mathbb{E}(f_a(\hat{a}) - f_a(a)) &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{1/cr}(1 - crt)e^{\frac{-t^2}{2}}\,dt \\
&\qquad + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r} \qquad \text{(II.12)} \\
&\geq \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r} \\
&\geq \frac{c'}{c}e^{-1/2} - \frac{\sqrt{2\pi}}{r},
\end{aligned}$$

where we used $cr > 1$ for the last inequality. Further we get

$$\begin{aligned}\frac{c'}{c} &= \frac{\sqrt{\|\hat{a}\|_2^2 - \langle a, \hat{a}\rangle^2}}{\langle a, \hat{a}\rangle} = \sqrt{\frac{\|\hat{a}\|_2^2 - \langle a, \hat{a}\rangle^2}{\langle a, \hat{a}\rangle^2}} \\ &= \sqrt{\left(\frac{\|\hat{a}\|_2 - \langle a, \hat{a}\rangle}{\langle a, \hat{a}\rangle}\right)\left(\frac{\|\hat{a}\|_2 + \langle a, \hat{a}\rangle}{\langle a, \hat{a}\rangle}\right)} \\ &= \sqrt{\frac{(2 - 2\langle a, \hat{a}/\|\hat{a}\|_2\rangle)(2 + 2\langle a, \hat{a}/\|\hat{a}\|_2\rangle)}{4\langle a, \hat{a}/\|\hat{a}\|_2\rangle^2}} \\ &= \sqrt{\frac{\|a - \hat{a}/\|\hat{a}\|_2\|_2^2 \cdot \|a + \hat{a}/\|\hat{a}\|_2\|_2^2}{4\langle a, \hat{a}/\|\hat{a}\|_2\rangle^2}} \\ &\geq \frac{1}{2}\frac{\|a - \hat{a}/\|\hat{a}\|_2\|_2}{\langle a, \hat{a}/\|\hat{a}\|_2\rangle}.\end{aligned} \tag{II.13}$$

Finally, combining (II.11), (II.12) and (II.13), we arrive at

$$\begin{aligned}&\frac{1}{\pi}\left(\frac{\|a - \hat{a}/\|\hat{a}\|_2\|_2}{\langle a, \hat{a}/\|\hat{a}\|_2\rangle}\frac{1}{2}e^{-1/2} - \frac{\sqrt{2\pi}}{r}\right) \\ &\qquad \leq \mathbb{E}(f_a(\hat{a}) - f_a(a)) \leq \varepsilon,\end{aligned}$$

which finishes the proof of the theorem. ■

III. Proofs

The main aim of this section is to prove Theorems II.1 and II.2. Before we come to that, we shall give a number of helpful Lemmas.

A. Concentration of $f_a(w)$

In this subsection we want to show that $f_a(w)$ does not deviate uniformly far from its expected value $\mathbb{E}f_a(w)$, i.e. we want to show that

$$\sup_{w \in K} |f_a(w) - \mathbb{E}f_a(w)|$$

is small with high probability. Therefore we will first estimate its mean

$$\mu := \mathbb{E}\left(\sup_{w \in K} |f_a(w) - \mathbb{E}f_a(w)|\right) \tag{III.1}$$

and then use a concentration inequality to prove Theorem II.1. The proof relies on standard techniques from [21] and [20] and is inspired by the analysis of 1-bit compressed sensing given in [24].

For $i = 1, \ldots, m$ let $\varepsilon_i \in \{+1, -1\}$ be i.i.d. Bernoulli variables with

$$\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2. \tag{III.2}$$

Let us put

$$\mathcal{A}_i(w) = [1 - y_i\langle x_i, w\rangle]_+, \quad \mathcal{A}(w) = [1 - y\langle x, w\rangle]_+, \tag{III.3}$$

where x is an independent copy of any of the x_i and $y = \operatorname{sign}(\langle x, a\rangle)$. Further, we will make use of the following lemmas.

Lemma III.1. *For $m \in \mathbb{N}$, i.i.d. Bernoulli variables $\varepsilon_1, \ldots, \varepsilon_m$ according to* (III.2) *and any scalars $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$ it holds*

$$\mathbb{P}\left(\sum_{i=1}^m \varepsilon_i[\lambda_i]_+ \geq t\right) \leq 2\mathbb{P}\left(\sum_{i=1}^m \varepsilon_i\lambda_i \geq t\right). \tag{III.4}$$

Proof: First we observe

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^m \varepsilon_i[\lambda_i]_+ \geq t\right) &= \mathbb{P}\left(\sum_{\lambda_i \geq 0} \varepsilon_i\lambda_i \geq t\right) \\ &= \mathbb{P}\left(\sum_{\lambda_i \geq 0} \varepsilon_i\lambda_i \geq t \text{ and } \sum_{\lambda_i < 0} \varepsilon_i\lambda_i \geq 0\right) \\ &\quad + \mathbb{P}\left(\sum_{\lambda_i \geq 0} \varepsilon_i\lambda_i \geq t \text{ and } \sum_{\lambda_i < 0} \varepsilon_i\lambda_i < 0\right).\end{aligned}$$

Now we can estimate the second of these two probabilities by the first one and we arrive at

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=1}^m \varepsilon_i[\lambda_i]_+ \ge t\Big) &\le 2\mathbb{P}\Big(\sum_{\lambda_i\ge 0}\varepsilon_i\lambda_i \ge t \text{ and } \sum_{\lambda_i<0}\varepsilon_i\lambda_i \ge 0\Big)\\ &\le 2\mathbb{P}\Big(\sum_{i=1}^m \varepsilon_i\lambda_i \ge t\Big).\end{aligned}$$

■

Lemma III.2. 1) *For Gaussian random variables* $x_1,\dots,x_m \in \mathbb{R}^d$ *according to* (II.3) *it holds*

$$\mathbb{E}\Big\|\frac{1}{m}\sum_{i=1}^m x_i\Big\|_\infty \le \frac{r\sqrt{2\log(2d)}}{\sqrt{m}}. \tag{III.5}$$

2) *Let the i.i.d. Bernoulli variables* $\varepsilon_1,\dots,\varepsilon_m$ *be according to* (III.2) *and let* $u>0$*. Then it holds*

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_{i=1}^m \varepsilon_i\Big| \ge u\Big) \le 2\exp\Big(\frac{-mu^2}{2}\Big). \tag{III.6}$$

3) *For* $x_1,\dots,x_m \in \mathbb{R}^d$ *and* $K\subset\mathbb{R}^d$ *according to* (II.3) *and* (II.6) *we denote*

$$\tilde{\mu} = \mathbb{E}\Big(\sup_{w\in K}\Big\langle \frac{1}{m}\sum_{i=1}^m x_i, w\Big\rangle\Big). \tag{III.7}$$

Then it holds

$$\mathbb{P}\Big(\sup_{w\in K}\Big\langle \frac{1}{m}\sum_{i=1}^m x_i, w\Big\rangle \ge \tilde{\mu}+u\Big) \le \exp\Big(\frac{-mu^2}{2r^2R^2}\Big). \tag{III.8}$$

Proof:

1) The statement follows from

$$\mathbb{E}\Big\|\frac{1}{m}\sum_{i=1}^m x_i\Big\|_\infty = \frac{r}{\sqrt{m}}\mathbb{E}\|\tilde{x}\|_\infty$$

with $\tilde{x}\sim\mathcal{N}(0,\mathrm{Id})$ and proposition 8.1 of [13]:

$$\frac{\sqrt{\log(d)}}{4} \le \mathbb{E}\|\tilde{x}\|_\infty \le \sqrt{2\log(2d)}. \tag{III.9}$$

2) The estimate follows as a consequence of Hoeffding's inequality [16].

3) Theorem 5.2 of [24] gives the estimate

$$\mathbb{P}\Big(\sup_{w\in K}\Big\langle \frac{1}{m}\sum_{i=1}^m x_i, w\Big\rangle \ge \tilde{\mu}+u\Big) \le \exp\Big(\frac{-u^2}{2\sigma^2}\Big)$$

with

$$\sigma^2 = \sup_{w\in K}\mathbb{E}\Big(\Big\langle \frac{1}{m}\sum_{i=1}^m x_i, w\Big\rangle^2\Big).$$

Since the $x_i's$ are independent we get

$$\frac{1}{m}\sum_{i=1}^m x_i = \frac{r}{\sqrt{m}}\tilde{x} \quad \text{with} \quad \tilde{x}\sim\mathcal{N}(0,\mathrm{Id})$$

and we end up with

$$\sigma^2 = \sup_{w\in K}\mathbb{E}\Big(\frac{r^2}{m}\langle\tilde{x},w\rangle^2\Big) = \frac{r^2}{m}\sup_{w\in K}\|w\|_2^2 = \frac{r^2R^2}{m}. \tag{III.10}$$

■

1) Estimate of the mean μ: To estimate the mean μ, we first derive the following symmetrization inequality, cf. [21, Chapter 6] and [24, Lemma 5.1].

