On Some Aspects of Recovery of Sparse Signals in High Dimensions from Nonlinear Measurements using Compressed Sensing

vorgelegt von

Master of Science Anton Kolleck

geb. in Berlin

Von der Fakultät II - Mathematik und Naturwissenschaften

der Technischen Universität Berlin

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

Dr. rer. nat.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Peter Bank
Gutachterin:	Prof. Dr. Gitta Kutyniok
Gutachter:	Prof. Dr. Holger Rauhut
Gutachter:	Dr. habil. Jan Vybíral

Tag der wissenschaftlichen Aussprache: 24.03.2017

Berlin 2017

Deutsche Zusammenfassung

Im Bereich der Datenverarbeitung stehen wir oft vor dem Problem, ein hochdimensionales Signal aus linearen Messungen rekonstruieren zu müssen. Dabei stellen wir in vielen Anwendungen, wie der bildgebenden Diagnostik (engl: medical imaging) in der Medizin, fest, dass die Signale oft dünnbesetzt (engl.: sparse) sind, d.h., dass die meisten Einträge des Signals null oder zumindest sehr klein sind. Vor ungefähr zehn Jahren hat sich Compressed Sensing als neuartige Methode zur Rekonstruktion von dünnbesetzten Signalen aus linearen Messungen hervorgetan.

In dem ersten Abschnitt dieser Dissertation beschäftigen wir uns mit der Theorie der s-Zahlen (engl.: s-numbers). Unser Hauptaugenmerkt liegt hierbei auf der Ungleichung von Carl, welche das asymptotische Verhalten einiger wichtiger s-Zahlen abschätzt. Das Hauptresultat in diesem Abschnitt ist ein Beweis der Ungleichung von Carl für den Fall von Gelfand Zahlen auf quasi-Banach Räumen. Im Kontext von Compressed Sensing können wir dieses Resultat dann insbesondere nutzen, um eine Schranke für die Mindestanzahl an benötigten linearen Messungen herzuleiten, aus welchen wir dünnbesetzte Signale zufriedenstellend rekonstruieren können.

Das Ausgangsproblem von Compressed Sensing beschäftigt sich mit der Rekonstruktion von Signalen aus linearen Messungen. In vielen Anwendungen erhalten wir allerdings nur Zugang zu nichtlinearen Messungen, mit welchen wir uns in den weiteren Abschnitten der Arbeit befassen. Zunächst betrachten wir das Problem des sogenannten 1-Bit Compressed Sensing. Hier erhalten wir lediglich die Vorzeichen der linearen Messungen und dadurch sehr viel weniger Informationen als im klassischen Compressed Sensing. Dennoch ist es möglich, das Signal (bis auf ein skalares Vielfaches) mit der gleichen Ordnung an Messungen zu rekonstruieren. Um das Problem des 1-Bit Compressed Sensing zu lösen, wurden bereits einige Rekonstruktionsalgorithmen vorgeschlagen. Wir werden uns dabei auf die ℓ_1-Support Vector Machines beschränken, welche häufig eine Anwendung bei Problemen des Maschinellen Lernens (engl.: Machine Learning) finden.

Nachdem wir das Problem des 1-Bit Compressed Sensing besprochen haben, widmen wir uns der Approximation von Ridge Funktionen. Diese können wir auf zwei verschiedene Arten interpretieren: Auf der einen Seite, im Kontext von Compressed Sensing, können die Ridge Funktionen als Verallgemeinerung des 1-Bit Compressed Sensing Problems verstanden werden. Anstelle der Vorzeichen der Messwerte erhalten wir hier eine beliebige, unbekannte nichtlineare Störung der Messwerte. Dabei ist allerdings zu bemerken, dass wir bei der Approximation von Ridge Funktionen die zusätzliche Annahme treffen, dass die Nichtlinearität differenzierbar ist und somit das Vorzeichen der Messwerte hier nicht als Nichtlinearität gewählt werden kann.

Auf der anderen Seite bemerken wir, dass die Approximation von Funktio-

nen in vielen Veränderlichen oft unter dem Fluch der Dimensionalität leidet. Bemerkenswert ist, dass selbst die gleichmäßige Approximation von beliebig oft differenzierbaren Funktionen unter diesem Fluch leidet. Da wir aber dennoch an der Approximation von Funktionen in vielen Veränderlichen interessiert sind, müssen wir weitere strukturelle Annahmen treffen. Eine der einfachsten dieser Annahmen ist die der Ridge Funktionen, nämlich, dass die Funktionen konstant entlang von Hyperebenen sind.

Abstract

In data processing we often aim for the recovery of signals in very high dimensions from linear measurements. In many applications such as medical imaging, it turns out that the signal of interest is sparse (or allows a sparse representation in a certain dictionary), i.e., most of its entries are zero or at least very small. Some ten years ago, compressed sensing emerged as a novel method for the recovery of sparse signals in high dimensions, where the sparsity assumption is used to heavily reduce the needed number of linear measurements.

In the first main part of this thesis we deal with the theory of s-numbers. Our main interest is the so-called Carl's inequality, which estimates the asymptotic behavior of important instances of s-numbers such as the Gelfand numbers. The main result in this part is given by Carl's inequality for Gelfand numbers on quasi-Banach spaces. In particular, in the context of compressed sensing Carl's inequality can be used to estimate the minimal needed number of linear measurements in order to obtain reasonable approximations of a sparse signal.

The basic setting of compressed sensing deals with the recovery of sparse signals from linear measurements. However, many applications gain only access to nonlinear measurements. Thus, in the remaining part of this thesis we discuss specific nonlinearities in the measurement process. More precisely, first we will discuss the problem of 1-bit compressed sensing, where we only obtain the signs of the linear measurements. Although getting less information from the measurement process, it is still possible to recover the signal (up to some scalar multiple) from the same amount of linear measurements as in the usual compressed sensing setting. Recently, several algorithms where proposed to solve the 1-bit compressed sensing problem. We will analyze the so-called ℓ_1-support vector machines, which are often used in machine learning applications.

After discussing the 1-bit compressed sensing problem, we will discuss the approximation of ridge functions, which we can interpret in the following two ways. On the one hand, in the context of compressed sensing we can interpret the approximation of ridge function as a generalization of the 1-bit compressed sensing problem. Instead of the sign, here the measurements get disturbed by some unknown nonlinearity. However, in the theory of ridge functions we usually assume a differentiability condition on the nonlinearity, which is obviously not fulfilled for the sign function.

On the other hand, we observe that the approximation of multivariate functions in high dimensions often suffers from the curse of dimensionality. It turns out that even the approximation of arbitrarily often differentiable multivariate functions is intractable in general. Hence, if we still want to approximate multivariate functions

in high dimensions, we have to assume further structure. One of the probably simplest structures we can think of is given by ridge functions, which are constant along hyperplanes.

Acknowledgement

First of all, I would like to thank my mentor Jan Vybíral. You gave me the insight into the theory of compressed sensing and opened very interesting aspects and open problems to me. Furthermore, it was always a pleasure for me to discuss with you and to learn from your brilliant ideas. You also let me learn a lot from your experience and your knowledge. I am very happy that you are my mentor, and my special thanks goes to you.

I would like to thank my supervisor Gitta Kutyniok. It was a pleasure to be in your working group, where I could learn a lot. It was also always very helpful to learn from your experience and to get your advices.

Furthermore, I would like to thank all members of our working group. Namely, I would like to thank Martin Schäfer, Irena Bojarovska, Jackie Ma, Philipp Petersen, Friedrich Philipp, Wang-Q Lim, Axel Flinth, Ali Hashemi, Maximilian März, Anja Hedrich, Anja Peter, Annika Preuß, Mones Raslan, Maximilian Leitheiser, Martin Genzel and Sandra Keiper. It is a pleasure for me to be a part of this group, where I have learned a lot during the seminars and further discussions, but where I also had a lot of fun during our social events and our daily lunch breaks.

I also would like to thank Aicke Hinrichs for the fruitful collaboration and the very nice time in Linz.

I would like to thank Martin Genzel, Sandra Keiper, Martin Schäfer, Jackie Ma and Annika Preuß for proofreading the thesis. You gave me fruitful remarks and suggestions to improve the presentation of this thesis.

Ein ganz besonderer Dank geht an meine Familie, insbesondere an meine beiden Eltern Bernd und Susanne, an Angelika, meine Geschwister Nina und Jakob, meine Nichte Luana und meinen Neffen Emil. Euer Rückhalt gab mir zu jedem Zeitpunkt ein Gefühl der Sicherheit, auf welches ich mich während der gesamten Studienzeit verlassen konnte. Dafür danke ich euch von ganzem Herzen.

Meiner Freundin Sandra möchte ich für die große Unterstützung danken. Gerade in der letzten Zeit der Promotion war es sicher nicht immer einfach, auf dich konnte ich mich aber immer verlassen.

Contents

Chapter 1

Introduction

1.1 Overview

In many applications such as medical imaging we want to deduce information of a certain signal from measured data. Through technological progress, we are facing two kinds of phenomena in this context. On the one hand, we are able to collect more and more data; on the other hand, we obtain efficient tools to compress data without losing quality. Typical examples are more and more improved sensing devices having a higher resolution, such as the MRI machine in medical imaging. But also companies as Google or Facebook increase the amount of stored data. Recent studies expect the total amount of worldwide collected data to double every other year, hence, the growth is exponentially fast. When describing this phenomenon with all of its aspects, the blurry keyword *Big Data* became popular. One of the main challenges in modern data sciences is to handle the huge amount of data we collect nowadays.

In many of those applications the measuring process follows a linear structure. Mathematically formulated, we aim for the reconstruction of the signal $x \in \mathbb{R}^d$ from linear measurements

$$y = Ax \tag{1.1}$$

for a known sensing matrix $A \in \mathbb{R}^{m,d}$ and observed data $y \in \mathbb{R}^m$. Clearly, basic linear algebra teaches us that we cannot recover x from y if we have less measurements than unknowns, i.e., if $m < d$. Nevertheless, to speed up sensing operations or even not to exceed computational power if the ambient dimension d is large, we want to reduce the amount of required measurements m. To still recover the signal x, we certainly have to introduce further assumptions, e.g., by incorporating prior knowledge on the signal.

In many applications the signals of interest follow a specific structure. For instance, the MRI machine wants to produce an image of a certain part of the human body which we know in advance. Using this prior knowledge, Lustig, Donoho, and Pauly were able to reduce the required number of measurements while keeping the resolution unchanged. This in the end leads to a reduction of the scanning time [38].

Introducing the structural assumption on the signal often allows a sparse representation in a certain dictionary which is known in advance. That is, if we represent the (unknown) signal with respect to this (known) dictionary, most of its coefficients

are zero or at least very small, in which case we call the signal *sparse* or *compressible*, respectively. For instance, medical images have a sparse representation with respect to a wavelet basis [31].

Compressed sensing emerged around 2006 [17, 19, 35, 37] as a novel method for the recovery of sparse vectors $x \in \mathbb{R}^d$ from linear measurements as in (1.1). Due to its solid mathematical background the theory has been intensively investigated with applications in many different areas such as image processing, biology, medicine, astronomy, radar communication, and material sciences.

Let us highlight that the main difficulty in compressed sensing is that we only know x to be sparse, i.e., that it only has a few nonzero entries. As we have no prior knowledge on the locations, we cannot delete the zeros and corresponding columns in A to get an overdetermined linear system. Hence, if we want to recover x, we have to reconstruct both, the entries and their locations.

Surprisingly, it turns out that we can exactly recover an s-sparse signal $x \in \mathbb{R}^d$ with $s << d$ nonzero entries from only $m = \mathcal{O}(s \log(d))$ linear measurements. Here, the number of measurements only depends logarithmically on the underlying dimension d leading to a heavy reduction of measurements compared to classical linear algebra methods requiring $m = d$. Clearly, by having less measurements than unknowns, classical linear algebra tools cannot be applied to solve the linear equation (1.1) for x. Since the emergence of compressed sensing, several reconstruction algorithms have been developed. In [25] the authors proposed to use the so-called *basis pursuit*

$$\Delta_1(y) := \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_1 \quad \text{subject to} \quad Aw = y,$$

which has emerged as a suitable method to recover sparse signals and we will, therefore, focus on. Other famous algorithms in the area of compressed sensing are the *orthogonal matching pursuit* (OMP) [88], the *compressive sampling matching pursuit* (CoSaMP) [81], and the *iterative hard thresholding* (IHT) [8], to mention just a few of them. In the following we state some of the ongoing research topics related to the area of compressed sensing which we will work on in this thesis.

1.1.1 Performance over Classes

To solve the linear system (1.1) for x, not only the reconstruction algorithm such as the basis pursuit is important, but also the choice of the measurement matrix A itself. For the usual problem of compressed sensing random matrices with subgaussian entries as normally distributed or Bernoulli variables turn out to perform very well with overwhelmingly high probability [6, 27]. Unfortunately, random matrices are often not useful in practice since the measurement process usually obeys a certain structure. To mention only one example, in MRI the acquired samples are given as Fourier coefficients of the signal [38].

However, in many applications we do not only have to take certain drawbacks from the measurement limitations into account, often also more information on the original signal is known. For instance, in some applications such as wireless communications, we have the additional prior knowledge that the original signal has to lie on a grid [100]. We refer to [41, 66] for recent results concerning the recovery of sparse signals lying on a grid.

Another example is given by group sparsity. Here we partition the indices $1, \dots, d$ into groups, where we assume coefficients belonging to the same group to tend to be zero or nonzero simultaneously. This phenomenon often appears in applications such as microarray analysis, where genes in a family share a similar sequence of DNA building blocks. The authors of [103] derived a bound on the number of samples needed to recover a block sparse signal. We also refer to [60] dealing with the advantages of group sparsity.

Concluding, for each practical application the measurement and reconstruction procedure has to consider given limitations of the particular sensing device, but also further prior knowledge other than sparsity has to be taken into account. Afterwards the performance of the particular measurement and reconstruction method has to be analyzed and compared to optimal benchmarks, which also have to be derived.

1.1.2 Nonlinear Compressed Sensing

The problem of compressed sensing was introduced to recover a signal x from an underdetermined linear system (1.1), where the signal structure, namely the sparsity, is taken into account. However, in many applications it might happen that we do not have access to linear measurements. A very simple example is given by analog to digital conversion, where the values get quantized which destroys the linear measurement process. Modeling the known or even unknown nonlinearity as function f instead of the linear system (1.1), we get the measurements

$$y = f(Ax),$$

which we will denote as *semiparametric single index model*. Driven by various applications, it is natural to ask how to proceed with these nonlinear measurements.

One particular application is given by *1-bit compressed sensing*, where we are left over with quantization in its extreme case, i.e., with the particular choice $f = \operatorname{sign}$. Here we only get the ± 1 measurements

$$y = \operatorname{sign}(Ax). \tag{1.2}$$

The problem of 1-bit compressed sensing was first introduced by Baraniuk and Boufounos in [10] with rapidly growing publications on this subject afterwards, where we, in particular, want to highlight the work of Plan and Vershynin [96]. Surprisingly, it turns out that we only need the same amount of measurements $m = \mathcal{O}(s \log(d))$ compared to the usual compressed sensing theory if we want to recover an s-sparse signal x from (1.2). From this point of view we can neglect the impact of extreme quantization, but let us point out that one clearly needs another recovery algorithm besides the classical basis pursuit.

Another example is given by X-ray crystallography, where one aims to recover the three dimensional structure of a certain crystal from diffraction patterns. In contrary to 1-bit compressed sensing, here we lose the information about the phase, that is, we obtain the measurements

$$y = |Ax|.$$

This problem is denoted as *phase retrieval* [4, 5, 18]. While traditionally no sparsity prior on the signal x is assumed here, some research has been done in this direction, cf. [111].

