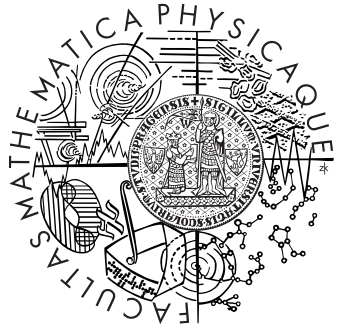
Functions and sequences in analysis and applications

Dr. Jan Vybíral, Ph.D.



Habilitation thesis

Faculty of Mathematics and Physics
of Charles University in Prague

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, 72, 83, 84, 111, 96, 116] 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 [126] 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 [108, 109, 110] (with some forerunners [55, 102]). 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. [117, 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 [53] and [54] (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. Fur-

thermore, 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 [53] and [54] 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. [119].

• 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 $\langle a_i, x\rangle, i = 1, \dots, m$, 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 [114]. The methods used in this area combine powerful techniques of concentration of measure [74], geometry of Banach spaces [75], optimization theory and linear programming [56]. 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 [126] 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

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

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

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

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 $0 < \{s\}^+ \le 1$. Let $s > 0$ and $1 \le p, q < \infty$. Then

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

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

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 [117, 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

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

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 $0 < p, q \le \infty$. 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 $0 < q \le \infty$. 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 [134].

(ii) Let $s \in \mathbb{R}$, $0 < p < \infty$ and $0 < q \le \infty$. 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 $-\infty < s_1 < s_0 < \infty, \quad 0 < p_0 < p_1 < \infty, \quad 0 < q_0 \le q_1 \le \infty$ 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 [52, 63] 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 [117, 96, 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 [118, 121, 122] and to [134].

(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

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

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,

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

(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 $H_p(\mathbb{R}^d), 0 < p \le \infty$ 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 $0 < p \leq 1$, 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 $0 < p \leq 1$ was given in [28] and generalized to $d > 1$ in [73]. It uses the notion of p-atoms on the real line.

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

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 $H^p(\mathbb{R})$, $0 < p \leq 1$ 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

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

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

1.3 Besov and Triebel-Lizorkin spaces

M. Frazier and B. Jawerth extended in [53, 54] 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 [54] 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 [54] 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 [54] 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 [117, 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 [54] through the assumption, that the support of the Fourier transform of φ and ψ stays away from zero. The new condition of [54] 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 [121]. 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 $0 < p, q \le \infty$, $s \in \mathbb{R}$ 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 [53] and [54].

Theorem 1.8. ([121], Theorem 1.19) *(i) Let $0 < p \le \infty$, $0 < q \le \infty$, $s \in \mathbb{R}$. Let $K \in \mathbb{N}_0, L \ge 0$ 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 $c \ge 1$, $a_{\nu m}$ 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 $0 < p < \infty$, $0 < q \le \infty$, $s \in \mathbb{R}$. Let $K \in \mathbb{N}_0, L \ge 0$ 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 $c \ge 1$, $a_{\nu m}$ 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, 85, 132] 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. ([120], 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

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

with

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

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 $\mathfrak{b}^s_{pq}$.

(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

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

with

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

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 [121, 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 [104]. This result (with many forerunners for which we refer to references given in [104]) states, that if Ω has Lipschitz boundary then there is a common bounded linear extension operator $\mathrm{Ext} : A^s_{p,q}(\Omega) \to A^s_{p,q}(\mathbb{R}^d)$ 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 [79], and radar technology [12] to sampling theory [88, 123], 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, 49, 51] 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

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

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

$$x_1^* \geq x_2^* \geq \cdots \geq x_n^* \geq 0.$$

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 $\|x\|_0 \leq k$ and the set of all k-sparse vectors is denoted by

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

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 $m \leq n$. 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.

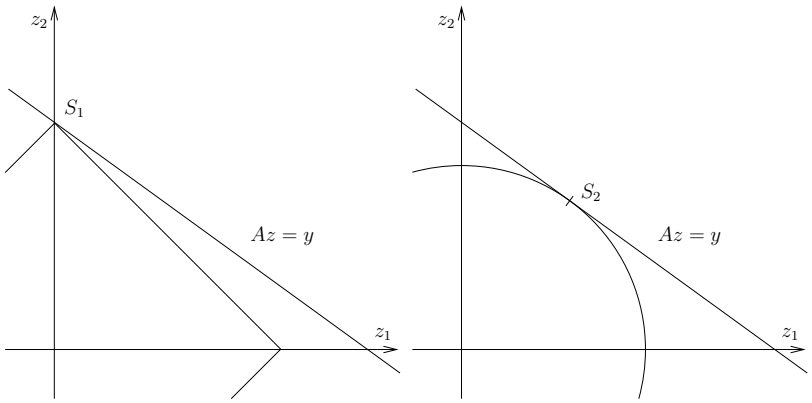


Figure 1: Solution of $S_p = \operatorname{argmin}_{z \in \mathbb{R}^2} \|z\|_p \quad \text{s.t.} \quad y = Az$ for $p = 1$ and $p = 2$

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 [56] 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

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

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 $T \subset \{1, \ldots, n\}$, then we denote by $T^c = \{1, \ldots, n\} \setminus T$ the complement of T in $\{1, \ldots, n\}$. 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 [26].