Lemma III.3 (Symmetrization)**.** *Let $\varepsilon_1, \dots, \varepsilon_m$ be i.i.d. Bernoulli variables according to* (III.2). *Under the "Standing assumptions" it holds for μ defined by* (III.1)

$$\mu \leq 2\mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i [1 - y_i \langle x_i, w \rangle]_+ \right|. \tag{III.11}$$

Proof: Let $\mathcal{A}_i(w)$ and $\mathcal{A}(w)$ be according to (III.3). Let x_i' and x' be independent copies of x_i and x. Then $\mathcal{A}_i'(w)$ and $\mathcal{A}'(w)$, generated in the same way (III.3) with x_i' and x' instead of x_i and x, are independent copies of $\mathcal{A}_i(w)$ and $\mathcal{A}(w)$. We denote by $\mathbb{E}'$ the mean value with respect to x_i' and x'. Using $\mathbb{E}'\big(\mathcal{A}_i'(w) - \mathbb{E}'\mathcal{A}'(w)\big) = 0$, we get

$$\begin{aligned}\mu &= \mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\mathcal{A}_i(w) - \mathbb{E}\mathcal{A}(w)\big) \right| \\ &= \mathbb{E} \sup_{w \in K} \Bigg| \frac{1}{m} \sum_{i=1}^{m} \big(\mathcal{A}_i(w) - \mathbb{E}\mathcal{A}(w)\big) \\ &\qquad - \mathbb{E}'\big(\mathcal{A}_i'(w) - \mathbb{E}'\mathcal{A}'(w)\big) \Bigg| \\ &= \mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}'\big(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\big) \right|.\end{aligned}$$

Applying Jensen's inequality we further get

$$\begin{aligned}\mu &\leq \mathbb{E}\,\mathbb{E}' \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \big(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\big) \right| \\ &= \mathbb{E}\,\mathbb{E}' \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \big(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\big) \right| \\ &\leq 2\mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \mathcal{A}_i(w) \right| \\ &= 2\mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i [1 - y_i \langle x_i, w \rangle]_+ \right|\end{aligned}$$

as claimed. ■

Equipped with this tool, we deduce the following estimate for μ.

Lemma III.4. *Under the "Standing assumptions" we have*

$$\mu = \mathbb{E} \sup_{w \in K} |f_a(w) - \mathbb{E}f_a(w)| \leq \frac{4\sqrt{8\pi} + 8rR\sqrt{2\log(2d)}}{\sqrt{m}}.$$

Proof: Using Lemma III.3 we obtain

$$\begin{aligned}\mu &= \mathbb{E} \sup_{w \in K} |f_a(w) - \mathbb{E}f_a(w)| \\ &\leq 2\mathbb{E} \sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i [1 - y_i \langle x_i, w \rangle]_+ \right| \\ &= 2 \int_0^\infty \mathbb{P}\left(\sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i [1 - y_i \langle x_i, w \rangle]_+ \right| \geq t \right) dt.\end{aligned}$$

Now we can apply Lemma III.1 to get

$$\begin{aligned}\mu &\leq 4 \int_0^\infty \mathbb{P}\left(\sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i (1 - y_i \langle x_i, w \rangle) \right| \geq t \right) dt \\ &\leq 4 \int_0^\infty \mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \right| \geq t/2 \right) \\ &\qquad + \mathbb{P}\left(\sup_{w \in K} \left| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i y_i \langle x_i, w \rangle \right| \geq t/2 \right) dt.\end{aligned}$$

Using the second part of Lemma III.2 we can further estimate

$$\begin{aligned}\mu &\le \frac{4\sqrt{8\pi}}{\sqrt{m}} + 4\int_0^\infty \mathbb{P}\biggl(\sup_{w\in K}\biggl|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i\langle x_i, w\rangle\biggr| \ge t/2\biggr)\,dt\\ &= \frac{4\sqrt{8\pi}}{\sqrt{m}} + 8\mathbb{E}\biggl(\sup_{w\in K}\biggl|\biggl\langle \frac{1}{m}\sum_{i=1}^m \varepsilon_i x_i, w\biggr\rangle\biggr|\biggr).\end{aligned}$$

Using the duality $\|\cdot\|_1' = \|\cdot\|_\infty$ and the first part of Lemma III.2 we get

$$= \frac{4\sqrt{8\pi}}{\sqrt{m}} + 8R\,\mathbb{E}\biggl\|\frac{1}{m}\sum_{i=1}^m x_i\biggr\|_\infty \le \frac{4\sqrt{8\pi}}{\sqrt{m}} + \frac{8rR\sqrt{2\log(2d)}}{\sqrt{m}}.$$

■

2) Concentration inequalities: In this subsection we will estimate the probability that $f_a(w)$ deviates anywhere on K far from its mean, i.e. the probability

$$\mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge \mu + t\biggr)$$

for some $t > 0$. First we obtain the following modified version of the second part of Lemma 5.1 of [24], cf. also [21, Chapter 6.1].

Lemma III.5 (Deviation inequality)**.** *Let $\varepsilon_1, \ldots, \varepsilon_m$ be i.i.d. Bernoulli variables according to* (III.2) *and let the "Standing assumptions" be fulfilled. Then, for $\mu \in \mathbb{R}$ according to* (III.1) *and any $t > 0$, it holds*

$$\begin{aligned}&\mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu + t\biggr)\\ &\qquad\le 4\mathbb{P}\biggl(\sup_{w\in K}\biggl|\frac{1}{m}\sum_{i=1}^m \varepsilon_i[1 - y_i\langle x_i, w\rangle]_+\biggr| \ge t/2\biggr).\end{aligned} \tag{III.12}$$

Proof: Using Markov's inequality let us first note

$$\begin{aligned}&\mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu\biggr)\\ &\qquad\le \frac{\mathbb{E}\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)|}{2\mu} = \frac{1}{2}.\end{aligned}$$

Using this inequality we get

$$\begin{aligned}&\frac{1}{2}\mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu + t\biggr)\\ &\le \biggl(1 - \mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu\biggr)\biggr)\\ &\qquad\cdot \mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu + t\biggr)\\ &= \mathbb{P}\biggl(\forall w \in K : |f_a(w) - \mathbb{E}f_a(w)| < 2\mu\biggr)\\ &\qquad\cdot \mathbb{P}\biggl(\exists w \in K : |f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu + t\biggr).\end{aligned}$$

Let $\mathcal{A}_i$ and ε_i be again defined by (III.2), (III.3) and let $\mathcal{A}_i'$ be independent copies of $\mathcal{A}_i$. We further get

$$\begin{aligned}&\frac{1}{2}\mathbb{P}\biggl(\sup_{w\in K}|f_a(w) - \mathbb{E}f_a(w)| \ge 2\mu + t\biggr)\\ &\le \mathbb{P}\biggl(\forall w \in K : \biggl|\frac{1}{m}\sum_{i=1}^m\Bigl(\mathcal{A}_i(w) - \mathbb{E}\mathcal{A}(w)\Bigr)\biggr| < 2\mu\biggr)\\ &\qquad\cdot \mathbb{P}\biggl(\exists w \in K : \biggl|\frac{1}{m}\sum_{i=1}^m\Bigl(\mathcal{A}_i'(w) - \mathbb{E}\mathcal{A}'(w)\Bigr)\biggr| \ge 2\mu + t\biggr)\\ &\le \mathbb{P}\biggl(\exists w \in K : \biggl|\frac{1}{m}\sum_{i=1}^m\Bigl(\bigl(\mathcal{A}_i(w) - \mathbb{E}\mathcal{A}(w)\bigr)\end{aligned}$$

$$
\begin{aligned}
&- \big(\mathcal{A}_i'(w) - \mathbb{E}\mathcal{A}'(w)\big)\bigg)\bigg| \geq t\bigg) \\
&= \mathbb{P}\bigg(\exists w \in K : \bigg|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i(\mathcal{A}_i(w) - \mathcal{A}_i'(w))\bigg| \geq t\bigg) \\
&\leq 2\mathbb{P}\bigg(\exists w \in K : \bigg|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \mathcal{A}_i(w)\bigg| \geq t/2\bigg),
\end{aligned}
$$

which yields the claim. ■

Combining the Lemmas III.1 and III.5 we deduce the following result.