In the previous two examples the nonlinearity f was known, for 1-bit compressed sensing we chose $f = \operatorname{sign}$ and in the phase retrieval problem f was taken to be the absolute value. Nevertheless, in many applications the function f is even unknown. For the general approach with unknown nonlinearity we refer to the recent work [98].

1.1.3 Approximation of Multivariate Functions

In many areas such as data analysis, learning theory, bioinformatics, parametric PDEs or financial mathematics, functions depending on a large number of variables play a crucial role. In these applications an immense amount of computational power is used to analyze those multivariate functions, thus results on the numerical behavior of multivariate functions become more and more important.

Unfortunately, it turns out that multivariate problems often suffer from the so-called *curse of dimensionality*, that is, the minimal number of operations needed to achieve a satisfying solution of the certain problem grows exponentially fast in the underlying dimension. Clearly, if the ambient dimension is large, exponentially many operations exceed every practicable computational power and the problem gets infeasible.

The curse of dimensionality was observed many times in the literature and appears in a vast amount of situations. One of the probably most impressive results is that even the uniform approximation of infinitely differentiable functions is intractable [83].

To mention another situation where the curse appears, let us stress that many chemical reactions run very fast. Hence, the function which returns the amount of chemical compounds in a given sample can be well approximated by an indicator function: Either the reaction took place and there are (almost) only chemical compounds of the outcome, or the reaction did not start. Unfortunately, it turns out that the approximation of monotone or convex functions also suffers from the curse of dimensionality [58], which transfers to the approximation of indicator functions of convex sets. Hence, we cannot tractably recover indicator functions of convex sets, which gets even worse when dealing with more complicated sets.

To overcome the curse of dimensionality, one might introduce structural conditions on the function of interest. In the area of *information based complexity* it was possible to achieve positive results on the tractability, e.g., by using tensor product constructions, where we refer to the monographs [84–86]. Another approach is to require the function f to allow a sparse representation in a known dictionary. For instance, shearlets provide optimally sparse approximations of so-called cartoon-like functions [49, 50, 70].

One might also reduce the complexity by considering functions which are defined on a d-dimensional space but only depend on some of the variables. That is, for some unknown indices $i_1, \dots, i_k \in \{1, \dots, d\}$ with $k \ll d$ we want to approximate the function

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \tilde{f}(x_{i_1}, \dots, x_{i_k}),$$

for some $\tilde{f}\colon \mathbb{R}^k \to \mathbb{R}$. For the study of functions of few variables in high dimensions, we refer to [33, 101, 114].

Another structural assumption to overcome the curse of dimensionality is given by considering functions which are constant along some unknown manifolds. Here one clearly has to make further restrictions on the manifolds, since every function is constant along its level sets. The probably simplest instances of nontrivial manifolds are given by hyperplanes and spheres, leading to so-called *ridge functions* and *translated radial functions*, which we will introduce later on in more detail.

1.2 Contributions

In the following subsections we collect and explain the main contributions of the present thesis.

1.2.1 Carl's Inequality for Quasi-Banach Spaces

The theory of the so-called *s-numbers* emerged from studies of geometry of Banach spaces and operators between them but found applications in many other areas. Our main interest is the so-called *Carl's inequality* [21], which relates the asymptotic behavior of *approximation-*, *Gelfand-*, and *Kolmogorov numbers* to the *entropy numbers* and is one of the most important tools in the theory of s-numbers.

More explicit, if $T\colon X \to Y$ is a bounded linear operator between two Banach spaces X and Y, Carl's inequality states that for any $\alpha > 0$ there exists a constant γ_α only depending on α such that for any $n \in \mathbb{N}$ it holds

$$\sup_{1\le k\le n} k^\alpha e_k(T) \le \gamma_\alpha \sup_{1\le k\le n} k^\alpha s_k(T),$$

where e_k denotes the k-th entropy number of T and s_k stands either for the k-th approximation, Kolmogorov, or Gelfand number, cf. Chapter 4 for exact definitions. Although the original proof of Carl only holds for the case of Banach spaces, it was already observed in [7, 47] or [39, Section 1.3.3] that Carl's inequality easily extends to the case of quasi-Banach spaces and approximation or Kolmogorov numbers.

Recently, the s-numbers and in particular Carl's inequality were used in the area of compressed sensing to provide general lower bounds for the performance of sparse recovery methods, cf. [17, 35] and also [9, 45]. In its basic setting, compressed sensing studies pairs of (linear) measurement maps $A\colon \mathbb{R}^d \to \mathbb{R}^n$ and (nonlinear) recovery maps $\Delta\colon \mathbb{R}^n \to \mathbb{R}^d$ such that the error $x - \Delta(Ax)$ is small for all vectors belonging to a certain set $K \subset \mathbb{R}^d$, e.g., the set of all s-sparse vectors. The search for the optimal recovery pair (A, Δ) is expressed as compressive n-width, which is defined as

$$E_n(K, Y) = \inf_{(A,\Delta)} \sup_{x\in K} \|x - \Delta(Ax)\|_Y,$$

where Y denotes a (quasi-)norm on $\mathbb{R}^d$. Based on previous work in approximation theory and information based complexity [79, 82, 95] it was observed in [27, 35, 65] that the compressive n-width are equivalent to the so-called Gelfand n-width of a symmetric and subadditive set $K \subset \mathbb{R}^d$, which itself are related to Gelfand numbers.

In the field of compressed sensing the unit balls B_p^d of the spaces ℓ_P^d usually serve as a good model for compressible signals and the error of reconstruction is often

measured in the euclidean ℓ_2-norm. Consequently, Donoho investigated the decay of the compressive n-width $E_n(B_p^d, \ell_2^d)$ for $0 < p < 1$ by using Carl's inequality for Gelfand numbers [35]. Unfortunately, the argument presented by Donoho contains a crucial flaw. In [21] Carl's inequality was only proven for the case of Banach spaces, hence, it cannot be applied in the case when $p < 1$. This gap was corrected in [44] with a completely different approach using techniques from compressed sensing. The question whether Carl's inequality also holds in the case of quasi-Banach spaces remained open, and indeed the authors of [44] expressed their belief that "Carl's theorem actually fails for Gelfand widths of general quasi-norm balls".

The main result of Chapter 4 is that Carl's inequality also holds for the case of quasi-Banach spaces and Gelfand numbers. As an application, we also fill the gap in Donohos argument and give an alternative proof for the bound of $E_n(B_p^d, \ell_2^d)$ contained in [44].

1.2.2 Non-Asymptotic Analysis of ℓ_1-Support Vector Machines

In 1-bit compressed sensing we aim to recover a sparse signal $x \in \mathbb{R}^d$ from nonlinear measurements of the form

$$y_i = \operatorname{sign}(\langle a_i, x \rangle), \quad i = 1, \dots, m.$$

Note that, instead of x, we could also aim to recover the hyperplane $E_x = \{w \in \mathbb{R}^d \mid \langle x, w \rangle = 0\} \subset \mathbb{R}^d$ through the origin with normal vector x, which separates the two sets $C_+ = \{x_i \mid y_i = +1\}$ and $C_- = \{x_i \mid y_i = -1\}$. Indeed, up to sign and scale, x and E_x are uniquely determined by each other.

In machine learning, *support vector machines (SVMs)* are standard classification methods which are constructed to find a plane $E \subset \mathbb{R}^d$ separating two classes $C_+, C_- \subset \mathbb{R}^d$. While there are usually several planes (or none) separating the two classes, SVMs search for the hyperplane which not only separates the two classes, but also maximizes the distance to them.

Since their introduction by Vapnik and Chervonenkis [108] SVMs where studied intensively and many different variants where developed. We will concentrate on the so-called *soft margin SVMs* [28], which allow for misclassification by introducing so-called *slack variables*. Due to their robustness against noise, they are the most frequently used SVMs nowadays. In its most common form, the soft margin SVM is given by the optimization problem

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

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

for a trade-off parameter $\lambda > 0$ and slack variables ξ_i.

The term $\|w\|_2^2$ reflects the search for a hyperplane maximizing the distance to the two groups. Hence, a reasonable requirement is the existence of such a hyperplane. So far, no further assumption on the true classifier is made. A certain drawback of SVMs, as the variant given above, is that they perform rather badly when the number of measurements is much smaller than the ambient dimension of sample points a_i, i.e., if $m << d$ [46]. To overcome this drawback, the authors of [46]

proposed the *bet on sparsity* principle, suggesting that one should "use a procedure that does well in sparse problems, since no procedure does well in dense problems".

In this spirit, to aim for sparse classifiers, the authors of [12] proposed to replace the Euclidean norm $\|w\|_2$ in the definition of the SVM by the ℓ_1-norm $\|w\|_1 = \sum_{j=1}^{d} |w_j|$, which was also motivated by the success of Lasso [107]. For the success of Lasso in the framework of 1-bit compressed sensing we also refer to the recent work [97].

Support vector machines with ℓ_1-penalty, which we will refer to as ℓ_1-*SVM*, were further popularized in [116] and various variants of it became a standard tool in analysis of high-dimensional classification problems with sparsity constraints. The ℓ_1-SVMs found numerous applications, e.g., in bioinformatics [56, 112, 115], and are closely related to other popular methods like elastic nets [117] or sparse principal component analysis (sPCA) [118].

The performance of SVMs was studied intensively in the literature, where we highlight the remarkable results of Steinwart stating that various SVMs are consistent, i.e., they recover the true classifier x if the number m of measurements tends to infinity [104, 105].

In Chapter 5 we analyze the ℓ_1-SVM in the framework of 1-bit compressed sensing. In particular, we show that the ℓ_1-SVM recovers an s-sparse signal x from only $\mathcal{O}(s \log(d))$ measurements, which is the same rate as for the usual compressed sensing and goes in hand with other recent results on this topic [96]. Further, we will consider a modification of the ℓ_1-SVM by adding an additional ℓ_2-constraint which is recalled as *doubly regularized support vector machine* [113] and which we will denote as $\ell_{1,2}$-SVM. We will show that it still recovers an s-sparse signal from only $m = \mathcal{O}(s \log(d))$ measurements, but the performance on other parameters improves.

1.2.3 Approximation of Ridge Functions

To overcome the curse of dimensionality in the reconstruction of multivariate functions, we restrict to the approximation of functions following a certain structure. We focus on the approximation of *ridge functions*, which are constant along an unknown hyperplane and can be written as

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto f(x) = g(\langle a, x \rangle)$$

for an unknown function $g\colon \mathbb{R} \to \mathbb{R}$, called the *ridge profile*, and some unknown $a \in \mathbb{R}^d$, called the *ridge vector*. The approximation of ridge functions from point queries was initiated by Cohen, Daubechies, DeVore, Kerkyacharian and Picard [26] and was further developed in a series of recent papers [42, 55, 80].

The study of ridge functions is by no means new in mathematics. For example, they very often appear in statistics such as econometrics in the frame of so-called *single-index models* [59] or in physics as *plane waves* [13]. A difference between our approach and the usual setting in statistical learning is that we suppose that we can freely choose the sampling points of f, i.e., they are not given in advance. Also, note that ridge functions appeared in mathematical analysis of neural networks [15, 93] and they form the building blocks of so-called *ridgelets*, which were introduced by Candés and Donoho [16].

Furthermore, in approximation theory the simple structure of ridge functions motivated the question whether general functions can be well approximated by sums

of ridge functions with pioneering work [76], where the term "ridge function" was first introduced, and also [72]. A survey on approximation by sums of ridge functions is given in [94].

The authors of [26] studied ridge functions defined on the cube $[0,1]^d$ with stochastic ridge vector, i.e., $a = (a_1, \dots, a_d)$ satisfies $a_j \geq 0$ and $\sum_{j=1}^d a_j = 1$. Under the further assumption that the ridge profile $g\colon [0,1] \to \mathbb{R}$ is a $\mathcal{C}^s$-function for some $s > 1$, they proposed an optimal recovery method. First, they sample f along the base points $t_i(1, \dots, 1)$ for t_i ranging from 0 to 1. This gives function values of g at t_i which can be used to find a suitable approximation $\hat{g}$ of g. Afterwards they sample f at the padding points $\eta(1, \dots, 1) + \mu b_i$ for some constants μ, η and random Bernoulli vectors b_i and used a compressing sensing approach for the recovery of a.

The approximation scheme presented in [26], which we have described above, heavily relies on the quite restrictive assumption $a_j \geq 0$. By using a completely different approach, in particular changing the approximation order, the authors of [42] were able to drop this assumption. Based on the simple observation

$$\nabla f(x) = g'(\langle a, x \rangle)a,$$

they first approximated the gradient of f at sampling points ξ_j from finite differences $(f(\xi_j + \varepsilon\varphi_i) - f(\xi_j))/\varepsilon$ for some random vectors φ_i. Afterwards, by employing techniques from compressed sensing, this gives an approximation $\hat{a}$ of a and the problem reduces to find an approximation of the univariate function g, which can be done by basic numerical algorithms as spline interpolation.

This approximation scheme requires the function f to be defined on the unit ball in $\mathbb{R}^d$ and g to be twice continuously differentiable. Further, it can also be used to find an approximation for the more general model

$$f(x) = g(Ax)$$

for some unknown arbitrary matrix $A \in \mathbb{R}^{k,d}$ with $\operatorname{rank}(A) = k \ll d$.

In this thesis we will close some gaps and answer some open questions left so far in the analysis of [42].

The ridge functions considered in [42] are defined on the unit ball of $\mathbb{R}^d$, whose geometry perfectly fits together with the structure of ridge functions using the scalar product. Although the possibility of extending the analysis to ridge functions defined on other domains was already mentioned, no further steps in this direction were done. We study ridge functions defined on the unit cube $[-1,1]^d$ in detail. To adapt the approximation scheme to this case, the crucial component will be the use of the sign of a vector, which is defined componentwise. Although the mapping $x \mapsto \operatorname{sign}(x)$ is obviously not continuous and the two vectors $\operatorname{sign}(x_1)$ and $\operatorname{sign}(x_2)$ may be far from each other, although x_1 and x_2 are close, we observe that for fixed $a \in \mathbb{R}^d$ the map $x \mapsto \langle a, \operatorname{sign}(x) \rangle$ is continuous in a. This observation allows us to imitate the approximation scheme of [42] for this setting and also partly shows how one can adapt it to further domains.

Another open question which was only briefly discussed in [42] is the approximation of ridge functions from noisy measurements, which is an important step for any practical application. Since we approximate the gradient of f from finite differences leading to approximation errors which we will denote as deterministic noise, we have

two different kinds of noise in this setting. To handle random noise we present an algorithm using the so-called *Dantzig selector* from [20], which was already proposed in [42].

The third main topic we will discuss is adaptability of the methods to similar function classes. More concrete, we consider translated radial functions

$$f(x) = g(\|a - x\|_2^2)$$

for an unknown univariate function g and a center point $a \in \mathbb{R}^d$. Translated radial functions are constant along spheres centered in a, in contrary to ridge functions which are constant along hyperplanes perpendicular to the ridge vector. Following the study of ridge functions we develop an approximation scheme which recovers the center point a and, in particular, even works when the function f has a singularity there.

Chapter 2

Preliminaries and Notation

In this chapter we collect basic notation which we will use in the remainder of this thesis.

In this thesis we restrict to real-valued signals. The dimensions of the underlying spaces are denoted by m and d, where we usually assume $m < d$. To avoid confusion, the entries of $x \in \mathbb{R}^d$ are denoted with subindex j, i.e., by x_j, $j = 1, \dots, d$ and the entries of $y \in \mathbb{R}^m$ are denoted with subindex i by y_i, $i = 1, \dots, m$. The entries of a matrix $A \in \mathbb{R}^{m,d}$ with m rows and d columns will be denoted by $a_{ij} \in \mathbb{R}$, where $1 \leq i \leq m$ and $1 \leq j \leq d$. Further, the rows of A will be denoted by $a_i \in \mathbb{R}^d$, $i = 1, \dots, m$.