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

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

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 $T = \text{supp}(v)$, 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 $\|\hat{x}\|_0 \leq \|x\|_0 \leq k$. 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 $\delta \geq 0$, such that

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

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 $\delta_1(A) \leq \delta_2(A) \leq \cdots \leq \delta_k(A)$ 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 2.8. *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.*

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 $k \leq n/2$. 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. [74, 75, 87].

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. [64]. 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 [71].

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 [51].

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 $T_1, \ldots, T_N$ of $\{1, \ldots, n\}$, such that*

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

(ii) $|T_i| = k$ *for all* $i = 1, \ldots, N$ *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 [52, 63] 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 $|\lambda_{\nu,m}| \leq |\lambda'_{\nu,m}|$ for all $\nu \in \mathbb{N}_0$ and $m \in \mathbb{Z}^d$,

then $\|\lambda|b^s_{p,q}\| \le \|\lambda'|b^s_{p,q}\|$. 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 [58], function spaces defined by their subatomic decompositions [106] and to spaces built upon Morrey spaces [59].

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. [100]. 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 [100, 23] for further details.

Let Ω be a bounded Lipschitz domain and let $0 < p_1, p_2, q_1, q_2 \le \infty$ 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

$$\mathcal{I}d : B^{s_1}_{p_1 q_1}(\Omega) \to B^{s_2}_{p_2 q_2}(\Omega) \tag{3.2}$$

is compact. Using Theorem 1.10 and the existence of a universal extension operator due to Rychkov [104], the question may be reduced to the corresponding problem on the sequence space level. We obtain

$$s_n(\mathcal{I}d : B^{s_1}_{p_1 q_1}(\Omega) \to B^{s_2}_{p_2 q_2}(\Omega)) \approx s_n(id : \mathfrak{b}^{s,\Omega}_{pq} \to \mathfrak{b}^{s,\Omega}_{pq}), \tag{3.3}$$

where $\mathfrak{b}^{s,\Omega}_{pq}$ is a certain variant of the spaces $\mathfrak{b}^s_{pq}$ 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 [80] and [78]. We refer also to [76] and [101] for the survey of the state of the art as it was in the second half of 1980's and to [77] 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 [112] 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

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

- 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 [99] and Tikhomirov [115].

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 [70].

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. [134]. 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 [66] 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

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

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 [53] and [54], 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 [66], 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

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

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

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

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

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

and its partially discretized counterpart

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

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 [103], 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 [51] 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 [64] 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. [61] or [62].

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, 81, 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 [71].

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. [113, 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 [105] and after the work of Oskolkov [95], 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, 86, 87, 89, 90] 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, 94, 65].

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. [91], 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. [92, 93], 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, 133] 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

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

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

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

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

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

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*

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

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 $|g'(0)| \geq \varkappa > 0$ 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 [124], 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

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi\in\mathbb{R}^m}} \frac{1}{2}\|w\|_2^2 + \lambda\sum_{i=1}^m \xi_i \quad \text{subject to} \quad y_i\langle x_i, w\rangle \geq 1-\xi_i \quad \text{and} \quad \xi_i \geq 0 \tag{6.1}$$

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

(iii) $y_i = \mathrm{sgn}(\langle x_i, a\rangle), \quad i = 1, \ldots, m$ 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 [74, 75], 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

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

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) [125, 98] 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 [125, 98]. 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 [136, 97, 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 $\sqrt{(1/N)\|\boldsymbol{P} - \boldsymbol{Dc}\|_2^2}$. 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).

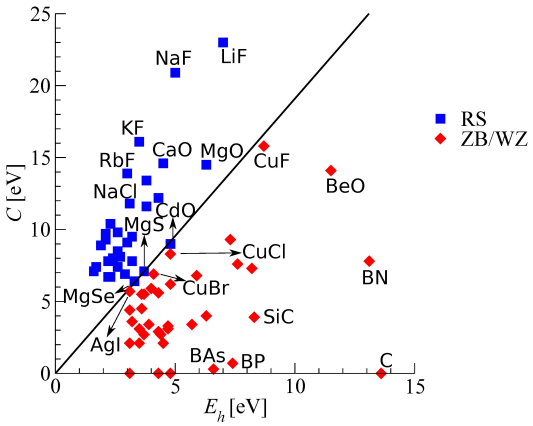


Figure 2: Ground-state structures of 68 octet binary compounds, arranged according to the two-dimensional descriptor introduced by Phillips and van Vechten [125, 98]. Both descriptors and classification derive from experimental data. Because of visibility reasons only 10 materials are labeled for each structure.