Lemma III.6. *Under the "Standing assumptions" it holds for μ and $\tilde{\mu}$ according to* (III.1) *and* (III.7) *and any $u > 0$*

$$
\begin{aligned}
&\mathbb{P}\bigg(\sup_{w\in K} |f_a(w) - \mathbb{E} f_a(w)| \geq 2\mu + 2\tilde{\mu} + u\bigg) \\
&\qquad \leq 8\bigg(\exp\bigg(\frac{-mu^2}{32}\bigg) + \exp\bigg(\frac{-mu^2}{32r^2R^2}\bigg)\bigg). \qquad \text{(III.13)}
\end{aligned}
$$

Proof: Applying Lemma III.5 and Lemma III.1 we get

$$
\begin{aligned}
&\mathbb{P}\bigg(\sup_{w\in K} |f_a(w) - \mathbb{E} f_a(w)| \geq 2\mu + 2\tilde{\mu} + u\bigg) \\
&\leq 4\mathbb{P}\bigg(\sup_{w\in K}\bigg|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i [1 - y_i\langle x_i, w\rangle]_+\bigg| \geq \tilde{\mu} + u/2\bigg) \\
&\leq 8\mathbb{P}\bigg(\sup_{w\in K}\bigg|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i (1 - y_i\langle x_i, w\rangle)\bigg| \geq \tilde{\mu} + u/2\bigg) \\
&\leq 8\mathbb{P}\bigg(\bigg|\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i\bigg| \geq u/4\bigg) \\
&\qquad + 8\mathbb{P}\bigg(\sup_{w\in K}\bigg|\bigg\langle \frac{1}{m}\sum_{i=1}^{m} x_i, w\bigg\rangle\bigg| \geq \tilde{\mu} + u/4\bigg).
\end{aligned}
$$

Finally, applying the second and third part of Lemma III.2 this can be further estimated from above by

$$
\leq 8\bigg(\exp\bigg(\frac{-mu^2}{32}\bigg) + \exp\bigg(\frac{-mu^2}{32r^2R^2}\bigg)\bigg),
$$

which finishes the proof. ■

Using the two Lemmas III.4 and III.6 we can now prove Theorem II.1.

Proof of Theorem II.1: Lemma III.6 yields

$$
\begin{aligned}
&\mathbb{P}\bigg(\sup_{w\in K} |f_a(w) - \mathbb{E} f_a(w)| \geq 2\mu + 2\tilde{\mu} + u\bigg) \\
&\qquad \leq 8\bigg(\exp\bigg(\frac{-mu^2}{32}\bigg) + \exp\bigg(\frac{-mu^2}{32r^2R^2}\bigg)\bigg).
\end{aligned}
$$

Using Lemma III.4 we further get

$$
\mu \leq \frac{4\sqrt{8\pi} + 8rR\sqrt{2\log(2d)}}{\sqrt{m}}.
$$

Invoking the duality $\|\cdot\|_1' = \|\cdot\|_\infty$ and the first part of Lemma III.2 we can further estimate $\tilde{\mu}$ by

$$
\tilde{\mu} = R\mathbb{E}\bigg\|\frac{1}{m}\sum_{i=1}^{m} x_i\bigg\|_\infty \leq \frac{rR\sqrt{2\log(2d)}}{\sqrt{m}}.
$$

Hence, with probability at least

$$
1 - 8\bigg(\exp\bigg(\frac{-mu^2}{32}\bigg) + \exp\bigg(\frac{-mu^2}{32r^2R^2}\bigg)\bigg)
$$

we have

$$\sup_{w\in K} |f_a(w) - \mathbb{E}f_a(w)| \le 2\mu + 2\tilde{\mu} + u$$
$$\le \frac{8\sqrt{8\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u$$

as claimed. ■

B. Estimate of the expected value

In this subsection we will estimate

$$\mathbb{E}(f_a(w) - f_a(a)) = \mathbb{E}[1 - y\langle x, w\rangle]_+ - \mathbb{E}[1 - y\langle x, a\rangle]_+$$

for some $w \in \mathbb{R}^d \backslash \{0\}$ with $\|w\|_1 \le R$. We will first calculate both expected values separately and later estimate their difference. We will make use of the following statements from probability theory.

Lemma III.7. *Let $a, x \in \mathbb{R}^d$ be according to* (II.4), (II.3) *and let $w \in \mathbb{R}^d \backslash \{0\}$. Then it holds*

1) $\langle x, a\rangle,\ \langle x, \frac{w}{\|w\|_2}\rangle \sim \mathcal{N}(0, r^2)$,
2) $\mathrm{Cov}(\langle x, a\rangle, \langle x, w\rangle) = r^2\langle a, w\rangle$.

Proof: The first statement is well known in probability theory as the 2-stability of normal distribution. For the second statement we get

$$\mathrm{Cov}(\langle x, a\rangle, \langle x, w\rangle) = \mathbb{E}(\langle x, a\rangle\langle x, w\rangle) = \sum_{i,j=1}^{d} a_i w_j \mathbb{E}(x_i x_j)$$
$$= r^2 \sum_{i=1}^{d} a_i w_i = r^2\langle a, w\rangle$$

as claimed. ■

It is very well known, cf. [14, Corollary 5.2], that projections of a Gaussian random vector onto two orthogonal directions are mutually independent.

Lemma III.8. *Let $x \sim \mathcal{N}(0, \mathrm{Id})$ and let $a, b \in \mathbb{R}^d$ with $\langle a, b\rangle = 0$. Then $\langle x, a\rangle$ and $\langle x, b\rangle$ are independent random variables.*

Applying these two lemmas to our case we end up with the following lemma.

Lemma III.9. *For $a \in \mathbb{R}^d$ according to* (II.4), *$x \sim \mathcal{N}(0, r^2\mathrm{Id})$ and $w \in \mathbb{R}^d$ we have*

$$\langle x, w\rangle = c\langle x, a\rangle + c'Z$$

for some $Z \sim \mathcal{N}(0, r^2)$ independent of $\langle x, a\rangle$ and

$$c = \langle a, w\rangle, \quad c' = \sqrt{\|w\|_2^2 - c^2}. \tag{III.14}$$

Remark III.10. Note that c' is well defined, since $c^2 \le \|w\|_2^2\|a\|_2^2 = \|w\|_2^2$.

Proof: If $c' = 0$, the statement holds trivially. If $c' \neq 0$, we set

$$Z = \frac{1}{c'}(\langle x, w\rangle - c\langle x, a\rangle) = \frac{1}{c'}\sum_{i=1}^{d} x_i\left(w_i - ca_i\right).$$

Hence, Z is indeed normally distributed with $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = r^2$. It remains to show that Z and $\langle x, a\rangle$ are independent. We observe that

$$\langle a, w - ca\rangle = \langle a, w\rangle - \langle a, w\rangle\|a\|_2 = 0$$

and, finally, Lemma III.8 yields the claim. ■

Lemma III.11. *Let $a \in \mathbb{R}^d$ and $f_a\colon \mathbb{R}^d \to \mathbb{R}$ be according to* (II.4), (II.7). *Then it holds*

1) $\mathbb{E}f_a(a) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\left[1 - r|t|\right]_+ e^{\frac{-t^2}{2}}\,dt$,
2) $\mathbb{E}f_a(w) = \frac{1}{2\pi}\int_{\mathbb{R}^2}\left[1 - cr|t_1| - c'rt_2\right]_+ e^{\frac{-t_1^2 - t_2^2}{2}}\,dt_1\,dt_2$, *where c and c' are defined by* (III.14).

Proof:

1) Let $\omega \sim \mathcal{N}(0,1)$ and use the first part of Lemma III.7 to obtain

$$\begin{aligned}\mathbb{E}f_a(a) &= \mathbb{E}[1-|\langle x,a\rangle|]_+ = \mathbb{E}[1-r|\omega|]_+\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \big[1-r|t|\big]_+ e^{\frac{-t^2}{2}}\,dt.\end{aligned}$$

2) Using the notation of Lemma III.9 we get

$$\begin{aligned}\mathbb{E}f_a(w) &= \mathbb{E}[1-\operatorname{sign}(\langle x,a\rangle)\langle x,w\rangle]_+\\ &= \mathbb{E}[1-\operatorname{sign}(\langle x,a\rangle)(c\langle x,a\rangle + c'Z)]_+\\ &= \mathbb{E}[1-c|\langle x,a\rangle| - c'\operatorname{sign}(\langle x,a\rangle)Z]_+\\ &= \mathbb{E}[1-c|\langle x,a\rangle| - c'Z]_+\\ &= \frac{1}{2\pi}\int_{\mathbb{R}^2}[1-cr|t_1| - c'rt_2]_+ e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2.\end{aligned}$$

■

Using this result we now can prove Theorem II.2.