The ℓ_p-norm of a given vector $x \in \mathbb{R}^d$ and some $0 < p < \infty$ is defined by

$$\|x\|_p := \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p}, \tag{2.1}$$

which gets complemented by putting for $p = \infty$ and $p = 0$

$$\|x\|_\infty := \max_{j \in [d]} |x_j| \quad \text{and} \quad \|x\|_0 := \#\{j \in [d] \mid x_j \neq 0\}. \tag{2.2}$$

Here $\#X$ denotes the cardinality of the set X and $[d] = \{1, \dots, d\}$ denotes the set of natural numbers from 1 to d. Note

$$\lim_{p \to 0} \|x\|_p^p = \|x\|_0 \quad \text{and} \quad \lim_{p \to \infty} \|x\|_p = \|x\|_\infty,$$

which makes the definition of the ℓ_0- and ℓ_∞-norm quite reasonable. It is well known that the ℓ_p-norm defines a norm for $1 \leq p \leq \infty$ and a quasi-norm for $0 < p < 1$. Even more, defining the scalar product

$$\langle x, y \rangle = y^T \cdot x = \sum_{j=1}^{d} x_j y_j \tag{2.3}$$

for $x, y \in \mathbb{R}^d$ makes $(\mathbb{R}^d, \|\cdot\|_2)$ a Hilbert space. In contrary, the ℓ_0-norm does not define a norm, since it is, for instance, not homogeneous, i.e., $\|\lambda x\|_0 \neq |\lambda| \cdot \|x\|_0$ for general $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^d$. Further, for $0 < p < \infty$ and $x \in \mathbb{R}^d$ we define the $\ell_{p,\infty}$-quasi-norm by

$$\|x\|_{p,\infty} := \max_{j \in [d]} j^{1/p} x_{(j)}, \tag{2.4}$$

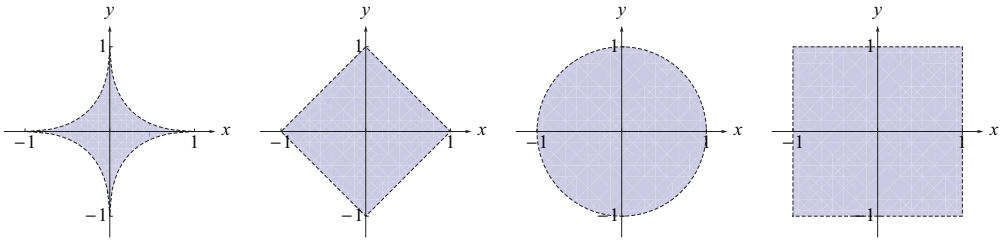

Figure 2.1: Sketch of the unit balls B_p^2 for $p \in \{1/2, 1, 2, \infty\}$.

where $(x_{(j)})$ denotes a non-increasing rearrangement of the absolute entries of x.

In the following we will denote $\mathbb{R}^d$ endowed with the (quasi-)norm $\|\cdot\|_p$ simply by ℓ_p^d and the (open) unit balls in ℓ_p^d by

$$B_p^d := \{x \in \mathbb{R}^d \mid \|x\|_p < 1\}. \tag{2.5}$$

Figure 2.1 illustrates the open unit balls B_p^d for different choices of p.

Applying a real-valued function to some vector $x \in \mathbb{R}^d$ has to be understood componentwise. For instance, the entries of the positive- and negative parts $x_+, x_- \in \mathbb{R}^d$ of x are given by

$$(x_+)_j = [x_j]_+, \qquad (x_-)_j = [-x_j]_+, \tag{2.6}$$

where we set $[t]_+ := \max\{t, 0\}$ for some $t \in \mathbb{R}$.

Chapter 3

Background of Compressed Sensing

In this chapter we recall basic concepts and tools from the area of compressed sensing, which we will need in the remainder of this thesis. As it is not our aim to develop the theory of compressed sensing but rather to use it, we will only give a very focused presentation and refer the interested reader to [9, 40] for recent overviews and to the detailed book [45] and references therein.

In the first section we introduce basic definitions and properties concerning the recovery of sparse vectors from linear measurements such as the nullspace property (NSP) and the restricted isometry property (RIP). We will further introduce the so-called basis pursuit as a particular recovery method.

The second section deals with tools from probability theory, where we will see that random matrices yield optimal recovery rates and, in particular, satisfy the NSP and the RIP with overwhelmingly high probability.

The third section of this chapter then treats the recovery of sparse vectors from noisy linear measurements. Here we distinguish between two different kinds of noise, namely deterministic noise, which we assume to be small with respect to some norm, and random Gaussian noise, where we assume the entries to be normally distributed with small variance.

3.1 Recovery of Sparse Vectors

The basic problem in compressed sensing is to recover signals $x \in \mathbb{R}^d$ from linear measurements

$$y_i = \langle a_i, x \rangle, \qquad i = 1, \dots, m \tag{3.1}$$

for some measurement vectors $a_1, \dots, a_m \in \mathbb{R}^d$, where we assume to have less measurements than unknowns, i.e., $m < d$. If $A \in \mathbb{R}^{m,d}$ denotes the matrix with rows a_i, we set $y = (y_i) \in \mathbb{R}^m$ and we can rewrite (3.1) as

$$y = Ax. \tag{3.2}$$

Since we have less measurements than unknowns, basic linear algebra knowledge tells us that (3.2) has infinitely many solutions. To still be able to recover x as

the unique solution, we have to infer further knowledge. In compressed sensing we impose the structural assumption of sparsity, i.e., the signal has only few nonzero entries where both, the entries and their positions are unknown.

Definition 3.1. A vector $x \in \mathbb{R}^d$ is called *s-sparse* if $\|x\|_0 \leq s$. The set of all s-sparse vectors in $\mathbb{R}^d$ is denoted by

$$\Sigma_s^d := \{x \in \mathbb{R}^d \mid \|x\|_0 \leq s\}.$$

The assumption of sparsity is very restrictive. To get closer to real-life applications, it is important to also consider vectors which are not truly sparse, but can only be approximated very well by sparse vectors. This concept will be referred to as *compressibility*, which we will measure by the best s-term approximation:

Definition 3.2. The *error of the best s-term approximation* of $x \in \mathbb{R}^d$ with respect to the ℓ_p-norm for some $p > 0$ is given by

$$\sigma_s^p(x) := \min_{w \in \Sigma_s^d} \|x - w\|_p.$$

Further, $\hat{x} \in \Sigma_s^d$ is called the *best s-term approximation* of x, if $\|x - \hat{x}\|_p = \sigma_s^p(x)$.

Remark 3.3. The compressibility of a certain vector $x \in \mathbb{R}^d$ will often be expressed by assuming the ratio $\|x\|_p / \|x\|_q$ to be small for $0 < p < q$ and $p \leq 1$. Indeed, in that case it follows (cf. [45, Proposition 2.3])

$$\sigma_s^q(x) \leq \frac{1}{s^{1/p-1/q}} \|x\|_p.$$

A first naive approach to recover the sparse vector x from (3.2) is to solve the minimization problem

$$\Delta_0(y) = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_0 \quad \text{subject to} \quad Aw = y, \tag{P_0}$$

i.e., to search for the sparsest among all solutions of the linear equation. Unfortunately, it turns out that any algorithm solving (P_0) can be used to solve the so-called exact cover problem, which is known to be NP-hard, cf. [9, Theorem 1]. Therefore, the minimization problem (P_0) turns out to be infeasible for practical implementation. To overcome this drawback, in the next sections we will treat the following two topics:

i) We are interested in a (convex) relaxation of (P_0), which can be solved in a reasonable time and gives at least a good approximation of the ground truth signal x.

ii) Not only the reconstruction algorithm is important, but also the measurement matrix A. Hence, we will discuss matrices A allowing a recovery of x from the underdetermined linear system (3.2).

3.1.1 Basis Pursuit

One of the first ideas to relax the ℓ_0-minimization (P_0) is to replace the ℓ_0-norm by the ℓ_p-norm for some $p > 0$ small, since for any $x \in \mathbb{R}^d$ it holds

$$\|x\|_0 = \lim_{p \to 0} \|x\|_p^p.$$

This leads to the minimization problem

$$\Delta_p(y) := \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_p \quad \text{subject to} \quad Aw = y. \tag{P_p}$$

As we can see in Figure 3.1, a first intuition suggests that this minimization problem indeed gives sparse solutions for $0 < p \leq 1$. Second, we observe that the minimization problem (P_p) is convex only for $p \geq 1$. Since convex optimization problems turn out to be practicably solvable [11], in the following we will restrict to the choice $p = 1$, which we will recall as *ℓ_1-minimization* or as *basis pursuit* [25]:

$$\Delta_1(y) := \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_1 \quad \text{subject to} \quad Aw = y. \tag{P_1}$$

Although we restrict to the convex case $p = 1$, there is also lots of work for $p < 1$ as well [24, 43, 77].

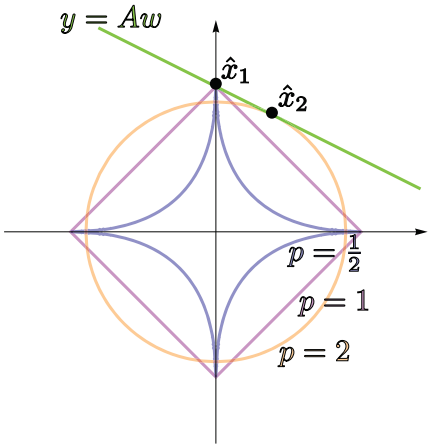

Figure 3.1: Solution of (P_p) for different values of $p \in \{1/2, 1, 2\}$. We observe that the solution $\hat{x}_1$ coincides for the choices $p = 1/2$ and $p = 1$ and, in particular, is 1-sparse in that case. Furthermore, for $p = 2$ the solution $\hat{x}_2$ differs and is not sparse. For the concrete construction of this figure, we have chosen the signal $x = (1, 0)$ and the measurement vector $a_1 = (1/2, 1)^T$. And indeed, the solution $\hat{x}_1$ coincides with x, but the solution $\hat{x}_2$ is quite different.

The next lemma shows that the basis pursuit is not only convex, but can even be reformulated as a linear program. Although this result is already well known, we will recall its proof, since we will use the linear reformulation for numerical experiments later on.

Lemma 3.4 ([25])**.** *The basis pursuit can be formulated as a linear problem, i.e., with linear functional and linear constraints.*

Proof. Splitting every vector $w \in \mathbb{R}^d$ into its positive- and negative part $w_+ := [w]_+ \in \mathbb{R}^d$ and $w_- := [-w]_+ \in \mathbb{R}^d$ (cf. (2.6)), we observe

$$\|w\|_1 = \sum_{j=1}^{d} (w_+)_j + \sum_{j=1}^{d} (w_-)_j,$$
$$y = Aw = Aw_+ - Aw_- = [A, -A] \begin{pmatrix} w_+ \\ w_- \end{pmatrix}.$$

The solution $w^* \in \mathbb{R}^d$ of the basis pursuit (P_1) and the solution $z^* \in \mathbb{R}^{2d}$ of

$$\arg\min_{z \in \mathbb{R}^{2d}} \sum_{j=1}^{2d} z_j \quad \text{subject to} \quad z \geq 0,\ [A, -A]z = y. \tag{3.3}$$

therefore satisfy $z^* = (w_+^*; w_-^*)$. □

3.1.2 Null Space Property

The success of recovery algorithms as the basis pursuit (P_1) not only depend on the sparsity of the signal $x \in \mathbb{R}^d$, but also on the amount of measurements m and the matrix $A \in \mathbb{R}^{m,d}$. For instance, assume that there is a $2s$-sparse vector $v \in \mathbb{R}^d$ in the kernel

$$\ker(A) := \{w \in \mathbb{R}^d \mid Aw = 0\}$$

of A. Then we can split v into two s-sparse vectors $v_1, v_2 \in \mathbb{R}^d$ with $v = v_1 - v_2$ and we obtain

$$0 = Av = A(v_1 - v_2) = Av_1 - Av_2,$$

i.e., $Av_1 = Av_2$. Hence, v_1 and v_2 give the same measurements $y = Av_1 = Av_2$ and we cannot recover v_1 or v_2 only from the knowledge of y and A using any kind of method. We conclude that the kernel of A must not contain any $2s$-sparse vector if we want to recover every s-sparse vector.

Lemma 3.5 ([27])**.** *Let $A \in \mathbb{R}^{m,d}$ and let $2s \leq m$. Then there exists a decoder $\Delta \colon \mathbb{R}^m \to \mathbb{R}^d$ with*

$$\Delta(Ax) = x$$

for all s-sparse signals $x \in \Sigma_s^d$ if and only if

$$\Sigma_{2s}^d \cap \ker(A) = \{0\}.$$

The previous lemma gives a sufficient and necessary condition for the existence of a decoder Δ which recovers all sparse vectors, but it is not clear how this decoder looks like. In particular, Δ can depend on the sparsity level s or on the measurement matrix A. In order to formulate a similar result using the basis pursuit as fixed recovery method, we have to strengthen the requirement on the kernel of A in the following way:

Definition 3.6. A matrix $A \in \mathbb{R}^{m,d}$ is said to have the *null space property* (NSP) of order s if for all $v \in \ker(A)\backslash\{0\}$ and all index sets $T \subset [d]$ with $\#T \leq s$, it holds

$$\|v_T\|_1 < \|v_{T^c}\|_1,$$

where $v_T \in \mathbb{R}^d$ denotes the vector v restricted to the indices of T, i.e., with $(v_T)_i = v_i$ if $i \in T$ and $v_i = 0$ otherwise.

If A satisfies the NSP of order s, it necessarily holds $\ker(A) \cap \Sigma_{2s}^d = \{0\}$. Indeed, assume that there exists a $2s$-sparse vector $v \in \ker(A)$. Let T be the index set of the s largest entries of v (in magnitude). It follows

$$\|v_T\|_1 \geq \|v_{T^c}\|_1$$

in contradiction to the NSP. Further, it turns out that the NSP indeed guarantees the recovery of sparse vectors using the ℓ_1-minimizer (P_1):

Theorem 3.7 ([45])**.** *Every s-sparse vector $x \in \mathbb{R}^d$ is the unique solution of the basis pursuit* (P_1) *if and only if the measurement matrix $A \in \mathbb{R}^{m,d}$ satisfies the null space property of order s.*

The null space property of order s is equivalent to the exact recovery of every s-sparse vector using the ℓ_1-minimizer. But since the vectors we aim to recover in a realistic setting are usually only sparse in idealized situations, we have to adapt the previous theorem for also allowing compressible signals. A first step in this direction is given by the following slightly strengthened version of the NSP:

Definition 3.8. A matrix $A \in \mathbb{R}^{m,d}$ is said to have the *stable null space property* (sNSP) of order s with constant $0 < \rho < 1$, if it holds

$$\|v_T\|_1 \leq \rho \|v_{T^c}\|_1$$

for every $v \in \ker(A)\backslash\{0\}$ and every set $T \subset [d]$ with $\#T \leq s$.

To measure the compressibility of a certain signal, we introduced the notion of the best s-term approximation. Consistently, we would like to control the error of reconstruction between the ground truth compressible signal x and the recovered $\hat{x}$ in terms of the error of the best s-term approximation of x.