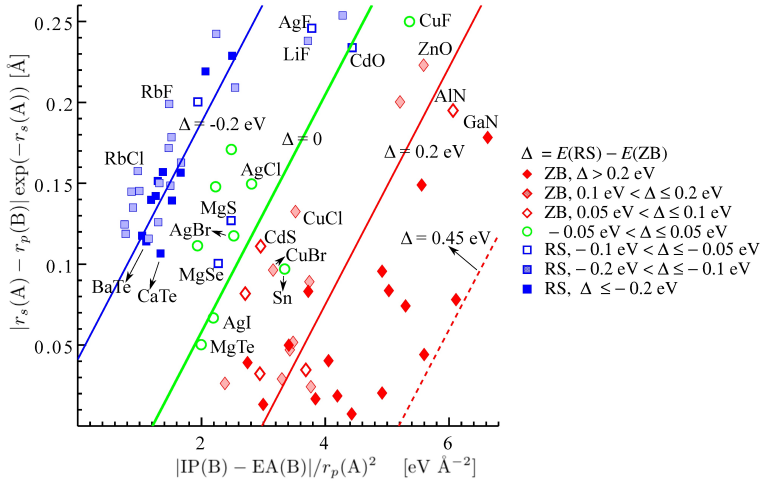


Figure 3: Calculated energy differences of the 82 octet binary materials, arranged according to our optimal two-dimensional descriptor. For visibility reasons, not all materials are labeled. Seven ZB materials with predicted $\Delta E > 0.5$ eV are outside the shown window.

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).

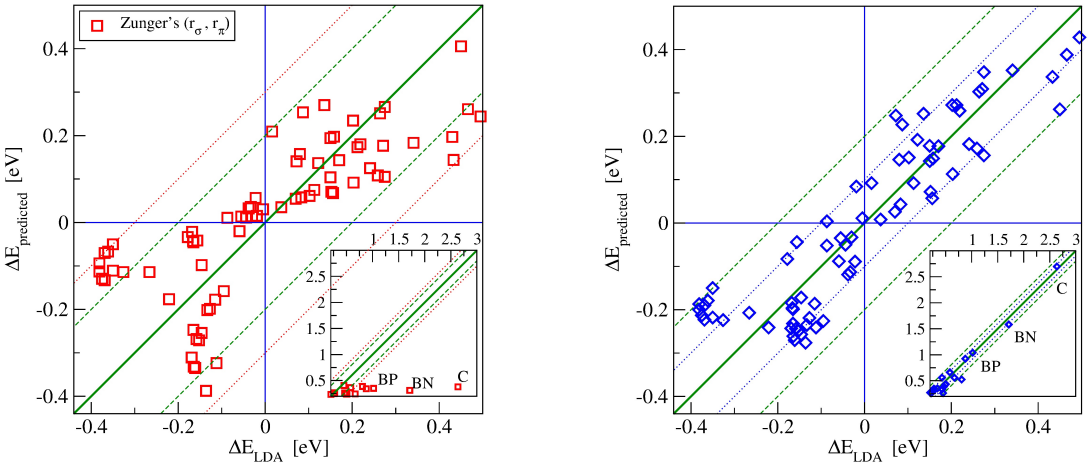


Figure 4: Error of a linear fit for Zunger's descriptors (left figure) and for our best pair (right figure). Each symbol represents one material, which was left out from training and afterwards forecasted by the description found. Especially materials with high ΔE_{LDA} are predicted by our method with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

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

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

References

[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, 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

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

[50] 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

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

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

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

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

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

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

[57] 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)

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

[59] 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.

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

[61] 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.

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

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

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

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

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

[67] 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.

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

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

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

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

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

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

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

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

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

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

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

[79] 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)

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

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

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

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

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

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

[86] 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.

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

[88] 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)

[89] 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.

[90] 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.

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

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

[93] 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.

[94] 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).

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

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

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

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

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

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

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

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

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

[104] 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.

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

[106] 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.

[107] 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

[108] 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).

[109] 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.

[110] 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).

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

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

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

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

[115] 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.

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

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

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

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

[120] 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.

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

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

[123] 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)

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

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

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

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

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

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

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

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

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

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

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

[135] 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.

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