Proof of Theorem II.2: Using Lemma III.11 we first observe

$$\begin{aligned}-\pi\mathbb{E}f_a(a) &= -\frac{\sqrt{\pi}}{\sqrt{2}}\int_{\mathbb{R}}[1-r|t|]_+ e^{\frac{-t^2}{2}}\,dt\\ &= -\sqrt{2\pi}\int_0^{\frac{1}{r}}\big(1-rt\big)e^{\frac{-t^2}{2}}\,dt\\ &\geq -\sqrt{2\pi}\int_0^{\frac{1}{r}} e^{\frac{-t^2}{2}}\,dt \geq -\frac{\sqrt{2\pi}}{r}.\end{aligned} \tag{III.15}$$

To estimate the expected value of $f_a(w)$ we now distinguish the two cases $c \leq 0$ and $c > 0$.

1. case: $c \leq 0$: In that case we get

$$\pi\mathbb{E}f_a(w) = \int_{\mathbb{R}}\int_0^{\infty}\big[1-crt_1 - c'rt_2\big]_+ e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2.$$

Since $-crt_1 \geq 0$ for $0 \leq t_1 < \infty$ we can further estimate

$$\begin{aligned}&\pi\mathbb{E}f_a(w) \geq \int_{\mathbb{R}}\int_0^{\infty}\big[1-c'rt_2\big]_+ e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2\\ &\geq \int_{-\infty}^0\int_0^{\infty}(1-c'rt_2)e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2\\ &= \int_{-\infty}^0\int_0^{\infty} e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2 + c'r\int_0^{\infty}\int_0^{\infty} t_2 e^{\frac{-t_1^2-t_2^2}{2}}\,dt_1\,dt_2\\ &= \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}}.\end{aligned}$$

As claimed, putting both terms together, we arrive at

$$\pi\mathbb{E}(f_a(w)-f_a(a)) \geq \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r}.$$

2. case: $c > 0$: First let us observe that $1-crt_1 - c'rt_2 \geq 0$ on $[0,1/cr]\times(-\infty,0] \subset \mathbb{R}^2$. Hence, we get

$$\begin{aligned}\pi\mathbb{E}f_a(w) &= \int_{\mathbb{R}^2}[1-crt_1 - c'rt_2]_+ e^{\frac{-t_1^2-t_2^2}{2}}\,dt_2\,dt_1\\ &\geq \int_0^{\frac{1}{cr}}\int_{-\infty}^0 (1-crt_1 - c'rt_2)e^{\frac{-t_1^2-t_2^2}{2}}\,dt_2\,dt_1\\ &= \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1-crt)e^{\frac{-t^2}{2}}\,dt + c'r\int_0^{\frac{1}{cr}} e^{\frac{-t^2}{2}}\,dt\\ &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1-crt)e^{\frac{-t^2}{2}}\,dt + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right).\end{aligned}$$

Combining this estimate with (III.15) we arrive at

$$\pi\mathbb{E}(f_a(w) - f_a(a)) \geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1-crt)e^{\frac{-t^2}{2}}\,dt \\ + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r}.$$

■

IV. ℓ_1-SVM WITH ADDITIONAL ℓ_2-CONSTRAINT

A detailed inspection of the analysis done so far shows that it would be convenient if the convex body K would not include vectors with large ℓ_2-norm. For example, in (III.10) we needed to calculate $\sup_{w\in K}\|w\|_2^2 = R^2$, although the measure of the set of vectors in K with ℓ_2-norm close to R is extremely small.

Therefore, we will modify the ℓ_1-SVM (II.5) by adding an additional ℓ_2-constraint, that is instead of (II.5) we consider the optimization problem

$$\min_{w\in\mathbb{R}^d}\sum_{i=1}^m[1-y_i\langle x_i, w\rangle]_+ \text{ s. t. } \|w\|_1 \leq R \text{ and } \|w\|_2 \leq 1. \tag{IV.1}$$

The combination of ℓ_1 and ℓ_2 constraints is by no means new - for example, it plays a crucial role in the theory of elastic nets [32]. Furthermore, let us remark that the set

$$\tilde{K} = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R \text{ and } \|w\|_2 \leq 1\} \tag{IV.2}$$

appears also in [24]. We get $\tilde{K} \subset K$ with K according to (II.6). Hence, Theorem II.1 and (II.8) still remain true if we replace K by $\tilde{K}$ and we obtain

$$\sup_{w\in\tilde{K}}|f_a(w) - \mathbb{E}f_a(w)| \leq \frac{8\sqrt{8\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u \tag{IV.3}$$

with high probability and

$$\mathbb{E}(f_a(\hat{a}) - f_a(a)) \leq 2\sup_{w\in\tilde{K}}|f_a(\hat{a}) - f_a(a)|, \tag{IV.4}$$

where $\hat{a}$ is now the minimizer of (IV.1).

It remains to estimate the expected value $\mathbb{E}(f_a(w) - f_a(a))$ in order to obtain an analogue of Theorem II.3 for (IV.1), which reads as follows.

Theorem IV.1. *Let $d \geq 2$, $0 < \varepsilon < 1/2$, $r > 2\sqrt{2\pi}(1-2\varepsilon)^{-1}$, $a \in \mathbb{R}^d$ according to* (II.4)*, $m \geq C\varepsilon^{-2}r^2R^2\log(d)$ for some constant C, $x_1, \ldots, x_m \in \mathbb{R}^d$ according to* (II.3) *and $\hat{a} \in \mathbb{R}^d$ a minimizer of* (IV.1)*. Then it holds*

$$\|a - \hat{a}\|_2^2 \leq \frac{C'\varepsilon}{r(1-\exp\left(\frac{-1}{2r^2}\right))} \tag{IV.5}$$

with probability at least

$$1 - \gamma\exp\left(-C''\log(d)\right)$$

for some positive constants γ, C', C''.

Remark IV.2.

1) As for Theorem II.3 we can write down the expressions explicitly, i.e. without the constants γ, C, C' and C''. That is, taking $m \geq 4\varepsilon^{-2}\left(8\sqrt{8\pi} + (18+t)rR\sqrt{2\log(2d)}\right)^2$ for some $t > 0$, we get

$$\|a - \hat{a}\|_2^2 \leq \frac{\sqrt{\pi/2}\;\varepsilon}{r\left(1-\exp\left(\frac{-1}{2r^2}\right)\right)}.$$

with probability at least

$$1 - 8\left(\exp\left(\frac{-t^2r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-t^2\log(2d)}{16}\right)\right).$$

2) The main advantage of Theorem IV.1 compared to Theorem II.3 is that the parameter r does not need to grow to infinity. Actually, (IV.5) is clearly not optimal for large r. Indeed, if (say) $\varepsilon < 0.2$, we can take $r = 10$, and obtain

$$\|a - \hat{a}\|_2^2 \leq \tilde{C}'\varepsilon$$

for $m \geq \tilde{C}\varepsilon^{-2}R^2\log(d)$ with high probability.

Proof: As in the proof of Theorem II.3 we first obtain $c' = \sqrt{\|\hat{a}\|_2^2 - \langle a, \hat{a}\rangle^2} > 0$ and $c = \langle a, \hat{a}\rangle > 0$. Using Lemma III.11 we get

$$\begin{aligned}
&\pi\mathbb{E}(f_a(w) - f_a(a)) \\
&\geq \int_0^{\frac{1}{r}} \int_{\mathbb{R}} \big((1 - crt_1 - c'rt_2) - (1 - rt_1)\big)e^{\frac{-t_1^2 - t_2^2}{2}}\,dt_2\,dt_1 \\
&= r(1-c)\sqrt{2\pi}\int_0^{\frac{1}{r}} te^{\frac{-t^2}{2}}\,dt
\end{aligned}$$

with

$$1 - c = 1 - \langle a, \hat{a}\rangle \geq \frac{1}{2}(\|a\|_2^2 + \|\hat{a}\|_2^2) - \langle a, \hat{a}\rangle = \frac{1}{2}\|a - \hat{a}\|_2^2.$$

The claim now follows from (IV.4) and (IV.3). ■

V. Numerical experiments

We performed several numerical tests to exhibit different aspects of the algorithms discussed above. In the first two parts of this section we fixed $d = 1000$ and set $\tilde{a} \in \mathbb{R}^d$ with 5 nonzero entries $\tilde{a}_{10} = 1$, $\tilde{a}_{140} = -1$, $\tilde{a}_{234} = 0.5$, $\tilde{a}_{360} = -0.5$, $\tilde{a}_{780} = 0.3$, Afterwards we normalized $\tilde{a}$ and set $a = \tilde{a}/\|\tilde{a}\|_2$ and $R = \|a\|_1$.

A. Dependency on r

We run the ℓ_1-SVM (II.5) with $m = 200$ and $m = 400$ for different values of r between zero and 1.5. The same was done for the ℓ_1-SVM with the additional ℓ_2-constraint (IV.1), which is called $\ell_{1,2}$-SVM in the legend of the figure. The average error of $n = 20$ trials between a and $\hat{a}/\|\hat{a}\|_2$ is plotted against r. We observe that especially for small r's the ℓ_1-SVM with ℓ_2-constraint performs much better than classical ℓ_1-SVM.