Theorem 3.9 ([45])**.** *If $A \in \mathbb{R}^{m,d}$ has the sNSP of order s with constant $0 < \rho < 1$, then, for every $x \in \mathbb{R}^d$, it holds*

$$\|x - \hat{x}\|_1 \leq \frac{2(1+\rho)}{1-\rho} \cdot \sigma_s^1(x), \tag{3.4}$$

where $\hat{x}$ denotes a solution of the basis pursuit (P_1)*.*

Remark 3.10. If x is s-sparse, the error of its best s-term approximation vanishes: $\sigma_s^1(x) = 0$. In that case the previous theorem yields $\|x - \hat{x}\|_1 = 0$, i.e., $\hat{x} = x$, which can also be seen by Theorem 3.7. In that sense the previous Theorem 3.9 can be understood as a generalization of Theorem 3.7, although we have a stricter requirement on A, namely that it satisfies the sNSP instead of the NSP.

3.1.3 Restricted Isometry Property

In this section we introduce the so-called *restricted isometry property*, which was first introduced by Candès and Tao [19] and gives another access to the recovery of sparse signals.

Definition 3.11. A matrix $A \in \mathbb{R}^{m,d}$ satisfies the *restricted isometry property* (RIP) of order s, if there exists a constant $\delta > 0$ such that

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \tag{3.5}$$

holds for every s-sparse $x \in \Sigma_s^d$. The *RIP-constant* $\delta_s > 0$ of A is defined as smallest constant δ, such that (3.5) holds.

If the matrix A satisfies the RIP of order s with small RIP constant δ_s, it almost behaves like an isometry on the set of s-sparse vectors. In particular, in that case the mapping $w \mapsto Aw$ is injective on the set Σ_s^d. Hence, the RIP should guarantee the recovery of sparse signals using a certain decoder. This simple intuition gets further manifested, since the RIP of order $2s$ trivially implies $\Sigma_{2s}^d \cap \ker(A) = \emptyset$, which is, according to Lemma 3.5, a necessary and sufficient condition for the recovery of s-sparse vectors. Moreover, it turns out that we can even use the basis pursuit for the reconstruction:

Theorem 3.12 ([14, 27])**.** *Let $\delta > 0$ and let $A \in \mathbb{R}^{m,d}$ satisfy the RIP of order $2s$ with RIP constant $\delta_{2s} \leq \delta < 1/3$. For every $x \in \mathbb{R}^d$ it then holds*

$$\|x - \Delta_1(Ax)\|_1 \leq C\sigma_s^1(x),$$

where the constant C only depends on δ.

Note that the previous theorem implies the exact reconstruction of s-sparse vectors, hence, combined with Theorem 3.7 we deduce that the RIP of order $2s$ implies the NSP of order s.

Corollary 3.13 ([9, 27])**.** *If $A \in \mathbb{R}^{m,d}$ satisfies the RIP of order $2s$ with RIP constant $\delta_{2s} < 1/3$ then A satisfies the NSP of order s.*

The RIP compares the length of $x \in \mathbb{R}^d$ and the measured $y = Ax \in \mathbb{R}^m$. Similarly we can also compare their scalar products in the following way:

Definition 3.14. A matrix $A \in \mathbb{R}^{m,d}$ is said to have the *restricted orthogonality* of order (s, s') if there exists a constant $\theta > 0$ such that

$$|\langle Av, Aw\rangle| \leq \theta\|v\|_2\|w\|_2 \tag{3.6}$$

holds for all s-sparse vectors $v \in \Sigma_s^d$ and all s'-sparse vectors $w \in \Sigma_{s'}^d$. The smallest value for $\theta > 0$ such that (3.6) holds will be called the *restricted orthogonality constant* and will be denoted by $\theta_{s,s'}$.

Lemma 3.15 ([45])**.** *Let $A \in \mathbb{R}^{m,d}$ satisfy the restricted orthogonality of order (s, s') with constant $\theta_{s,s'}$. Then A also satisfies the RIP of order $s + s'$ and it holds*

$$\theta_{s,s'} \leq \delta_{s+s'} \leq \frac{1}{s+s'}\left(s\delta_s + s'\delta_{s'} + 2\sqrt{ss'}\theta_{s,s'}\right).$$

In particular, if $s' = s$, we get $\theta_{s,s} \leq \delta_{2s}$ and $\delta_{2s} \leq \delta_s + \theta_{s,s}$.

3.2 Tools from Probability Theory

So far we introduced the null space property (NSP), the stable null space property (sNSP) and the restricted isometry property (RIP) as criteria on the matrix A in order to formulate stability results for the recovery of sparse or compressible signals using the basis pursuit. But we have not yet discussed how we can construct matrices satisfying one of the mentioned properties. Furthermore, we have also not yet discussed the dependency on the amount of measurements m needed for the reconstruction, although this is one of the crucial ingredients in the theory of compressed sensing.

To answer both questions, a major breakthrough is obtained by the use of random matrices. It will turn out that random (subgaussian) matrices yield optimal recovery rates and satisfy the NSP and RIP with overwhelmingly high probability, provided that $m = \mathcal{O}(s \log(d))$. This means, that the amount of measurements m only has to grow logarithmically in the underlying dimension d. In particular, if the dimension d is very large, this leads to a heavy reduction of needed measurements compared to usual methods from linear algebra, where one typically needs $m \geq d$.

The aim of this section is to formulate and discuss related results for random matrices. We start introducing basic notations and tools from probability theory, with a final discussion on concentration inequalities such as Hoeffding's inequality or the Bernstein inequality. For a more detailed introduction to probability theory in the framework of compressed sensing, we refer the interested reader to [45, Chapter 7 and 8].

3.2.1 Preliminaries from Probability Theory

Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space over Ω with σ-algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ and probability measure $\mathbb{P}$, where $\mathcal{P}(\Omega)$ denotes the power set of Ω and the probability of an event $B \in \mathcal{A}$ will be denoted by $\mathbb{P}(B)$. A real valued measurable function X over Ω will be called *random variable*. The *expected value* of X is given by

$$\mathbb{E}(X) := \int_\Omega X(\omega) \, \mathrm{d}\, \mathbb{P}(\omega). \tag{3.7}$$

Note $\mathbb{P}(X \geq t) = \mathbb{E}\left(\mathcal{X}_{\{X \geq t\}}\right)$, where $\mathcal{X}_B \colon \Omega \to \mathbb{R}$ denotes the characteristic function of a set $B \in \mathcal{A}$, i.e.,

$$\mathcal{X}_B(\omega) = \begin{cases} 1, & \text{if } \omega \in B, \\ 0, & \text{else.} \end{cases}$$

The function $t \mapsto \mathbb{P}(X \leq t)$ is called *distribution* of X and it holds

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq t) \, \mathrm{d}\, t - \int_0^\infty \mathbb{P}(X \leq -t) \, \mathrm{d}\, t. \tag{3.8}$$

Theorem 3.16 (Markov's inequality)**.** *Let X be a random variable. Then it holds*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}(|X|)}{t} \tag{3.9}$$

for every $t > 0$.

The quantity

$$\mathrm{Var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

is called *variance* of the random variable X and the *covariance* $\mathrm{Cov}(X_1, X_2)$ of two random variables X_1, X_2 is given by $\mathrm{Cov}(X_1, X_2) = \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)))$.

A function $\phi\colon \mathbb{R} \to [0, \infty)$ is called *density function* of X if

$$\mathbb{P}(t_1 \leq X \leq t_2) = \int_{t_1}^{t_2} \phi(t)\,\mathrm{d}\,t$$

holds for all $t_1 < t_2$. If X has a density function ϕ and $g\colon \mathbb{R} \to \mathbb{R}$ is measurable, it follows

$$\mathbb{E}\left(g(X)\right) = \int_{\mathbb{R}} g(t)\phi(t)\,\mathrm{d}\,t. \tag{3.10}$$

We will call the random variables $X_1, \ldots, X_m$ to be *independent*, if

$$\mathbb{P}(X_1 \leq t_1, \ldots, X_n \leq t_m) = \mathbb{P}(X_1 \leq t_1) \cdot \ldots \cdot \mathbb{P}(X_m \leq t_m)$$

holds for every $t_1, \ldots, t_m \in \mathbb{R}$. In particular, if X_1, X_2 are independent, it follows

$$\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\,\mathbb{E}(X_2)$$

and if they have density functions ϕ_1, ϕ_2, their joint density function ϕ is given by

$$\phi(t_1, t_2) = \phi_1(t_1)\phi_2(t_2). \tag{3.11}$$

That is, for any measurable set $B \subset \mathbb{R}^2$, it holds

$$\mathbb{P}((X_1, X_2) \in B) = \int_B \phi_1(t_1)\phi_2(t_2)\,\mathrm{d}\,t_1\,\mathrm{d}\,t_2.$$

A collection $X_1, \ldots, X_m$ of independent random variables all of which have the same distribution are called *independent and identically distributed* (i.i.d.).

If the entries of a vector $X \in \mathbb{R}^d$ are random variables, X is called a *random vector* and if the entries a_{ij} of a matrix $A \in \mathbb{R}^{m,d}$ are random variables, we call A a *random matrix*.

Theorem 3.17 (Jensen's inequality)**.** *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex and let $X \in \mathbb{R}^d$ be a random vector. Then it holds*

$$f(\mathbb{E}\,X) \leq \mathbb{E}\,f(X).$$

3.2.2 Bernoulli and Gaussian Variables

A random variable X is called *subgaussian*, if there exist some constants $\beta, \kappa > 0$, such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t^2} \tag{3.12}$$

holds for every $t > 0$. In this section we will concentrate on two important examples of subgaussians, namely Gaussian and Bernoulli variables.

Definition 3.18. i) A random variable ξ given by

$$\xi = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2 \end{cases} \tag{3.13}$$

is called *Bernoulli variable*.

ii) A random variable g with $\mathbb{E}\, g = \mu$, $\operatorname{Var} g = \sigma^2$ and density function ϕ given by

$$\phi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right), \tag{3.14}$$

is called *normally distributed* or *Gaussian variable*. In that case we simply write $g \sim \mathcal{N}(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$, i.e., $g \sim \mathcal{N}(0, 1)$, we will call g *standard normally distributed*.

Remark 3.19. i) If a_{ij} are i.i.d. Bernoulli or Gaussian variables, we will call $A = (a_{ij}) \in \mathbb{R}^{m,d}$ *Bernoulli-* or *Gaussian matrix*, respectively. Further, the scaled matrix $A/\sqrt{m}$ will be called normalized Bernoulli- or Gaussian matrix.

ii) In the literature Bernoulli variables are often allowed to have arbitrary entries from a discrete set, each with a constant probability. Here we restrict ourselves to the entries ± 1, both with a probability of $1/2$. These are often also called *Rademacher variables*.

iii) A *Gaussian vector* $g \in \mathbb{R}^d$ with normally distributed entries $g_1, \dots, g_d$ will be denoted by

$$g \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma \in \mathbb{R}^{d,d}$ denotes the symmetric covariance matrix with entries $\sigma_{ij} = \operatorname{Cov}(g_i, g_j)$ and $\mu \in \mathbb{R}^d$ is the vector of expected values $\mu_j = \mathbb{E}\, g_j$.

Bernoulli and Gaussian variables are indeed subgaussian: For $t > 0$ and a Bernoulli variable ξ we get

$$\mathbb{P}(|\xi| \geq t) = \begin{cases} 1, & \text{if } t \leq 1, \\ 0, & \text{if } t > 1. \end{cases}$$

In both cases we can estimate this probability from above by $e^1 e^{-t^2}$. For a mean-zero Gaussian variable $g \sim \mathcal{N}(0, \sigma^2)$ we get

$$\begin{aligned} \mathbb{P}(|g| \geq t) &= \frac{2}{\sqrt{2\pi\sigma^2}} \int_t^\infty \exp\left(\frac{-s^2}{2\sigma^2}\right) \mathrm{d}\, s = \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \int_0^\infty \exp\left(\frac{-(s+t)^2}{2\sigma^2}\right) \mathrm{d}\, s \\ &\leq \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \exp\left(\frac{-t^2}{2\sigma^2}\right) \int_0^\infty \exp\left(\frac{-s^2}{2\sigma^2}\right) \mathrm{d}\, s = \exp\left(\frac{-t^2}{2\sigma^2}\right). \end{aligned}$$

Let $g \sim \mathcal{N}(0, \sigma^2)$ be normally distributed. Then $\operatorname{sign}(g)$ is a Bernoulli variable and if ξ is a Bernoulli variable independent of g, their product $\xi g \sim \mathcal{N}(0, \sigma^2)$ is again normally distributed. Further, if $g \sim \mathcal{N}(\mu, \sigma^2)$, we get

$$\lambda(g - \tau) \sim \mathcal{N}\left(\lambda(\mu - \tau), \lambda^2\sigma^2\right) \tag{3.15}$$

for any $\lambda, \tau \in \mathbb{R}$ and if $g_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $g_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, it holds

$$g_1 + g_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right), \tag{3.16}$$

which is recalled as *2-stability of Gaussian variables*. Combining (3.15) and (3.16), for any $w \in \mathbb{R}^d$ we in particular get

$$\frac{1}{\|w\|_2} \cdot \left\langle \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \tilde{a}_i, w \right\rangle \sim \mathcal{N}(0,1) \tag{3.17}$$

for i.i.d. random vectors $\tilde{a}_1, \dots, \tilde{a}_m \in \mathbb{R}^d$ with $\tilde{a}_i \sim \mathcal{N}(0, \mathrm{id})$. The next lemma collects further basic properties of Gaussian variables and vectors which we will need later on, cf. [52, Corollary 5.2] and [45, Proposition 8.1].

Lemma 3.20. *i) For $g \sim \mathcal{N}(0,1)$ it holds*

$$\mathbb{E}\,|g| = \frac{\sqrt{2}}{\sqrt{\pi}}. \tag{3.18}$$

ii) If $g \in \mathbb{R}^d$ has i.i.d. entries $g_j \sim \mathcal{N}(0,1)$ and $v, w \in \mathbb{R}^d$ are orthogonal, the random variables $\langle g, w\rangle$ and $\langle g, v\rangle$ are independent.

iii) Let $g \in \mathbb{R}^d$ be a random Gaussian vector with entries $g_j \sim \mathcal{N}(0,1)$. Then it holds

$$\frac{\sqrt{2d}}{\sqrt{\pi}} \le \mathbb{E}\,\|g\|_2 \le \sqrt{d} \quad \text{and} \quad \mathbb{E}\,\|g\|_\infty \le \sqrt{2\log(2d)}. \tag{3.19}$$

Further, if the g_j are independent and $d \ge 2$, we obtain

$$\frac{\sqrt{\log(d)}}{4} \le \mathbb{E}\,\|g\|_\infty. \tag{3.20}$$

3.2.3 Concentration Inequalities

The *concentration of measure* phenomenon describes that (Lipschitz-) functions on a high-dimensional probability space often heavily concentrate around their means. One of the typical examples in this direction (without randomness) is that most of the volume of the unit ball in $\mathbb{R}^d$ concentrates around its equator.

In the area of compressed sensing we often encounter sums of random variables: If we choose the sensing matrix $A = (a_{ij}) \in \mathbb{R}^{m,d}$ to be a random Bernoulli or Gaussian matrix, we obtain the measurements

$$y_i = (Ax)_i = \langle a_i, x\rangle = \sum_{j=1}^{d} a_{ij} x_j.$$

Hoeffding's inequality and the Bernstein inequality provide useful tools to bound the tails of such sums and are two out of many examples for the concentration of measure phenomenon. In particular, Bernstein inequality can be used to prove the RIP for subgaussian matrices. For a detailed overview on the concentration of measure phenomenon we refer the interested reader to [74].

Theorem 3.21 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_m$ be a sequence of independent mean-zero random variables with $|X_i| \leq B_i$ almost surely for all $i = 1, \dots, m$ and constants $B_i > 0$. Then, for all $t > 0$, it holds*

$$\mathbb{P}\left(\sum_{i=1}^{m} X_i \geq t\right) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{m} B_i^2}\right)$$

and

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} X_i\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{m} B_i^2}\right).$$

Applying Hoeffding's inequality to a sum of Bernoulli variables we deduce the following corollary.

Corollary 3.22. *Let $\xi_1, \dots, \xi_m$ be i.i.d. Bernoulli variables. Then it holds*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} \xi_i\right| \geq u\right) \leq 2\exp\left(\frac{-mu^2}{2}\right) \tag{3.21}$$

for any $u > 0$.

The Bernstein inequality provides a useful generalization of Hoeffding's inequality to sums of unbounded independent random variables as Gaussian variables. However, note that the following theorem is only one out of many possible versions, which can be found in [45, Corollary 7.32].

Theorem 3.23 (Bernstein inequality)**.** *Let $X_1, \dots, X_m$ be a sequence of independent mean-zero random variables with $\mathbb{P}(|X_i| \geq t) \leq \beta e^{-\kappa t}$ for some constants $\beta, \kappa > 0$, all $t > 0$ and $i = 1, \dots, m$. Then it holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} X_i\right| \geq t\right) \leq 2\exp\left(\frac{-(\kappa t)^2/2}{2\beta m + \kappa t}\right).$$

Using the Bernstein inequality we deduce the following concentration inequality, which is the main ingredient to show the RIP for subgaussian matrices, cf. [45, Lemma 9.8].

Lemma 3.24. *Let $A = (a_{ij}) \in \mathbb{R}^{m,d}$ be a random matrix with i.i.d. subgaussian entries such that $\mathbb{E}\langle a_i, x\rangle^2 = \|x\|_2^2$ holds for all $x \in \mathbb{R}^d$. It follows*

$$\mathbb{P}\left(\left|m^{-1}\|Ax\|_2^2 - \|x\|_2^2\right| \geq t\|x\|_2^2\right) \leq 2\exp\left(-c't^2 m\right) \tag{3.22}$$

for all $x \in \mathbb{R}^d$ and $0 < t < 1$, where the constant c' only depends on the subgaussian parameter β, κ.

Remark 3.25. If A is a random Bernoulli or Gaussian matrix with standard normally distributed entries, it follows

$$\mathbb{E}\left(\langle a_i, x\rangle^2\right) = \sum_{j,k=1}^{d} \mathbb{E}(a_{ij}a_{ik})x_j x_k = \sum_{j=1}^{d} x_j^2 = \|x\|_2^2$$

and we can apply the previous Lemma 3.24. In that case, the right hand side of (3.22) can even be improved by $2\exp\left(-m(t^2/4 - t^3/6)\right)$, cf. [45, Exercise 9.2].

The next theorem gives a deviation inequality for the supremum of Gaussian processes, cf. [74, Theorem 7.1] or [96, Theorem 5.2].

Theorem 3.26 (Gaussian Concentration Inequality)**.** *Let T be a finite set and $G = (G_t)_{t\in T}$ a centered Gaussian process, i.e., G_t is normally distributed with $\mathbb{E}(G_t) = 0$ for any $t \in T$. Then, for any $u > 0$, one has*

$$\mathbb{P}\left(\sup_{t\in T} G_t \geq \mathbb{E}\sup_{t\in T} G_t + u\right) \leq \exp\left(\frac{-u^2}{2\sigma^2}\right),$$

where $\sigma^2 = \sup_{t\in T} \mathbb{E}\, G_t^2 < \infty$.

Remark 3.27. To avoid measurability issues on the supremum of random variables, we assumed the set T to be finite. However, the result can be extended to (infinite) separable sets of a metric space, e.g. for subsets T of the open unit ball $B_p^d \subset \mathbb{R}^d$.

3.2.4 RIP and NSP for Random Matrices

Based on the concentration inequality (3.22) we obtain that random subgaussian matrices satisfy the RIP with overwhelmingly high probability, cf. [6, Theorem 5.2].

Theorem 3.28 ([6])**.** *For $0 < \delta < 1$ let $m \geq c_0 s\log(d/s)$ and let $A \in \mathbb{R}^{m,d}$ be a random matrix satisfying* (3.22)*. Then $A/\sqrt{m}$ satisfies the RIP of order s with RIP constant $0 < \delta_s \leq \delta$ with probability at least $1 - 2e^{-c_1 m}$, where the constants c_0 and c_1 may only depend on δ.*

Remark 3.29. Following Lemma 3.24, subgaussian matrices, such as Gaussian- or Bernoulli matrices, satisfy the concentration inequality (3.22), so they also satisfy the RIP with high probability.

From Theorem 3.28 we obtain that random subgaussian matrices indeed satisfy the RIP and the NSP with overwhelmingly high probability. But there is also another surprising fact: the number of measurements only has to scale as $m = \mathcal{O}(s\log(d/s))$. That is, we only need about $s\log(d)$ measurements to exactly recover an s-sparse signal in $\mathbb{R}^d$, i.e., by not knowing the position of the nonzero entries we only lose a factor $\log(d)$. Furthermore, ignoring the logarithmic term, m has to scale linear with the unknowns s, which clearly is the best we can hope for.

Figure 3.2 shows a numerical example for the dependency of m on d and on s. Here we can also observe another typical phenomenon in the theory of compressed sensing, namely that a *phase transition phenomenon* appears, i.e., that there is a sharp boundary between configurations where the recovery works and where it does not work. This goes, in particular, in hand with the theory of the concentration of measure phenomenon, where one observes that random constructions either work or fail with overwhelmingly high probability in high dimensions.

Closely related to the RIP is the concentration inequality (3.22), which shows that the length of a vector x is almost preserved by a subgaussian matrix A. Applied to the pairwise distances of n points $x_1, \ldots, x_n \in \mathbb{R}^d$, we obtain the following lemma of Johnson and Lindenstrauss.

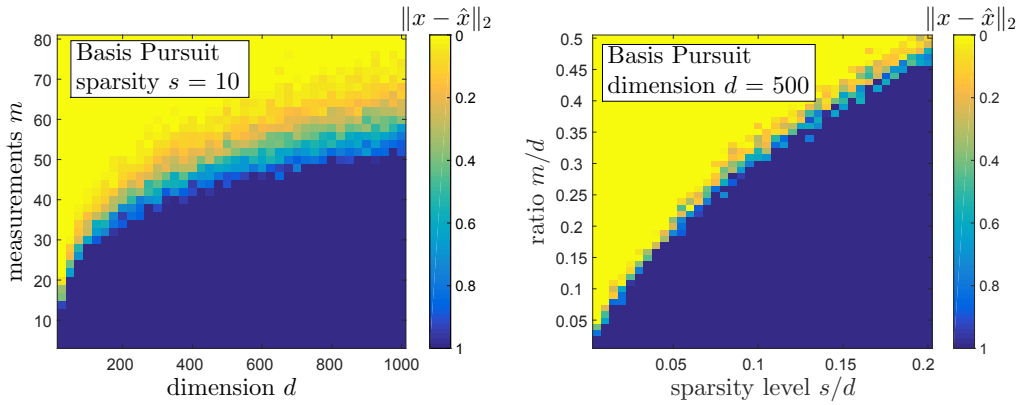

Figure 3.2: Illustration of the dependency of m on d (left) and s (right) of the basis pursuit with respect to a subgaussian measurement matrix. For the left image we choose $s = 10$ fixed. Afterwards, for values of d ranging from 0 to 1000 and values of m ranging from 0 to 80 we did the following experiment $n = 100$ times and plotted the average error: For each pair (d, m) we generated a random s-sparse vector $x \in \mathbb{R}^d$ using the MATLAB command **sprandn**. Afterwards, we generated a normalized Gaussian matrix $A \in \mathbb{R}^{m,d}$ (cf. Remark 3.19) using the MATLAB command **randn**. Once we have A and x, we set $y = Ax$ and we then run the basis pursuit (P_1) to obtain the recovered vector $\hat{x}$. Afterwards, we just calculated their Euclidean distance $\|x - \hat{x}\|_2$. The right images were constructed analogously, but there the dimension $d = 500$ was fixed and we let the ratios s/d range from 0 to 0.2 and m/d range from 0 to 0.5.

Theorem 3.30 ([62]). *Let $0 < \varepsilon < 1$ and $x_1, \dots, x_n \in \mathbb{R}^d$. If $m \geq C \log(n)\varepsilon^{-2}$ for some constant C, then there exists a Lipschitz mapping $L\colon \mathbb{R}^d \to \mathbb{R}^m$ such that*

$$(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|L(x_i) - L(x_j)\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$$

holds for all $i, j = 1, \dots, n$.

3.3 Recovery from Noisy Measurements

In this section we focus on the recovery of sparse vectors $x \in \mathbb{R}^d$ from noisy linear measurements of the form

$$y = Ax + \eta + z.$$

The term $\eta \in \mathbb{R}^m$ represents *deterministic noise*, e.g., given by rounding errors or calculations up to machine precision. Those errors can typically be controlled, since we know their appearance in advance. Hence, we assume η to be small with respect to some norm on $\mathbb{R}^m$. The second term $z \in \mathbb{R}^m$ represents *random Gaussian noise* with normally distributed entries.

3.3.1 Deterministic Noise

In this subsection we deal with deterministic noise $\eta \in \mathbb{R}^m$ only, i.e., we want to recover the signal x from linear measurements

$$y = Ax + \eta. \tag{3.23}$$

Here we assume η to be small in the sense that $\|\eta\| \leq \varepsilon$ for some $\varepsilon > 0$ and some norm $\|\cdot\|$ on $\mathbb{R}^m$. We aim for a reconstruction scheme such that the error between the recovered and the true signal can be controlled by the noise level ε and the compressibility of x. For this purpose we replace the basis pursuit (P_1) by the convex minimization problem

$$\Delta_{1,\varepsilon}(y) = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_1 \quad \text{subject to} \quad \|Aw - y\| \leq \varepsilon, \tag{$\mathrm{P}_{1,\varepsilon}$}$$

which we will recall as *$\ell_{1,\varepsilon}$-minimizer* or as *basis pursuit denoising*. The following strengthened version of the stable null space property guarantees the satisfying reconstruction of x using ($\mathrm{P}_{1,\varepsilon}$).

Definition 3.31. The matrix $A \in \mathbb{R}^{m,d}$ is said to satisfy the *robust null space property* (rNSP) of order s with constants $0 < \rho < 1$ and $\tau > 0$ and with respect to the norm $\|\cdot\|$, if for all $T \subset [d]$ with $\#T \leq s$ and all $v \in \mathbb{R}^d$, it holds

$$\|v_T\|_1 \leq \rho \|v_{T^c}\|_1 + \tau \|Av\|.$$

We obtain the following result, cf. [45, Theorem 4.19].

Theorem 3.32. *Suppose that $A \in \mathbb{R}^{m,d}$ satisfies the rNSP of order s with constants ρ, τ. Any solution $\hat{x} \in \mathbb{R}^d$ of the $\ell_{1,\varepsilon}$-minimizer ($\mathrm{P}_{1,\varepsilon}$) with $y = Ax + \eta$ and $\|\eta\| \leq \varepsilon$ then satisfies*

$$\|x - \hat{x}\|_1 \leq \frac{2(1+\rho)}{1-\rho} \sigma_s^1(x) + \frac{4\tau}{1-\rho} \varepsilon. \tag{3.24}$$

Remark 3.33. The previous Theorem 3.32 does not require any assumption on the signal x. However, the error estimate depends on the best s-term approximation of x, hence, it only gives a useful error bound if x is compressible.

Furthermore, if $\varepsilon = 0$, the recovery algorithm ($\mathrm{P}_{1,\varepsilon}$) coincides with ($\mathrm{P}_1$). In that case also the estimate (3.24) of Theorem 3.32 coincides with the estimate (3.4) of Theorem 3.9, where we assumed A to satisfy the sNSP instead of the rNSP. If additionally x is even s-sparse, Theorem 3.32 yields a perfect reconstruction of x as Theorem 3.7 for the NSP.

The following theorem shows that the RIP with sufficient small RIP constant implies the rNSP. Consequently, the RIP yields error bounds on the reconstruction, cf. [45, Theorem 6.12 and 6.13].

Theorem 3.34. *If $A \in \mathbb{R}^{m,d}$ satisfies the RIP of order $2s$ with RIP constant $\delta_{2s} < 4/\sqrt{41}$, then A satisfies the rNSP of order s with constants ρ, τ only depending on δ_{2s}.*

Theorem 3.35. *Let $A \in \mathbb{R}^{m,d}$ satisfy the RIP of order $2s$ with RIP constant $\delta_{2s} \leq 4/\sqrt{41}$. For any $x \in \mathbb{R}^d$ let $\hat{x} \in \mathbb{R}^d$ be a solution of the $\ell_{1,\varepsilon}$-minimizer $(\mathrm{P}_{1,\varepsilon})$ with $y = Ax + \eta$ and $\|\eta\|_2 \leq \varepsilon$. Then it follows*

$$\|x - \hat{x}\|_1 \leq C\sigma_s^1(x) + C'\sqrt{s}\varepsilon, \tag{3.25}$$

$$\|x - \hat{x}\|_2 \leq \frac{C'}{\sqrt{s}}\sigma_s^1(x) + C'\varepsilon \tag{3.26}$$

for some constants $C, C' > 0$ only depending on δ_{2s}.

The $\ell_{1,\varepsilon}$-minimizer $(\mathrm{P}_{1,\varepsilon})$ needs prior knowledge of some $\varepsilon > 0$ such that $\|\eta\| \leq \varepsilon$. Since η corresponds to deterministic noise as rounding errors, in some situations such an ε can be calculated in advance. However, in some applications it might happen that we cannot give an explicit choice for ε, hence, we cannot use the $\ell_{1,\varepsilon}$-minimizer there. A reasonable hope in these applications is that the small error η does not have a big impact, so that we can still use the basis pursuit (P_1) to achieve satisfactory reconstructions.

The idea to formulate an appropriate theorem is to search for a small vector $u \in \mathbb{R}^d$ in the preimage of η under A such that

$$y = Ax + \eta = A(x + u).$$

Then we can use the ℓ_1-minimizer to recover the compressible vector $x + u$ from y, which should be a good approximation to x.

Let us stress that the existence of $u \in \mathbb{R}^d$ is not the main problem here, since $A \in \mathbb{R}^{m,d}$ maps from the higher-dimensional space $\mathbb{R}^d$ to lower-dimensional space $\mathbb{R}^m$ and should be at least surjective. But the existence of u is not what we actually need, since we have to ensure $x+u$ to be compressible, which in general only holds if u has a small length. Before going further let us first introduce the so-called J-norm, which is given by

$$\|\eta\|_J := \max\left\{\sqrt{m}\|\eta\|_\infty; \sqrt{\frac{m}{\log(d/m)}}\|\eta\|_2\right\}. \tag{3.27}$$

The closed unit ball with respect to the J-norm will be denoted by $U_J := \{\eta \in \mathbb{R}^m \mid \|\eta\|_J \leq 1\}$.

Theorem 3.36 ([34, 73])**.** *Let $A \in \mathbb{R}^{m,d}$ be a normalized Bernoulli matrix and let $d \geq \log(6)^2 m$. Then there exists a constant C such that with probability at least*

$$1 - e^{-\sqrt{md}}$$

for every $\eta \in U_J$ there exists a $u \in \mathbb{R}^d$ with $Au = \eta$ and $\|u\|_1 \leq C$.