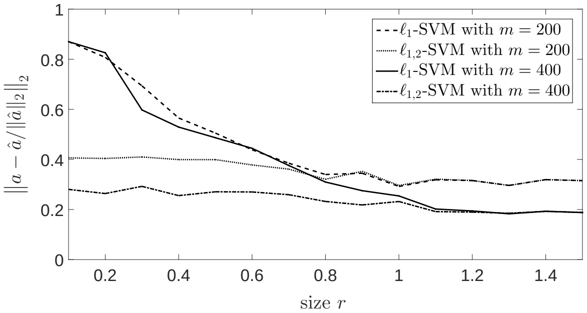


Figure 1. Dependency on r

B. Dependency on m and comparison with 1-Bit CS

In the second experiment, we run ℓ_1-SVM with and without the extra ℓ_2-constraint for two different values of r, namely for $r = 0.75$ and for r depending on m as $r = \sqrt{m}/30$. We plotted the average error of $n = 40$ trials for each value. The last method used is 1-bit Compressed Sensing [24], which is given as the maximizer of

$$\max_{w \in \mathbb{R}^d} \sum_{i=1}^m y_i\langle x_i, w\rangle \quad \text{subject to} \quad \|w\|_2 \leq 1,\ \|w\|_1 \leq R. \tag{V.1}$$

Note that maximizer of (V.1) is independent of r, since it is linear in x_i. First, one observes that the error of ℓ_1-SVM does not converge to zero if the value of $r = 0.75$ is fixed. This is in a good agreement with Theorem II.3 and the error estimate (II.9). This drawback disappears when $r = \sqrt{m}/30$ grows with m, but ℓ_1-SVM still performs quite badly. The two versions of $\ell_{1,2}$-SVM perform essentially better than ℓ_1-SVM, and slightly better than 1-bit Compressed Sensing.

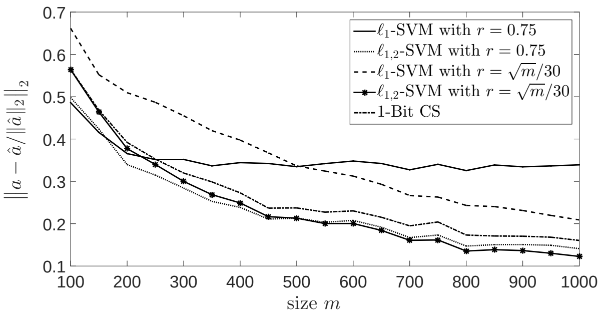


Figure 2. Comparison of ℓ_1-SVM with 1-Bit CS.

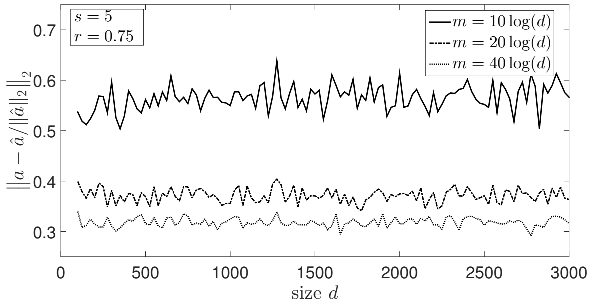


Figure 3. Dependency on d.

C. Dependency on d

In figure 3 we investigated the dependency of the error of ℓ_1-SVM on the dimension d. We fixed the sparsity level $s = 5$ and for each d between 100 and 3000 we draw an s-sparse signal a and measurement vectors x_i at random. Afterwards we run the ℓ_1-SVM with the three different values $m = m_i \log(d)$ with $m_1 = 10$, $m_2 = 20$ and $m_3 = 40$. We plotted the average errors between a and $\hat{a}/\|\hat{a}\|_2$ for $n = 60$ trials.

We indeed see that to achieve the same error, the number of measurements only needs to grow logarithmically in d, explaining once again the success of ℓ_1-SVM for high-dimensional classification problems.

VI. Discussion

In this paper we have analyzed the performance of ℓ_1-SVM (II.5) in recovering sparse classifiers. Theorem II.3 shows, that a good approximation of such a sparse classifier can be achieved with small number of learning points m if the data is well spread. The geometric properties of well distributed learning points are modelled by independent Gaussian vectors with growing variance r and it would be interesting to know, how ℓ_1-SVM performs on points chosen independently from other distributions. The number of learning points needs to grow logarithmically with the underlying dimension d and linearly with the sparsity of the classifier. On the other hand, the optimality of the dependence of m on ε and r remains open. Another important question left open is the behavior of ℓ_1-SVM in the presence of missclasifications, i.e. when there is a (small) probability that the signs $y_i \in \{-1, +1\}$ do not coincide with $\mathrm{sign}(\langle x_i, a\rangle)$. Finally, we proposed a modification of ℓ_1-SVM by incorporating an additional ℓ_2-constraint.

Acknowledgment

We would like to thank A. Hinrichs, M. Omelka, and R. Vershynin for valuable discussions.

References

[1] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin, "One-bit compressed sensing with non-Gaussian measurements", *Linear Algebra Appl.*, vol. 441, pp. 222–239, 2014.

[2] K.P. Bennett and O.L. Mangasarian, "Robust linear programming discrimination of two linearly independent inseparable sets", *Optimization Methods and Software*, pp. 23–34, 1992.

[3] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, "A survey of compressed sensing", Applied and Numerical Harmonic Analysis, Birkhäuser, Boston, 2015.

[4] P.T. Boufounos and R.G. Baraniuk, "1-Bit compressive sensing", In 42nd Annual Conference on Information Sciences and Systems, 2008.

[5] P.S. Bradley and O.L. Mangasarian, "Feature selection via mathematical programming", *INFORMS J. Comput.*, vol. 10, pp. 209–217, 1998.

[6] P.S. Bradley and O.L. Mangasarian, "Feature selection via concave minimization and support vector machines", In Proceedings of the 13th International Conference on Machine Learning, pp. 82–90, 1998.

[7] E. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”, *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, 2006.

[8] C. Cortes and V. Vapnik, “Support-vector networks”, *Machine Learning*, vol. 20, no.3, pp. 273–297, 1995.

[9] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[10] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok, *Introduction to compressed sensing*, in Compressed sensing, Cambridge Univ. Press, Cambridge, pp. 1–64, 2012.

[11] D.L. Donoho, “Compressed sensing”, *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[12] M. Fornasier and H. Rauhut, *Compressive sensing*, In: *Handbook of Mathematical Methods in Imaging*, Springer, pp. 187–228, 2011.

[13] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, Applied and Numerical Harmonic Analysis, Birkhäuser, Boston, 2013.

[14] W. Härdle and L. Simar, *Applied multivariate statistical analysis*, Springer, Berlin, 2003.

[15] M. Hilario and A. Kalousis, “Approaches to dimensionality reduction in proteomic biomarker studies”, *Brief Bioinform*, vol. 9, no. 2, pp. 102–118, 2008.

[16] W. Hoeffding, “Probability inequalities for sums of bounded random variables”, *J. Amer. Stat. Assoc.*, vol. 58, pp. 13–30, 1963.

[17] K. Knudson, R. Saab, and R. Ward, *One-bit compressive sensing with norm estimation*, preprint, available at `http://arxiv.org/abs/1404.6853`.

[18] O.L. Mangasarian, “Arbitrary-norm separating plane”, *Oper. Res. Lett.*, vol. 24, pp. 15–23, 1999.

[19] O.L. Mangasarian, “Support vector machine classification via parameterless robust linear programming”, *Optim. Methods Softw.*, vol. 20, pp. 115–125, 2005.

[20] M. Ledoux, *The Concentration of Measure Phenomenon*, Am. Math. Soc., 2001.

[21] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, Springer, Berlin, 1991.

[22] Y. Plan, R. Vershynin, and E. Yudovina, “High-dimensional estimation with geometric constraints”, available at `http://arxiv.org/abs/1404.3749`.

[23] Y. Plan and R. Vershynin, “One-bit compressed sensing by linear programming”, *Comm. Pure Appl. Math.*, vol. 66, pp. 1275–1297, 2013.

[24] Y. Plan and R. Vershynin, “Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach”, *IEEE Trans. Inform. Theory*, vol. 59, pp. 482–494, 2013.

[25] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, Berlin, 2008.

[26] R. Tibshirani, “Regression shrinkage and selection via the Lasso”, *J. Royal Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[27] V. Vapnik and A. Chervonenkis, “A note on one class of perceptrons”, *Automation and Remote Control*, vol. 25, no. 1, 1964.