This theorem gives a bound on the ℓ_1-norm on u. With a slight modification we can also bound the ℓ_2-norm of u as follows, cf. [45, Chapter 11] and also [34]:

Theorem 3.37 ([34, 45])**.** *Let $A \in \mathbb{R}^{m,d}$ be a normalized Bernoulli matrix according to Remark 3.19 with $d \geq \log(6)^2 m$ and let $\delta > 0$. Then there exist some universal*

constants C_1, C_2 and some constants C_3, C_4 depending on δ such that with probability at least

$$1 - 2e^{-C_1 m} - e^{-\sqrt{md}}$$

for each $\eta \in U_J$, there exists a $u \in \mathbb{R}^d$ with $Au = \eta$ and

$$\|u\|_1 \leq C_3, \qquad \|u\|_2 \leq C_4 \sqrt{\log(d/m)/m}.$$

3.3.2 Gaussian Noise

Now we want to discuss the reconstruction of sparse vectors from noisy measurements of the form

$$y = Ax + z, \tag{3.28}$$

where the entries $z_i \sim \mathcal{N}(0, \sigma^2)$ of z are i.i.d. normally distributed with small variance $\sigma > 0$. In contrary to deterministic noise e from the previous section, the expected norm of z grows with the number of measurements m:

$$\mathbb{E}\,\|z\|_2^2 = \sum_{i=1}^{m} \mathbb{E}\, z_i^2 = \sum_{i=1}^{m} \sigma^2 = m\sigma^2.$$

In this case the approaches from deterministic noise lead to error bounds depending on the norm of z, hence they grow with the number of measurements m. This phenomenon is recalled as *noise folding*, cf. [3] for more details. To overcome the drawback of noise folding, Candès and Tao suggested to use the so-called Dantzig selector [20].

Definition 3.38. For a matrix $A \in \mathbb{R}^{m,d}$ and constants $\lambda_d, \sigma > 0$ the *Dantzig selector* $\Delta_{DS}(y)$ of some input vector $y \in \mathbb{R}^m$ is defined as

$$\Delta_{DS}(y) := \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \|w\|_1 \quad \text{subject to} \quad \|A^T(y - Aw)\|_\infty \leq \lambda_d \sigma. \tag{3.29}$$

Assume that $A \in \mathbb{R}^{m,d}$ is a normalized Bernoulli matrix with entries $\pm 1/\sqrt{m}$ and let $y = Ax + z$ be according to (3.28). For $1 \leq j \leq d$ we then get

$$(A^T z)_j = \sum_{i=1}^{m} a_{ij} z_i = \sigma g_j$$

for some i.i.d. $g_j \sim \mathcal{N}(0, 1)$, where we used the 2-stability of Gaussian variables (3.16). We apply the estimate (3.19) on the ℓ_∞-norm of a Gaussian vector to end up with

$$\mathbb{E}\,\|A^T(y - Ax)\|_\infty = \mathbb{E}\,\|A^T z\|_\infty \leq \sigma \sqrt{2\log(2d)}.$$

Following this simple calculation, we get the idea that λ_d should, up to some constants, scale as $\sqrt{\log(d)}$ in the definition of the Dantzig selector.

Theorem 3.39 ([20]). *Let $A \in \mathbb{R}^{m,d}$ be a normalized Bernoulli matrix with RIP constant δ_{2s} and restricted orthogonality constant $\theta_{s,2s}$ obeying*

$$\delta_{2s} + \theta_{s,2s} < 1.$$

For $x \in \mathbb{R}^d$ with $\|x\|_{p,\infty} \leq R$ for some $R > 0$, $p \leq 1$, $z \in \mathbb{R}^m$ with i.i.d. entries $z_i \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$ and $\lambda_d = \sqrt{2\log(d)}$, it then holds

$$\|\Delta_{DS}(Ax + z) - x\|_2^2 \leq \min_{1 \leq t \leq s} 2C \log(d) \left(t\sigma^2 + R^2 t^{1-2/p} \right) \tag{3.30}$$

with high probability and a constant C only depending on δ_{2s} and $\theta_{s,2s}$.

Remark 3.40.

i) If $m = \mathcal{O}(s \log(d))$, Theorem 3.28 shows that a normalized Bernoulli matrix $A \in \mathbb{R}^{m,d}$ has the RIP of order $3s$ such that $\delta_{2s} + \theta_{s,2s} < 1$ holds with high probability. Hence, in the presence of random Gaussian noise, we need the same order of measurements as in the noiseless case.

ii) The main advantage of Theorem 3.39 compared to previous results for deterministic noise is that the error bound does not grow with the number of measurements m, although the expected norm of the noise z does. Nevertheless, getting more information on the signal by increasing the number of measurements should lead to better reconstruction and smaller error, but the estimate (3.30) only implicitly depends on m via s.

Chapter 4

s-Numbers and Carl's Inequality

In this chapter we will give a short introduction to the theory of s-numbers for operators between quasi-Banach spaces. The main motivation in the context of this thesis is given by Carl's inequality [21], which compares the asymptotic behavior of the entropy numbers with certain important so-called s-functions. One particular instance of those s-functions are the Gelfand numbers, which are equivalent to the so-called compressive n-width. The compressive n-width describe the performance of the measure and reconstruction procedure in compressed sensing and by applying Carl's inequality we can deduce a lower bound on the minimal needed amount of measurements.

In this chapter we mainly present the results of [57], which was joint work with Aicke Hinrichs and Jan Vybíral.

4.1 Quasi-Banach Spaces

In this section we collect some basic concepts about quasi-Banach spaces. For more details we refer the interested reader to [64].

If X is a real vector space, we say that a map $\|\cdot\|_X \colon X \to [0,\infty)$ defines a *quasi-norm* on X if

i) $\|x\|_X = 0$ only holds for $x = 0$,

ii) $\|\lambda x\|_X = |\lambda| \cdot \|x\|_X$ holds for any $\lambda \in \mathbb{R}$ and $x \in X$ and

iii) if there exists a constant $C_X \geq 1$ such that the *quasi-triangle inequality*

$$\|x_1 + x_2\|_X \leq C_X(\|x_1\|_X + \|x_2\|_X) \tag{4.1}$$

holds for all $x_1, x_2 \in X$. We will recall C_X as *quasi-triangle constant* of X.

The open unit ball of X endowed with the quasi-norm $\|\cdot\|_X$ will be denoted by

$$B_X := \{x \in X \mid \|x\|_X < 1\}.$$

If the quasi-norm $\|\cdot\|_X$ satisfies the so-called *p-triangle inequality*

$$\|x_1 + x_2\|_X^p \leq \|x_1\|_X^p + \|x_2\|_X^p \quad \text{for all } x_1, x_2 \in X \tag{4.2}$$

for some $0 < p \leq 1$, we will call it a *p-norm*. The fundamental Theorem of Aoki-Rolewicz states that every quasi-norm is at least equivalent to some p-norm:

Theorem 4.1 ([1, 99])**.** *Every quasi-norm is equivalent to some p-norm.*

Definition 4.2. Let $\|\cdot\|_X$ be equivalent to some p-norm $|||\cdot|||_X$ on X. If X is complete with respect to the metric induced by $|||\cdot|||_X^p$, it is called a *quasi-Banach space*. Further, if the norm $\|\cdot\|_X$ itself defines a p-norm, we will recall X as a *p-Banach space*.

If $M \subset X$ is a linear subspace of X, we define the quotient space X/M as set of all equivalence classes $[x]$, where $x_1, x_2 \in X$ are said to be equivalent if $x_1 - x_2 \in M$. The usual definition of the addition $[x_1] + [x_2] := [x_1 + x_2]$ and scalar multiplication $\lambda \cdot [x] := [\lambda x]$ makes X/M again a vector space. Further, if $\|\cdot\|_X$ defines a (quasi)-norm on X, the *quotient norm* $\|\cdot\|_{X/M}$ on X/M is defined by

$$\|[x]\|_{X/M} := \inf_{y \in M} \|x - y\|_X.$$

Lemma 4.3. *Let X be a p-Banach space and let M be a linear subspace of X. Then $\|\cdot\|_{X/M}$ defines a p-norm on X/M.*

Proof. To show the p-triangle inequality let $[x_1], [x_2] \in X/M$ and let $\varepsilon > 0$. Then, by definition of $\|\cdot\|_{X/M}$, there exist $v_1, v_2 \in M$ such that

$$\|x_1 - v_1\|_X \leq (1+\varepsilon)\|[x_1]\|_{X/M} \quad \text{and} \quad \|x_2 - v_2\|_X \leq (1+\varepsilon)\|[x_2]\|_{X/M}.$$

Since $\|\cdot\|_X$ defines a p-norm on X we deduce

$$\begin{aligned}\|[x_1 + x_2]\|_{X/M}^p &\leq \|x_1 + x_2 - v_1 - v_2\|_X^p \leq \|x_1 - v_1\|_X^p + \|x_2 - v_2\|_X^p \\ &\leq (1+\varepsilon)^p(\|[x_1]\|_{X/M}^p + \|[x_2]\|_{X/M}^p).\end{aligned}$$

Letting $\varepsilon \to 0$ yields the claim. □

If $T\colon X \to Y$ is a linear operator between the two quasi-Banach spaces X and Y, its operator norm is given by

$$\|T\| := \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \|Tx\|_Y.$$

We will call the operator T to be bounded, if its operator norm is finite. The set of all bounded linear operators from X to Y will be denoted by $\mathcal{L}(X, Y)$.

Lemma 4.4. *Let $T, T'\colon X \to Y$ and $S\colon Y \to Y_0$ be bounded linear operator between the quasi-Banach spaces X, Y, Y_0. Then it holds*

i) $\|Tx\|_Y \leq \|T\| \cdot \|x\|_X$ *for every* $x \in X$,

ii) $\|ST\| \leq \|S\| \cdot \|T\|$,

iii) $\|T + T'\| \leq C_Y(\|T\| + \|T'\|)$, *where C_Y denotes the quasi-triangle constant of Y and*

iv) it holds

$$\|T\| = \inf\{\varepsilon > 0 \mid T(B_X) \subset \varepsilon B_Y\}.$$

Proof. Although the properties listed above are well known, we recall their proofs to verify their validity for quasi-Banach spaces.

i) For any $x \in X$ we get

$$\|Tx\|_Y = \left\| T\left(\frac{x}{\|x\|_X}\right)\right\|_Y \cdot \|x\|_X \leq \sup_{\substack{x' \in X \\ \|x'\|_X \leq 1}} \|Tx'\|_Y \cdot \|x\|_X = \|T\| \cdot \|x\|_X.$$

ii) By plugging in the definition of the operator norm and using i), we obtain

$$\|ST\| = \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \|STx\|_{Y_0} \leq \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \|S\| \cdot \|Tx\|_Y = \|S\| \cdot \|T\|.$$

iii) We obtain

$$\begin{aligned}\|T + T'\| = \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \|(T + T')x\|_Y &\leq \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} C_Y(\|Tx\|_Y + \|T'x\|_Y) \\ &\leq C_Y\Big(\sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \|Tx\|_Y + \sup_{\substack{x' \in X \\ \|x'\|_X \leq 1}} \|T'x'\|_Y\Big) = C_Y(\|T\| + \|T'\|),\end{aligned}$$

as claimed.

iv) Using i), for every $x \in B_X$ we get

$$\|T(x)\|_Y \leq \|T\| \cdot \|x\|_X \leq \|T\|,$$

hence, $T(B_X) \subset \|T\| \cdot B_Y$ and it follows

$$\inf\{\varepsilon > 0 \mid T(B_X) \subset \varepsilon B_Y\} \leq \|T\|.$$

To show the reverse inequality, let $0 < \varepsilon < \|T\|$. By definition of the operator norm, there then exists some $x_0 \in B_X$ with $\|Tx_0\|_Y > \varepsilon$, which implies $T(B_X) \not\subset \varepsilon B_Y$.

□

4.2 s-Numbers

For the sake of completeness, we start this section with the general setting of the theory of s-numbers and afterwards, we will have a closer look on the approximation-, entropy and Gelfand numbers. For a detailed introduction to the theory of s-numbers we refer the interested reader to [23, 90–92].

Let X, Y denote two quasi-Banach spaces. A map s which maps any bounded linear operator $T \in \mathcal{L}(X, Y)$ to the real valued sequence $(s_n(T))_{n \in \mathbb{N}}$ is called *s-function* or *s-scale* and $s_n(T)$ is called n-th *s-number* of T, if it holds

i) $\|T\| = s_1(T)$,

ii) $0 \leq s_{n+1}(T) \leq s_n(T)$,

iii) $s_n(T+T_0) \leq C_Y\big(s_n(T)+\|T_0\|\big)$, where C_Y denotes the quasi-triangle constant of Y and $T_0\colon X \to Y$ is another bounded linear operator,

iv) $s_n(LTR) \leq \|L\| \cdot s_n(T) \cdot \|R\|$ for all bounded linear operators $L\colon Y \to Y_0$ and $R\colon X_0 \to X$ for some quasi-Banach spaces X_0, Y_0,

v) $\operatorname{rank}(T) < n$ implies $s_n(T) = 0$ and

vi) $s_n(\mathrm{id}\colon \ell_2^n \to \ell_2^n) = 1$

for every $n \in \mathbb{N}$. If only conditions i)-iv) are satisfied, s is called a *pseudo s-function.*

4.2.1 Approximation Numbers

Definition 4.5. The n-th *approximation number* $e_n(T)$ of a bounded linear operator $T\colon X \to Y$ between the two quasi-Banach spaces X and Y is defined by

$$a_n(T) := \inf_{\substack{S\colon X\to Y \\ \operatorname{rank}(S)<n}} \|T-S\|, \tag{4.3}$$

where $\operatorname{rank}(S) := \dim(S(X))$ denotes the dimension of the image of S.

Lemma 4.6. *The approximation number defines an s-number.*

Proof. To demonstrate that the properties i)-vi) of s-numbers easily follow from Lemma 4.4, we only give a proof of property iii). For this purpose let $T_1, T_2\colon X \to Y$ be two bounded linear operators and let $n \in \mathbb{N}$. It follows

$$\begin{aligned} a_n(T_1+T_2) &= \inf_{\substack{S\colon X\to Y \\ \operatorname{rank}(S)<n}} \|T_1+T_2-S\| \leq \inf_{\substack{S\colon X\to Y \\ \operatorname{rank}(S)<n}} C_Y\big(\|T_1-S\|+\|T_2\|\big) \\ &= C_Y\big(a_n(T_1)+\|T_2\|\big), \end{aligned}$$

which we wanted to show. □

Corollary 4.7. *Let s be any s-number and let $T\colon X \to Y$ be a bounded operator. For any $n \in \mathbb{N}$ it then holds*

$$s_n(T) \leq C_Y a_n(T),$$

where C_Y denotes the quasi-triangle constant of Y.

Proof. Let s be any s-function and let $T\colon X \to Y$ be a bounded linear operator. Further, for $n \in \mathbb{N}$ let $S\colon X \to Y$ be any linear operator with $\operatorname{rank}(S) < n$. Using property iii) from the definition of s-numbers, it follows

$$s_n(T) = s_n(T-S+S) \leq C_Y\big(s_n(S)+\|T-S\|\big) = C_Y\|T-S\|$$

and taking the infimum over S yields the claim. □

4.2.2 Entropy Numbers

Definition 4.8. The n-th *entropy number* $e_n(T)$ of a bounded linear operator $T\colon X \to Y$ between the two quasi-Banach spaces X and Y is defined by

$$e_n(T) := \inf\left\{\varepsilon > 0 \,\middle|\, T(B_X) \subset \bigcup_{j=1}^{2^{n-1}} (y_j + \varepsilon B_Y) \text{ for some } y_1, \dots, y_{2^{n-1}} \in Y\right\}.$$

In the definition of the entropy numbers $e_n(T)$ we can either work with the closed or the open unit ball on X. While usually the closed unit ball is used, we will work with the open unit ball for technical reasons.

If T has finite rank less than n, we still need some balls to cover the image of T. Hence, in that case we get

$$0 = C_Y a_n(T) < e_n(T),$$

and according to Corollary 4.7 the entropy numbers do not define an s-scale. However, the next lemma collects basic properties of entropy numbers in the spirit of s-number theory.

Lemma 4.9. *Let $T\colon X \to Y$ be a bounded linear operator between the two quasi-Banach spaces X and Y. Then:*

i) It holds $e_1(T) \le \|T\|$ with equality if Y is a Banach space.

ii) The entropy numbers are monotonically decreasing, i.e., for $n \in \mathbb{N}$ it holds

$$e_n(T) \ge e_{n+1}(T).$$

iii) For every $n \in \mathbb{N}$ and $T' \in \mathcal{L}(X,Y)$ it holds

$$e_n(T+T') \le C_Y\big(e_n(T) + \|T'\|\big).$$

iv) It holds

$$e_n(LTR) \le \|L\| \cdot e_n(T) \cdot \|R\|$$

for every $n \in \mathbb{N}$ and bounded operators $R\colon X_0 \to X$, $L\colon Y \to Y_0$.

Proof. i) First we observe

$$\begin{aligned} e_1(T) &= \inf\{\varepsilon > 0 \mid T(B_X) \subset (y + \varepsilon B_Y) \text{ for some } y \in Y\} \\ &\le \inf\{\varepsilon > 0 \mid T(B_X) \subset \varepsilon B_Y\} = \|T\|, \end{aligned}$$

so it remains to show the reverse inequality if Y is a Banach space. For this purpose let $y \in Y$ and $\varepsilon > 0$ be such that $T(B_X) \subset (y + \varepsilon B_Y)$. For every $x \in B_X$ we then get

$$\|Tx\|_Y = \frac{1}{2}\|Tx - y + y + Tx\|_Y \le \frac{1}{2}\big(\|Tx - y\|_Y + \|Tx + y\|_Y\big) < \varepsilon,$$

hence, $T(B_X) \subset \varepsilon B_Y$ and the claim follows from

$$\{\varepsilon > 0 \mid T(B_X) \subset (y + \varepsilon B_Y) \text{ for some } y \in Y\} \subset \{\varepsilon > 0 \mid T(B_X) \subset \varepsilon B_Y\}.$$

ii) This statement directly follows from the definition, since

$$\left\{\varepsilon > 0 \,\middle|\, T(B_X) \subset \bigcup_{j=1}^{2^{n-1}} (y_j + \varepsilon B_Y) \text{ for some } y_1, \dots, y_{2^{n-1}} \in Y\right\}$$
$$\subset \left\{\varepsilon > 0 \,\middle|\, T(B_X) \subset \bigcup_{j=1}^{2^{n}} (y_j + \varepsilon B_Y) \text{ for some } y_1, \dots, y_{2^{n}} \in Y\right\}.$$

iii) For any $\varepsilon > 0$ let $\eta = e_n(T) + \varepsilon$. Due to the definition of entropy numbers, there exists some $y_1, \dots, y_{2^{n-1}} \in Y$ such that for any $x \in B_X$ there is some $1 \le j \le 2^{n-1}$ with $\|Tx - y_j\|_Y < \eta$. It follows

$$\|(T + T')x - y_j\|_Y \le C_Y\big(\|Tx - y_j\|_Y + \|T'x\|_Y\big) \le C_Y\big(\eta + \|T'\|\big).$$

Hence,

$$(T + T')(B_X) \subset \bigcup_{j=1}^{2^{n-1}} y_j + C_Y\big(\eta + \|T'\|\big) B_Y,$$

which implies $e_n(T + T') \le C_Y\big(\eta + \|T'\|\big)$. Letting $\eta \to e_n(T)$, i.e., $\varepsilon \to 0$, yields the claim.

iv) For $R\colon X_0 \to X$ it holds $R(B_{X_0}) \subset \|R\| B_X$, which gives

$$T\big(R(B_{X_0})\big) \subset \|R\| \cdot T(B_X) \quad \Rightarrow \quad e_n(TR) \le \|R\| \cdot e_n(T).$$

To show the second claim $e_n(LT) \le \|L\| \cdot e_n(T)$, we set $\eta = e_n(T) + \varepsilon$ for some $\varepsilon > 0$. Then there exists some $y_1, \dots, y_{2^{n-1}} \in Y$, such that for every $x \in B_X$ there is some $1 \le j \le 2^{n-1}$ with $\|Tx - y_j\|_Y \le \eta$. It follows

$$\|LTx - Ly_j\|_{Y_0} \le \|L\| \cdot \|Tx - y_j\|_Y \le \eta \|L\|,$$

hence

$$LT(B_X) \subset \bigcup_{j=1}^{2^{n-1}} Ly_j + \eta \|L\| B_{Y_0},$$

which implies $e_n(LT) \le \eta\|L\| = \big(\varepsilon + e_n(T)\big) \cdot \|L\|$. Letting $\varepsilon \to 0$ yields the claim.

□

Remark 4.10. By similar arguments, property iii) of the previous Lemma 4.9 can be generalized to

$$e_{n+n'-1}(T + T') \le C_Y\big(e_{n-1}(T) + e_{n'}(T')\big)$$

for $n, n' \in \mathbb{N}$ and $T, T' \in \mathcal{L}(X, Y)$, cf. [47, Theorem 14].

Closely related to entropy numbers are the so-called ε-nets, which we define as follows:

Definition 4.11. A set $\mathcal{N} \subset X$ is called an *ε-net* of $K \subset X$, if for all $x \in K$ there exists a $z \in \mathcal{N}$ such that

$$\|x - z\|_X < \varepsilon.$$

Although one usually only requires $\|x - z\|_X \leq \varepsilon$ in the definition of ε-nets, we will work with the strict inequality to avoid technical issues later on. The next lemma is an analogue of [90, Proposition 12.1.13] for entropy numbers of identity mappings between quasi-Banach spaces, cf. [57, Lemma 2.1].

Lemma 4.12 ([57])**.** *Let X be a real d-dimensional p-Banach space. Then it holds*

$$e_n(\mathrm{id}\colon X \to X) \leq 4^{1/p} 2^{-\frac{n-1}{d}}$$

for all $n \in \mathbb{N}$.

Proof. First of all let us note that $e_n(T) \leq 1$ holds for every $n \in \mathbb{N}$. If we choose $n \in \mathbb{N}$ such that $(n-1)/d \leq 2/p$, it follows

$$e_n(\mathrm{id}\colon X \to X) \leq 1 = 4^{1/p} 2^{-2/p} \leq 4^{1/p} 2^{-\frac{n-1}{d}}.$$

Hence, it only remains to consider the case $(n-1) > 2d/p$. For $\varepsilon > 0$ we choose a maximal subset $\{x_1, \ldots, x_N\}$ of B_X such that

$$\|x_i - x_j\|_X > \epsilon$$

holds for all $i \neq j$. It follows:

i) $B_X \subset \bigcup_{i=1}^N (x_i + \varepsilon B_X)$, so the x_i form an ε-net of B_X.

ii) The sets $x_i + \frac{\varepsilon}{2^{1/p}} B_X$ are mutually disjoint.

Using the p-triangle inequality, for any $y \in B_X$ and $i = 1, \ldots, N$ we further get

$$\left\| x_i + \frac{\varepsilon}{2^{1/p}} \cdot y \right\|_X \leq \left(\|x_i\|_X^p + \frac{\varepsilon^p}{2} \cdot \|y\|_X^p \right)^{1/p} \leq \left(1 + \frac{\varepsilon^p}{2} \right)^{1/p},$$

which implies

$$\bigcup_{i=1}^N \left(x_i + \frac{\varepsilon}{2^{1/p}} B_X \right) \subset \left(1 + \frac{\varepsilon^p}{2} \right)^{1/p} B_X.$$

Comparing the volumes of both sets with respect to any translation invariant measure on X yields

$$\begin{aligned} N \left(\frac{\varepsilon}{2^{1/p}} \right)^d \mathrm{vol}(B_X) &= \mathrm{vol} \left(\bigcup_{i=1}^N \left(x_i + \frac{\varepsilon}{2^{1/p}} B_X \right) \right) \leq \mathrm{vol} \left(\left(1 + \frac{\varepsilon^p}{2} \right)^{1/p} B_X \right) \\ &= \left(1 + \frac{\varepsilon^p}{2} \right)^{d/p} \mathrm{vol}(B_X), \end{aligned}$$

where we used the σ-additivity of the volume for the first equality. It follows

$$N \leq \left(\frac{2^{1/p}}{\varepsilon}\right)^d \left(1+\frac{\varepsilon^p}{2}\right)^{d/p} = \left(1+\frac{2}{\varepsilon^p}\right)^{d/p}.$$

Now we choose $\varepsilon > 0$ to be

$$\varepsilon = \left(\frac{2}{2^{\frac{p(n-1)}{d}}-1}\right)^{1/p} < 1, \quad \text{such that} \quad N \leq \left(1+\frac{2}{\varepsilon^p}\right)^{d/p} = 2^{n-1},$$

where we used $p(n-1)/d > 2$ to ensure $\varepsilon < 1$. Hence, we found a covering of B_X using less then 2^{n-1} balls of radius $\varepsilon > 0$, and we end up with

$$e_n(\mathrm{id}\colon X \to X) \leq \varepsilon = \left(\frac{2}{2^{\frac{p(n-1)}{d}}-1}\right)^{1/p} \leq 4^{1/p} 2^{-\frac{n-1}{d}}.$$

□

Remark 4.13. With similar arguments as above, we can show that a bounded operator $T\colon X \to Y$ is of rank r if and only if

$$c \cdot 2^{\frac{-(n-1)}{r}} \leq e_n(T) \leq C \cdot \|T\| \cdot 2^{\frac{-(n-1)}{r}}$$

for some universal constants $c, C > 0$ and all $n \in \mathbb{N}$, cf. [23, (1.3.36)] or [47, Lemma 23].

4.2.3 Gelfand Numbers

Definition 4.14. The n-th *Gelfand number* $c_n(T)$ of a bounded linear operator $T \in \mathcal{L}(X,Y)$ between the two quasi-Banach spaces X and Y is defined by

$$c_n(T) := \inf_{\substack{M \subset X \\ \operatorname{codim} M < n}} \sup_{\substack{x \in M \\ \|x\|_X \leq 1}} \|Tx\|_Y,$$

where the infimum is taken over all linear subspaces M of X with $\operatorname{codim}(M) := \dim(X/M) < n$.

Using the quasi-triangle inequality and basic properties of the operator norm collected in Lemma 4.4, one easily observes that the Gelfand numbers indeed define an s-function.

Lemma 4.15. *The Gelfand numbers define an s-function.*

The next lemma collects two further properties of Gelfand numbers, which we will use later on.

Lemma 4.16. *Let $T\colon X \to Y$ be a bounded linear operator between the two quasi-Banach spaces X and Y. Then:*

i) If X is d-dimensional, we get $c_n(T) = 0$ for every $n > d$.

ii) If X and Y are d-dimensional and T is bijective, it holds $c_d(T) = \|T^{-1}\|^{-1}$.

Proof. i) Choosing $M = \{0\} \subset X$, we get $\operatorname{codim} M = d$. For any $n > d$ it follows

$$0 \leq c_n(T) \leq \sup_{x \in \{0\}} \|Tx\|_Y = 0.$$

ii) The claim follows from

$$c_d(T) = \inf_{\substack{M \subset X \\ \operatorname{codim} M < d}} \sup_{\substack{x \in M \\ \|x\|_X \leq 1}} \|Tx\|_Y = \inf_{x \in X} \frac{\|Tx\|_Y}{\|x\|_X} = \inf_{y \in Y} \frac{\|y\|_Y}{\|T^{-1}y\|_X} = \|T^{-1}\|^{-1}.$$

□

4.3 Carl's Inequality

For the sake of completeness, we define the n-th *Kolmogorov number* $d^n(T)$ of T by

$$d^n(T) := \inf_{\substack{L \subset Y \\ \dim L < n}} \sup_{\substack{x \in X \\ \|x\|_X \leq 1}} \inf_{z \in L} \|Tx - z\|_Y, \tag{4.4}$$

where the infimum is taken over all linear subspaces L of Y with dimension less than n. To avoid confusion, we denoted the Kolmogorov numbers with upper index n, since later on we will also discuss the so-called Gelfand n-width, which we will denote by d_n.

If $T\colon X \to Y$ is a bounded linear operator between the two (quasi-)Banach spaces X and Y, Carl's inequality [21] states that for any $\alpha > 0$ there exists a constant $\gamma_\alpha > 0$, such that for every $n \in \mathbb{N}$ it holds

$$\sup_{k \in [n]} k^\alpha e_k(T) \leq \gamma_\alpha \sup_{k \in [n]} k^\alpha s_k(T).$$

Here $s_k(T)$ either stands for the k-th Gelfand-, approximation-, or Kolmogorov number. Although the original proof of Carl heavily relies on the Hahn-Banach Theorem, which fails for quasi-Banach spaces [63], it easily extends to the case of quasi-Banach spaces and approximation- or Kolmogorov numbers, cf. [7, 47] and also [39, Section 1.3.3]. Unfortunately, the proof does not extend to the case of Gelfand numbers, which we will explain in the next section. Afterwards, we will fill the remaining gap by showing Carl's inequality for quasi-Banach spaces and Gelfand numbers using a new approach.

4.3.1 Carl's Inequality for Banach Spaces

In this section we assume X and Y to be Banach spaces. We will describe the original proof of Carl and point out that it heavily relies on the Hahn-Banach Theorem, which is used for both spaces, X and Y.