[28] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, 1995.

[29] V. Vapnik, *Statistical Learning Theory*, Wiley, Chichester, 1998.

[30] L. Wang, J. Zhu, and H. Zou, “Hybrid huberized support vector machines for microarray classification and gene selection”, *Bioinformatics*, vol. 24, no. 3, pp. 412–419, 2008.

[31] H. H. Zhang, J. Ahn, X. Lin, and Ch. Park, “Gene selection using support vector machines with non-convex penalty”, *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[32] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net”, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.

[33] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis”, *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, 2006.

[34] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines”, In Proc. Advances in Neural Information Processing Systems, vol. 16, pp. 49–56, 2004.

Anton Kolleck received his M.S. in Mathematics at Technical University Berlin, Germany in 2013, where he now continues as Ph.D. student. His research concentrates on sparse recovery and compressed sensing and their applications in approximation theory.

Jan Vybíral received his M.S. in Mathematics at Charles Univeristy, Prague, Czech Republic in 2002. He earned the Dr. rer. nat. degree in Mathematics at Friedrich-Schiller University, Jena, Germany in 2005. He had postdoc positions in Jena, Austrian Academy of Sciences, Austria, and Technical University Berlin, Germany. He is currently an Assistant Professor of Mathematics at Charles University. His core interests are in functional analysis with applications to sparse recovery and compressed sensing.

Big Data of Materials Science: Critical Role of the Descriptor

Luca M. Ghiringhelli,[1,*] Jan Vybiral,[2] Sergey V. Levchenko,[1] Claudia Draxl,[3] and Matthias Scheffler[1]

[1]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin-Dahlem, Germany*
[2]*Department of Mathematical Analysis, Charles University, 18675 Prague, Czech Republic*
[3]*Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, 12489 Berlin, Germany*

(Received 14 April 2014; revised manuscript received 20 October 2014; published 10 March 2015)

Statistical learning of materials properties or functions so far starts with a largely silent, nonchallenged step: the choice of the set of descriptive parameters (termed *descriptor*). However, when the scientific connection between the descriptor and the actuating mechanisms is unclear, the causality of the learned descriptor-property relation is uncertain. Thus, a trustful prediction of new promising materials, identification of anomalies, and scientific advancement are doubtful. We analyze this issue and define requirements for a suitable descriptor. For a classic example, the energy difference of zinc blende or wurtzite and rocksalt semiconductors, we demonstrate how a meaningful descriptor can be found systematically.

DOI: 10.1103/PhysRevLett.114.105503 PACS numbers: 61.50.-f, 02.60.Ed, 71.15.Mb, 89.20.Ff

Using first-principles electronic-structure codes, a large number of known and hypothetical materials have been studied in recent years, and currently, the amount of calculated data increase exponentially with time. The targets of these studies are, for example, the stable structure of solids or the efficiency of potential photovoltaic, thermoelectric, battery, or catalytic materials. Utilizing such data like a reference book (query and read out what was stored) is an avail. Finding the actuating mechanisms of a certain property or function and describing it in terms of a set of physically meaningful parameters (henceforth termed *descriptor*) is the desired science. A most impressive and influential example for the importance and impact of finding a descriptor is the periodic table of elements, where the elements are categorized (described) by two numbers, the table row and column. The initial version had several "white spots," i.e., elements that had not been found at that time, but the chemical properties of these elements were roughly known already from their position in the table. Interestingly, the physical meaning of this two-dimensional descriptor became clear only later. Below we will use an example from materials science to discuss and demonstrate the challenge of finding meaningful descriptors: the prediction of the crystal structure of binary compound semiconductors, which are known to crystallize in zinc blende (ZB), wurtzite (WZ), or rocksalt (RS) structures. The structures and energies of ZB and WZ are very close and for the sake of clarity we will not distinguish them here. The energy difference between ZB and RS is larger, though still very small, namely just 0.001% or less of the total energy of a single atom. Thus, high accuracy is required to predict this difference. This refers to both steps, the explicit calculations and the identification process of the appropriate descriptor (see below). The latter includes the representation of the descriptor-property relation.

For brevity, we only write "property," characterized by a number P_i in the following, with i denoting the actually calculated material, but we mean the materials function(s) as well. In general, the property will be characterized by a string of numbers, but here we like to keep the discussion simple. Analogously, the multidimensional descriptor is denoted as a vector $\boldsymbol{d}_i$, with dimension Ω. The generalization of the discrete data set $\{P_i, \boldsymbol{d}_i\}$ to a continuous function $P(\boldsymbol{d})$ has been traditionally achieved in terms of physical models, or mathematical fits. Scientific understanding of the descriptor $\boldsymbol{d}$ and of the relationship between $\boldsymbol{d}$ and P is needed for deciding with confidence what new materials should be studied next as the most promising novel candidates and for identifying interesting anomalies.

In 1970, Phillips and van Vechten (Ph-vV) [1,2] analyzed the classification challenge of ZB or WZ versus RS structures. They came up with a two-dimensional descriptor, i.e., two numbers that are related to the experimental dielectric constant and nearest-neighbor distance in the crystal [1,2]. Figure 1 shows their conclusion. Clearly, in this representation ZB or WZ and RS structures separate nicely: Materials in the upper left part crystallize in the RS structure, those in the lower right part are ZB or WZ. Thus, based on the ingenious descriptor $\boldsymbol{d} = (E_h, C)$ one can predict the structure of unknown compounds without the need of performing explicit experiments or calculations. Several authors have taken up the Ph-vV challenge and identified alternative descriptors [3–5]. We will come back to this below.

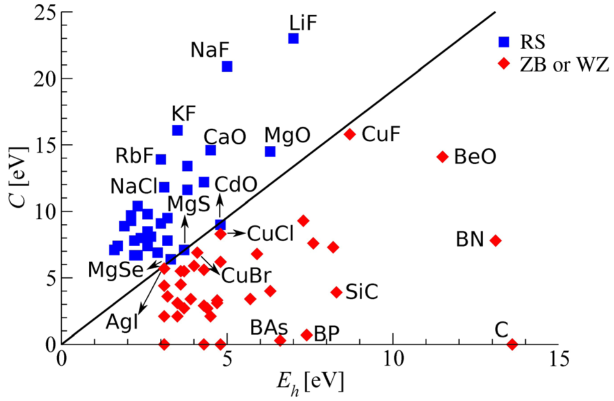


FIG. 1 (color online). Experimental ground-state structures of 68 octet binary compounds, arranged according to the two-dimensional descriptor introduced by Phillips [2] and van Vechten [1]. Because of visibility reasons, only 10 materials are labeled for each structure. See the Supplemental Material for more details [6].

In recent years, the demand for finding the function $P(\boldsymbol{d})$ employed statistical learning theory, which is the focus of this Letter. This strategy has been put forward by several authors in materials science [14–18], as well as in bio- and cheminformatics (see, e.g., Ref. [19] and references therein). Most of these works employed the kernel ridge regression (KRR) approach. For a Gaussian kernel, the fitted property is expressed as a weighted sum of Gaussians: $P(\boldsymbol{d}) = \sum_{i=1}^{N} c_i \exp(-\|\boldsymbol{d}_i - \boldsymbol{d}\|_2^2/2\sigma^2)$, where N is the number of training data points. The coefficients c_i are determined by minimizing $\sum_{i=1}^{N}[P(\boldsymbol{d}_i) - P_i]^2 + \lambda \sum_{i,j=1}^{N,N} c_i c_j \exp(-\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2/2\sigma^2)$, where $\|\boldsymbol{d}_i - \boldsymbol{d}_j\|_2^2 = \sum_{\alpha=1}^{\Omega}(d_{i,\alpha} - d_{j,\alpha})^2$ is the squared ℓ_2 norm of the difference of descriptors of different materials, i.e., their "similarity." The regularization parameter λ and σ are chosen separately, usually with the help of leave-some-out cross validation [20], i.e., by leaving some of the calculated materials out in the training process and testing how the predicted values for them agree with the actually calculated ones.

In essentially all previous materials studies the possibly multidimensional descriptor was introduced *ad hoc*, i.e., without demonstrating that it was the best (in some sense) within a certain broad class (see Ref. [17] for an impressive exception). In this Letter, we describe an approach for finding descriptors for the accurate prediction of a given property of a class of materials, where we restrict ourselves to *ab initio* data.

For the example shown in Fig. 1, statistical learning is unnecessary, because one can determine the classification by visual inspection of the 2D plot. In this Letter, we add the quantitative energy difference between ZB and RS to the original Ph-vV challenge. In general, the descriptor will be higher dimensional. Additionally, the scientific question will be typically more complex than the structural classification. We define the conditions that a proper descriptor must fulfill in order to be suitable for causal "learning" of materials properties, and we demonstrate how the descriptor with the lowest possible dimensionality can be identified. Specifically, we will use the least absolute shrinkage and selection operator (LASSO) for feature selection [21].

All data shown in this study have been obtained with density-functional theory using the local-density approximation (LDA) for the exchange-correlation interaction. Calculations were performed using the all-electron full-potential code FHI-aims [7] with highly accurate basis sets, **k** meshes, and integration grids. For the task discussed in this Letter, the quality of the exchange-correlation functional is irrelevant. Nevertheless, we stress that the LDA provides a good description of the studied materials. In particular, we have computed equilibrium lattice constants and total energies for all three considered lattices (ZB, WZ, RS) of a set of 82 binary materials. The full list of these materials and all calculated properties can be found in the Supplemental Material [6], and all input and output files can be downloaded from the NoMaD repository [22]. Furthermore, we calculated several properties of the isolated neutral atoms and dimer molecules (see below).

Let us start with a simple example that demonstrates the necessity of *validation* in the search for descriptors. The nuclear numbers of a binary semiconductor AB, Z_A and Z_B, unambiguously identify the lowest energy structure: They define the many-body Hamiltonian, and its total energies for the different structures give the stable and metastable structures. Figure 2 (top) displays the total-energy differences of the ZB and RS structures as function of Z_A and Z_B. When using the KRR approach, the data can be fitted well (see Supplemental Material [6]) when the whole set is used for learning. However, the predictive power of KRR based on the descriptor $\boldsymbol{d} = (Z_A, Z_B)$ is bad, as tested by leave-some-out cross validation (see Table I and Supplemental Material [6]). Obviously, the relation between $\boldsymbol{d} = (Z_A, Z_B)$ and the property that we need to learn is by far too complex.

For a descriptor, we consider the following properties to be important, if not necessary.

(a) A descriptor $\boldsymbol{d}_i$ uniquely characterizes the material i as well as property-relevant elementary processes.

(b) (b) Materials that are very different (similar) should be characterized by very different (similar) descriptor values.

(c) The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted.

(d) The dimension Ω of the descriptor should be as low as possible (for a certain accuracy request).

Although the Ph-vV descriptor $\boldsymbol{d} = (E_h, C)$ fulfills conditions (a), (b), and (d), it falls short on condition (c). In contrast, $\boldsymbol{d} = (Z_A, Z_B)$ fails for conditions (b) and (d).

In order to identify a good descriptor, we start with a large number M of candidates (the "feature space") for the

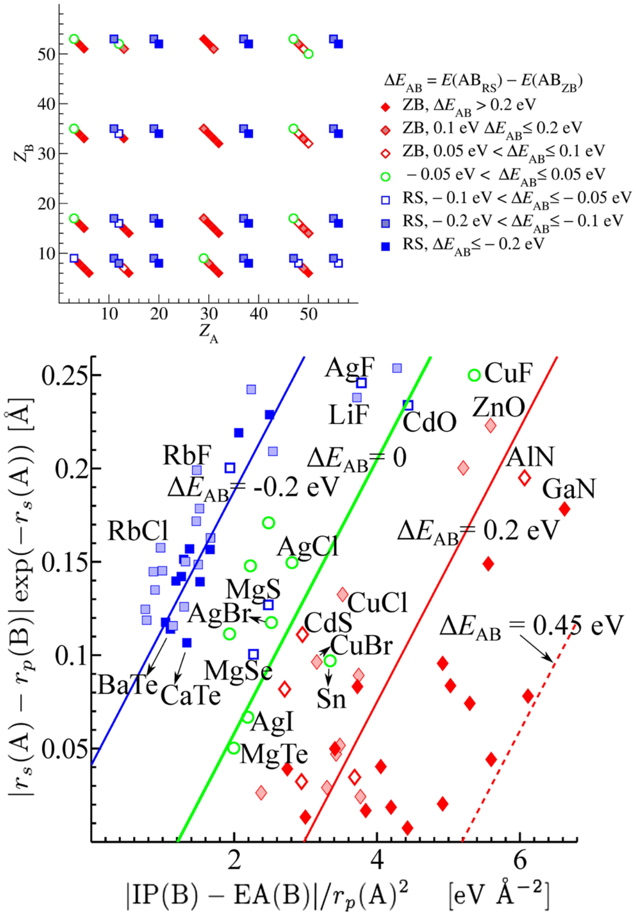


FIG. 2 (color online). Calculated energy differences between RS and ZB structures of the 82 octet binary AB materials, arranged by using the nuclear numbers (Z_A, Z_B) as descriptor (top) and according to our optimal two-dimensional descriptor (bottom). In the bottom panel, seven ZB materials with predicted $\Delta E_{AB} > 0.5$ eV are outside the shown window (see Supplemental Material [6]).

components of $\boldsymbol{d}$. We then look for the Ω-dimensional ($\Omega = 1, 2, \ldots$) descriptor $\boldsymbol{d}$ that gives the best linear fit of $P(\boldsymbol{d})$: $P(\boldsymbol{d}) = \boldsymbol{dc}$, where $\boldsymbol{c}$ is the Ω-dimensional vector of coefficients. It is determined by minimizing the *loss* function $\|\boldsymbol{P} - \boldsymbol{Dc}\|_2^2$, where $\boldsymbol{D}$ is a matrix with each of the N rows being the descriptor $\boldsymbol{d}_i$ for each training data point, and $\boldsymbol{P}$ is the vector of the training values P_i. We emphasize that the choice of a linear fitting function for $P(\boldsymbol{d})$ is not restrictive since, as we will show below, nonlinearities are included in a controlled way in the formation of the candidate components of $\boldsymbol{d}$. The function $P(\boldsymbol{d})$ is then determined by only Ω parameters.

The task is now to find, among all the Ω-tuples of candidate features, the Ω-tuple that yields the smallest $\|\boldsymbol{P} - \boldsymbol{Dc}\|_2^2$. Unfortunately, a computational solution for such a problem is infeasible (NP-hard) [23]. LASSO [21] provides sparse (i.e., low-dimensional) solutions by recasting the NP-hard problem into a convex minimization problem

$$\operatorname*{argmin}_{\boldsymbol{c} \in \mathbb{R}^M} \|\boldsymbol{P} - \boldsymbol{Dc}\|_2^2 + \lambda \|\boldsymbol{c}\|_1, \qquad (1)$$

where the use of the ℓ_1-norm ($\|\boldsymbol{c}\|_1 = \sum_{\alpha=1}^{M} |c_\alpha|$) is crucial. The larger we choose $\lambda > 0$, the smaller the ℓ_1-norm of the solution of Eq. (1) and vice versa. There is actually a smallest $\bar{\lambda} > 0$, such that the solution of Eq. (1) is zero. If $\lambda < \bar{\lambda}$, one or more coordinates of $\boldsymbol{c}$ become nonzero.

We note that the so-called "feature selection" is a widespread set of techniques that are used in statistical analysis in different fields [24], and LASSO is one of them. LASSO was successfully demonstrated in Ref. [17], for identifying the low-dimensional representation of the formation energy of an alloy, within the cluster expansion of the Hamiltonian. Obviously, when a well-identified basis set, such as the cluster expansion, is not available for the property to be modeled, the feature space must be constructed differently. In this Letter, we start from scientific insight, i.e., defining physically motivated primary features that form the basis for a large feature space. We then search for a low-dimensional descriptor that minimizes the RMSE, given by $\sqrt{(1/N)\|\boldsymbol{P} - \boldsymbol{Dc}\|_2^2}$, for our $N = 82$ binary compounds. The property P that we aim to predict is the difference in the LDA energies between RS and ZB for the given atom pair AB, ΔE_{AB}. The order of the two atoms is such that element A has the smallest Mulliken electronegativity: EN $= -(\mathrm{IP} + \mathrm{EA})/2$. IP and EA are atomic ionization potential and electron affinity.

For constructing the feature space, i.e., the candidate components of the descriptor, and then selecting the most

TABLE I. Root-mean-square error (RMSE) and maximum absolute error (MaxAE) in eV for the least-squares fit of all data (first two lines) and for the test set in a leave-10%-out cross validation (L-10%-OCV), averaged over 150 random selections of the training set (last two lines). The errors for (Z_A, Z_B) and (r_σ, r_π) [3] are for a KRR fit at hyperparameters (λ, σ) that minimize the RMSE for the L-10%-OCV (see Supplemental Material [6]). The errors for the $\Omega = 1, 2, 3, 5$ (noted as 1D, 2D, 3D, 5D) descriptors are for the LASSO fit. In the L-10%-OCV for the latter descriptors, the overall LASSO-based selection procedure of the descriptor (see text) is repeated at each random selection of the test set.

Descriptor	Z_A, Z_B	r_σ, r_π	1D	2D	3D	5D
RMSE	2×10^{-4}	0.07	0.14	0.10	0.08	0.06
MaxAE	8×10^{-4}	0.25	0.32	0.32	0.24	0.20
RMSE, CV	0.19	0.09	0.14	0.11	0.08	0.07
MaxAE, CV	0.43	0.17	0.27	0.18	0.16	0.12

relevant of them, we implemented an *iterative* approach. At first we defined primary features. These are (for atom A): $\mathrm{IP}(A)$ and $\mathrm{EA}(A)$, $\mathrm{H}(A)$, and $\mathrm{L}(A)$, i.e., the energies of the highest-occupied and lowest-unoccupied Kohn-Sham levels, as well as $r_s(A)$, $r_p(A)$, and $r_d(A)$, i.e., the radii where the radial probability density of the valence s, p, and d orbitals are maximal. The same was done for atom B. In addition to these atomic data, we offered information on AA, BB, and AB dimers, namely, their equilibrium distance, binding energy, and HOMO-LUMO Kohn-Sham gap. Altogether, these are 23 primary features.

Next, we define rules for linear and nonlinear combinations of the primary features. One can easily generate a huge number of candidate descriptors, e.g., all thinkable but not violating basic physical rules. In the present study, we used about 10 000 candidates grouped in subsets that are used in the different iterations (see Supplemental Material [6]). A more detailed discussion will be given in Ref. [25]. In the language of KRR, this approach designs a kernel, done here by using physical insight. Not surprisingly, LASSO (and actually any other method) has difficulties in selecting among highly correlated features [26]. In these cases, it is not ensured that the first Ω selected features form the best Ω-dimensional descriptor. Although checking correlations between pairs is straightforward and computationally reasonably inexpensive, discovering correlations between triples and more-tuples is computationally prohibitive. Therefore, we adopted a different strategy: The first 25–30 features proposed by LASSO were selected and a batch of least-squares fits performed [when the descriptor is fixed, i.e., the nonzero components of $\boldsymbol{c}$ are fixed, Eq. (1) reduces to a linear, least-squares, fit], taking in turn as $\boldsymbol{D}$ each single feature, each pair, etc. We confirmed that this strategy always finds the best descriptor by running the mentioned extensive search for several different subsets of hundreds of features.

Our procedure identifies as best (i.e., lowest RMSE) 1D, 2D, and 3D descriptors, the first, the first two, and all three of the following features:

$$\frac{\mathrm{IP}(B)-\mathrm{EA}(B)}{r_p(A)^2}, \qquad \frac{|r_s(A)-r_p(B)|}{\exp[r_s(A)]}, \qquad \frac{|r_p(B)-r_s(B)|}{\exp[r_d(A)]}. \tag{2}$$

Note that, mathematically, the descriptor does not necessarily need to build up incrementally in this way; e.g., the 1D one may not be a component of the 2D one. However, in our study, it does. The RMSE and MaxAE for the 1D, 2D, 3D descriptors are reported in Table I. By adding further dimensions to the descriptor, the decrease of the RMSE becomes smaller and smaller.

We tested the robustness of our descriptor by performing a leave-10%-out cross validation (L-10%-OCV). Thereby, the overall procedure of selecting the descriptor is repeated from scratch on a learning set obtained by randomly selecting 90% of the materials. The resulting fitted linear model is applied to the excluded materials and the prediction errors on this set, averaged over 150 random selections, are recorded. The results are shown in Table I. Not only the RMSE, but also the selection of the descriptor, proved very stable. In fact, the 2D descriptor was selected 100% of the times, while the 1D descriptor was the same in 90% of the cases.

The errors for the 2D descriptor introduced by Zunger (Refs. [3,5] and Supplemental Material [6]), based on sums and absolute differences of r_s's and r_p's, are also reported in Table I. The cross-validation error of the linear fit with our 2D descriptor is as small as the highly nonlinear KRR fit with Zunger's 2D descriptor. However, our descriptor bears the advantage that it was derived from a broad class of options by a well-defined procedure, providing a basis for a systematic improvement (with increasing Ω). Our LASSO-derived descriptor contains physically meaningful quantities, like the band gap of B in the numerator of the first component and the size mismatch between valence s and p orbitals (numerators of the second and third component). We note that the components of the descriptors are not symmetric with respect to exchange between A and B. Symmetric features were included in the feature space, but never emerged as prominent, and, for the selected descriptors, symmetrized versions were explicitly constructed, tested, and systematically found to perform worse. This reflects that the symmetry was explicitly broken in the construction of the test set, as the order AB in the compound is such that $\mathrm{EN}(A) < \mathrm{EN}(B)$. Furthermore, we find that d orbitals appear only in the third or higher dimension. In Fig. 2 (bottom) we show the calculated and predicted ΔE_{AB}, according to our best 2D descriptor. It is evident that our 2D descriptor fulfills all above noted conditions, where conditions (a), (c), and (d) are in fact ensured by construction.

In order to further test the robustness and the physical meaningfulness of the identified descriptor, we performed tests by perturbing the value of the property ΔE_{AB} by adding uniform noise in the interval $[-0.1, 0.1]$ eV. The 2D descriptor of Eq. (2) was identified 93% of the times, with an increase of the RMSE by 10% only. More details are reported in Ref. [25]. This test shows that the model allows for some uncertainty in the measured property. Larger noise terms, however, destroy the reliable identification of the descriptor (see Ref. [25]). This analysis implies that the descriptor identified by LASSO contains the important physically meaningful ingredients for the prediction of ΔE_{AB}, even though a physical model that justifies the $P(\boldsymbol{d})$ mapping is not transparent.

Finally, we comment on the causality of the learned descriptor-property relationship. The simplicity of our model is in sharp contrast with what is yielded by, for instance, KRR, where as many fit parameters as observed points are, in principle, necessary. As an indication of having identified a causal (physically meaningful) descriptor for the property ΔE_{AB}, we use the stability of the selection of the descriptor upon both L-10%-OCV and perturbation of the values of the property, under the condition that the $P(\boldsymbol{d})$ dependence has a small number of fit parameters and a simple functional form [see Eq. (2) and Supplemental Material [6]).

Financial support from the Einstein Foundation Berlin is appreciated. J. V. acknowledges financial support by the ERC-CZ Project No. LL1203 of the Ministry of Education, Youth and Sports of the Czech Republic. We thank Krishna Rajan for bringing the relevance of the Ph-vV analysis to our attention and Judea Pearl for inspiring discussions on the concept of causality in the context of statistical inference.

*ghiringhelli@fhi-berlin.mpg.de

[1] J. A. van Vechten, Phys. Rev. **182**, 891 (1969).

[2] J. C. Phillips, Rev. Mod. Phys. **42**, 317 (1970).

[3] A. Zunger, Phys. Rev. B **22**, 5839 (1980).

[4] D. G. Pettifor, Solid State Commun. **51**, 31 (1984).

[5] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).

[6] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.114.105503, which includes Refs. [3,5,7–13], for details on the iterative LASSO procedure, the performance of several descriptors with KRR and linear-least-square regression, details on the cross-validation strategy, details on the LDA calculations, list of calculated atomic and primary features for all considered materials.

[7] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Commun. **180**, 2175 (2009).

[8] D. Donoho, IEEE Trans. Inf. Theory **52**, 1289 (2006).

[9] E. J. Candés, J. Romberg, and T. Tao, IEEE Trans. Inf. Theory **52**, 489 (2006).

[10] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Springer, New York, 2013).

[11] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[12] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).

[13] E. van Lenthe, E. J. Baerends, and J. G. Snijders, J. Chem. Phys. **101**, 9783 (1994).

[14] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).

[15] K. Rajan, Annu. Rev. Mater. Res. **38**, 299 (2008).

[16] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 2810 (2013).

[17] L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B **87**, 035125 (2013).

[18] T. Mueller, E. Johlin, and J. C. Grossman, Phys. Rev. B **89**, 115202 (2014).

[19] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[20] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009).

[21] R. Tibshirani, J. R. Stat. Soc. Ser. B **58**, 267 (1996).

[22] The NoMaD (Novel Materials Discovery) repository contains full input and output files of calculations in materials science, http://nomad-repository.eu.

[23] S. Arora and B. Barak, *Computational Complexity: A Modern Approach* (Cambridge University Press, Cambridge, England, 2009).

[24] I. Guyon and A. Elisseeff, J. Mach. Learn. Res. **3**, 1157 (2003).

[25] L. M. Ghiringhelli, J. Vybiral, S. Levchenko, C. Draxl, and M. Scheffler (to be published).

[26] Two columns of $\boldsymbol{D}$ are correlated if the absolute value of their Pearson's correlation index is (about) 1.