In the first step, Carl showed the inequality only for the case of approximation numbers as follows:

Theorem 4.17 ([21, 23])**.** *Let $T\colon X \to Y$ be a bounded linear operator between the Banach spaces X, Y. Then for any $\alpha > 0$ there exists a constant $\gamma_\alpha > 0$ such that for any $n \in \mathbb{N}$, it holds*

$$\sup_{k \in [n]} k^\alpha e_k(T) \leq \gamma_\alpha \sup_{k \in [n]} k^\alpha a_k(T).$$

Proof. By monotonicity, it is enough to prove the statement only for the case $n = 2^N$ and $N \in \mathbb{N}$, which we explain in iv) in the proof of Theorem 4.22. For $j = 1, \dots N$ let $S_j \in \mathcal{L}(X, Y)$ be such that

$$\operatorname{rank}(S_j) < 2^j \quad \text{and} \quad \|T - S_j\| \leq 2a_{2^j}(T),$$

where we set $S_0 = 0$. Then it follows

$$\operatorname{rank}(S_j - S_{j-1}) < 2^j + 2^{j-1} < 2 \cdot 2^j = 2^{j+1}$$

and we can decompose T as

$$T = T - S_N + \sum_{j=1}^{N} (S_j - S_{j-1}).$$

For natural numbers $n_1, \dots, n_N$, which we will determine later, we deduce

$$\begin{aligned} e_{n_1+\dots+n_N-(N-1)}(T) &= e_{n_1+\dots+n_N-(N-1)} \left(T - S_N + \sum_{j=1}^{N} (S_j - S_{j-1}) \right) \\ &\leq \|T - S_N\| + \sum_{j=1}^{N} e_{n_j}(S_j - S_{j-1}), \end{aligned} \tag{4.5}$$

where we applied Remark 4.10 for the last inequality. Since $\operatorname{rank}(S_j - S_{j-1}) < 2^{j+1}$, we can apply Remark 4.13 to get

$$\begin{aligned} e_{n_j}(S_j - S_{j-1}) &\leq C \cdot 2^{\frac{-(n_j-1)}{2^{j+1}}} \cdot \|S_j - S_{j-1}\| \\ &\leq C \cdot 2^{\frac{-(n_j-1)}{2^{j+1}}} \cdot (\|S_j - T\| + \|T - S_{j-1}\|) \\ &\leq C \cdot 2^{\frac{-(n_j-1)}{2^{j+1}}} \cdot 2\left(a_{2^j}(T) + a_{2^{j-1}}(T)\right) \leq 4C \cdot 2^{\frac{-(n_j-1)}{2^{j+1}}} \cdot a_{2^{j-1}}(T) \end{aligned} \tag{4.6}$$

for some constant $C > 0$. Combining the previous two estimates (4.5), (4.6) yields

$$e_{n_1+\dots+n_N-(N-1)}(T) \leq 2a_{2^N}(T) + 4C \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}}} a_{2^{j-1}}(T). \tag{4.7}$$

Next we bound both summands of the right hand side of (4.7). For the second summand we obtain

$$\begin{aligned} \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}}} a_{2^{j-1}}(T) &= \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}} - \alpha(j-1)} \cdot 2^{\alpha(j-1)} a_{2^{j-1}}(T) \\ &\leq \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}} - \alpha(j-1)} \cdot \sup_{1 \leq k \leq N} \left(2^{(k-1)}\right)^{\alpha} a_{2^{k-1}}(T) \\ &\leq \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}} - \alpha(j-1)} \cdot \sup_{1 \leq k \leq 2^N} k^{\alpha} a_k(T) \end{aligned} \tag{4.8}$$

and for the first summand we recognize

$$a_{2^N}(T) \leq 2^{-\alpha N} \sup_{1\leq k\leq 2^N} k^\alpha a_k(T). \tag{4.9}$$

Combining (4.7), (4.8) and (4.9), we conclude

$$e_{n_1+\ldots+n_N-(N-1)}(T) \leq \left(4C \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}}-\alpha(j-1)} + 2^{-\alpha N+1} \right) \cdot \sup_{1\leq k\leq 2^N} k^\alpha a_k(T).$$

For $K \in \mathbb{N}$ with $1+\alpha \leq K \leq 2+\alpha$ we now put

$$n_j = 1 + K(N-j)2^{j+1}.$$

By induction, one easily comprehends

$$\begin{aligned} n_1 + \ldots + n_N - (N-1) &= 1 + K\sum_{j=1}^{N}(N-j)2^{j+1} = 1 + 4K(2^N - N - 1) \\ &\leq 4K2^N \end{aligned}$$

and, since $1+\alpha \leq K \leq 2+\alpha$, it follows

$$\begin{aligned} \sum_{j=1}^{N} 2^{\frac{-(n_j-1)}{2^{j+1}}-\alpha(j-1)} &= \sum_{j=1}^{N} 2^{-K(N-j)-\alpha(j-1)} = 2^{-KN+\alpha}\sum_{j=1}^{N} 2^{(K-\alpha)j} \\ &= 2^{-KN+\alpha}\cdot 2^{K-\alpha}\frac{2^{N(K-\alpha)}-1}{2^{K-\alpha}-1} \leq \frac{2^K 2^{-\alpha N}}{2^{K-\alpha}-1} \leq 42^{-\alpha N}2^\alpha. \end{aligned}$$

Using the monotonicity of the entropy numbers, we finally end up with

$$e_{n_1+\ldots+n_N-(N-1)}(T) \leq e_{4K2^N}(T) \leq 16C2^{-\alpha N}2^\alpha \sup_{1\leq k\leq 2^N} k^\alpha a_k(T).$$

The claim now follows using further monotonicity arguments. □

Remark 4.18. Using the Theorem of Aoki-Rolewicz 4.1, the proof easily extends to the case of quasi-Banach spaces, cf. [47, Proposition 24].

After showing the inequality for approximation numbers, Carl used the so-called injectivity and surjectivity of the entropy numbers [90, Proposition 12.1.8], to conclude the validity also for Gelfand and Kolmogorov numbers. Although these statements extend to the case of Kolmogorov numbers [7, Lemma 1], they completely fail for the case of Gelfand numbers, which we will discuss now.

Let $Y^* := \mathcal{L}(Y, \mathbb{R})$ denote the dual space of some Banach space Y. Endowed with the operator norm makes Y^* again a Banach space, whose closed unit ball $\overline{B_{Y^*}} = \{T \in Y^* \mid \|T\| \leq 1\}$ will be denoted in the following simply by S. Then we define the operator

$$\iota\colon Y \to \ell_\infty(S), \quad y \mapsto (y'(y))_{y'\in S}, \tag{4.10}$$

where the intuitive definition

$$\|s\|_{\ell_\infty(S)} := \sup_{y'\in S} |s_{y'}|, \qquad s = (s_{y'})_{y'\in S} \in \ell_\infty(S)$$

makes $\ell_\infty(S)$ a Banach space. Using the Theorem of Hahn-Banach, for any $y \in Y$ we obtain

$$\|\iota(y)\|_{\ell_\infty(S)} = \sup_{y' \in S} |y'(y)| = \|y\|_Y, \tag{4.11}$$

so ι defines an isometric embedding. Further, let us observe that the space $\ell_\infty(S)$ has the so-called *extension property*. That is, for any subspace M of a Banach space X let $\tilde{U}\colon M \to \ell_\infty(S)$ be a bounded and linear operator. Then there exists an extension $U\colon X \to \ell_\infty(S)$ of $\tilde{U}$ such that

$$\|\tilde{U}\| = \|U\| \quad \text{and} \quad \tilde{U}x = Ux \text{ for every } x \in M.$$

Indeed, since $\tilde{U}$ is bounded, for any $y' \in S$ the mapping

$$\tilde{U}_{y'} : M \to \mathbb{R}, \quad x \mapsto y'(\tilde{U}x)$$

is an element of the dual space of M. By the Theorem of Hahn-Banach there exists an extension $U_{y'}\colon X \to \mathbb{R}$ of $\tilde{U}_{y'}$, i.e., it holds $\|\tilde{U}_{y'}\| = \|U_{y'}\|$ and for every $x \in M$ we get $\tilde{U}_{y'}(x) = U_{y'}(x)$. Finally, the operator

$$U\colon X \to \ell_\infty(S), \quad x \mapsto (U_{y'}(x))_{y' \in S}$$

yields the desired properties of an extension of $\tilde{U}$. Equipped with the isometric embedding ι and the extension property, we now deduce the following lemma.

Lemma 4.19. *Let X, Y be two Banach spaces and let $T \in \mathcal{L}(X,Y)$. For every $k \in \mathbb{N}$ it holds*

$$\begin{aligned} &i) \qquad e_k(T) \le 2e_k(\iota \circ T), \\ &ii) \qquad c_k(T) = a_k(\iota \circ T), \end{aligned}$$

where $\iota\colon Y \to \ell_\infty(S)$ denotes the isometric embedding from (4.10).

Proof. i) Let $(\iota \circ T)(B_X)$ be covered by $N = 2^{k-1}$ balls of radius ε and center points $s_1, \dots, s_N \in \ell_\infty(S)$, i.e.,

$$(\iota \circ T)(B_X) \subset \bigcup_{j=1}^{N} B(\varepsilon, s_j).$$

Since we do not know if the points s_j are in the image of $\iota \circ T$, for every $j = 1, \dots, N$ we choose some $x_j \in X$ such that $(\iota \circ T)(x_j) \in B(\varepsilon, s_j)$. Using the triangle inequality, for every $s \in B(\varepsilon, s_j)$ we obtain

$$\|s - (\iota \circ T)(x_j)\|_{\ell_\infty(S)} \le \|s - s_j\|_{\ell_\infty(S)} + \|s_j - (\iota \circ T)(x_j)\|_{\ell_\infty(S)} < 2\varepsilon,$$

hence $B(\varepsilon, s_j) \subset B(2\varepsilon, (\iota \circ T)(x_j))$. Since ι is isometric, we found a covering of $T(B_X)$ with center points $T(x_j)$ and radius 2ε, which shows the claim.

ii) We prove the claim in two steps:

"$\leq$:" By Corollary 4.7, the approximation numbers are larger than the Gelfand numbers. It follows

$$\begin{aligned} c_k(T) &= \inf_{\operatorname{codim} M<k} \sup_{\substack{x\in M\\ \|x\|\leq 1}} \|Tx\|_Y = \inf_{\operatorname{codim} M<k} \sup_{\substack{x\in M\\ \|x\|\leq 1}} \|(\iota\circ T)x\|_{\ell_\infty(S)} \\ &= c_k(\iota\circ T) \leq a_k(\iota\circ T). \end{aligned}$$

"$\geq$:" For $n\in\mathbb{N}$ let M be any subspace of X with $\operatorname{codim} M < n$. Using the extension property of $\ell_\infty(S)$, we can extend $\tilde{U} := \iota\circ T|_M$ to an operator $U\colon X\to\ell_\infty(S)$ such that

$$\|U\| = \|\tilde{U}\| = \|\iota\circ T|_M\| = \|T|_M\|.$$

The operator $L := \tilde{U} - U$ has rank strictly smaller than n and we conclude

$$a_n(\iota\circ T) \leq \|\iota\circ T - L\| = \|U\| = \|T|_M\|.$$

The claim now follows by taking the infimum over all subspaces M.

□

Using the previous Lemma 4.19, Carl's inequality with respect to Gelfand numbers now simply follows from

$$\sup_{1\leq k\leq n} k^\alpha e_k(T) \leq 2 \sup_{1\leq k\leq n} k^\alpha e_k(\iota\circ T) \leq 2\gamma_\alpha \sup_{1\leq k\leq n} k^\alpha a_k(\iota\circ T) \leq 2\gamma_\alpha \sup_{1\leq k\leq n} k^\alpha c_k(T).$$

Let us again highlight the use of the Theorem of Hahn-Banach: While showing that ι defines an isometric embedding, we used the Theorem of Hahn-Banach for the Banach space Y. And to prove the second claim of Lemma 4.19, we used the extension property, which was shown by using the Theorem of Hahn-Banach for the Banach space X.

4.3.2 Carl's Inequality for Quasi-Banach Spaces

In this section we will fill the remaining gap by showing Carl's inequality for Gelfand numbers on quasi-Banach spaces with an alternate proof, which is based on the following lemma, cf. [57, Lemma 3.2].

Lemma 4.20 ([57])**.** *Let $T\colon X\to Y$ be a bounded operator between a p-Banach space X and a q-Banach space Y. Further, let*

a) $(M_n)_{n\in\mathbb{N}}$ be a sequence of finite codimensional subspaces of X,

b) $\delta_n = \prod_{i=1}^n \varepsilon_i$, where $(\varepsilon_n)_{n\in\mathbb{N}}$ is a sequence of positive numbers with $\varepsilon_n \leq 1$ for each $n\in\mathbb{N}$ and where we set $\delta_0 = 1$,

c) for every $n\in\mathbb{N}$ let $\mathcal{M}_n \subset B_{X/M_n}$ be an ε_n-net of the unit ball of X/M_n and

d) for every $n\in\mathbb{N}$ let $\mathcal{N}_n \subset 2^{1/p} B_X$ be a lifting of $\mathcal{M}_n$, i.e., $\mathcal{N}_n$ contains exactly one element of each equivalence class $[x]\in\mathcal{M}_n$.

Then, for every $x \in B_X$, there exist sequences $(x_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$, such that for every $n \in \mathbb{N}$ it holds $x_n \in \mathcal{N}_n$, $z_n \in M_n$ and

i) $\|z_n\|_X < 4^{1/p}$,

ii) $\|x - \sum_{k=1}^{n} \delta_{k-1}(x_k + z_k)\|_X < \delta_n$ *and*

iii) $\|Tx - \sum_{k=1}^{n} \delta_{k-1} Tx_k\|_Y^q \le (\|T\|\delta_n)^q + 4^{q/p} \sum_{k=1}^{n} \delta_{k-1}^q \|T|_{M_k}\|^q$,

where $\|T|_{M_k}\|$ denotes the operator norm of T restricted to the subspace M_k.

Proof. First we will prove i) and ii) by induction and from that we will deduce iii).

Let $x \in B_X$ and $n = 1$. Since $\mathcal{N}_1 \subset X$ is a lifting of the ε_1-net $\mathcal{M}_1$, there exists $x_1 \in \mathcal{N}_1$ and $z_1 \in M_1$ such that

$$\|x - \delta_0(x_1 + z_1)\|_X = \|x - (x_1 + z_1)\|_X < \varepsilon_1 = \delta_1.$$

Using the p-triangle inequality on X we further get

$$\|z_1\|_X^p \le \|x - (x_1 + z_1)\|_X^p + \|x\|_X^p + \|x_1\|_X^p < \varepsilon_1^p + 1 + 2 \le 4.$$

Next we assume that the conditions i) and ii) hold for some $n \in \mathbb{N}$ and we verify them for $n + 1$. First we observe

$$\left\| \frac{1}{\delta_n} \left(x - \sum_{k=1}^{n} \delta_{k-1}(x_k + z_k) \right) \right\|_X < 1,$$

and since $\mathcal{N}_{n+1}$ is a lifting of the ε_{n+1}-net $\mathcal{M}_{n+1}$, there exists some $x_{n+1} \in \mathcal{N}_{n+1}$ and some $z_{n+1} \in M_{n+1}$ such that

$$\left\| \frac{1}{\delta_n} \left(x - \sum_{k=1}^{n} \delta_{k-1}(x_k + z_k) \right) - (x_{n+1} + z_{n+1}) \right\|_X < \varepsilon_{n+1},$$

which proves ii). The bound for z_{n+1} again easily follows from the p-triangle inequality by

$$\begin{aligned} &\|z_{n+1}\|_X^p \\ &\le \frac{1}{\delta_n^p} \left(\left\| x - \sum_{k=1}^{n+1} \delta_{k-1}(x_k + z_k) \right\|_X^p + \left\| x - \sum_{k=1}^{n} \delta_{k-1}(x_k + z_k) \right\|_X^p + \|\delta_n x_{n+1}\|_X^p \right) \\ &\le \frac{1}{\delta_n^p} \left(\delta_{n+1}^p + \delta_n^p + 2\delta_n^p \right) = \varepsilon_{n+1}^p + 1 + 2 \le 4. \end{aligned}$$

It remains to deduce iii) from i) and ii). Here we use the q-triangle inequality on Y to conclude

$$\begin{aligned} \left\| Tx - \sum_{k=1}^{n} \delta_{k-1} Tx_k \right\|_Y^q &\le \left\| Tx - \sum_{k=1}^{n} \delta_{k-1} T(x_k + z_k) \right\|_Y^q + \sum_{k=1}^{n} \delta_{k-1}^q \|Tz_k\|_Y^q \\ &\le \|T\|^q \left\| x - \sum_{k=1}^{n} \delta_{k-1}(x_k + z_k) \right\|_X^q + \sum_{k=1}^{n} \delta_{k-1}^q \|T|_{M_k}\|^q \|z_k\|_Y^q \\ &\le (\|T\|\delta_n)^q + \sum_{k=1}^{n} \delta_{k-1}^q 4^{q/p} \|T|_{M_k}\|^q. \end{aligned}$$

□

Taking the infimum over all linear subspaces M_k and using the bound for the entropy numbers from Lemma 4.12, we can deduce the following estimate for the entropy numbers by terms of the Gelfand numbers, cf. [57, Theorem 3.3].

Lemma 4.21 ([57])**.** *Let $T\colon X \to Y$ be a bounded operator between the p-Banach space X and the q-Banach space Y with $0 < p, q \leq 1$ and let $(k_j)_{j\in\mathbb{N}}$ and $(m_j)_{j\in\mathbb{N}}$ be sequences in $\mathbb{N}$. Then, for any $n \in \mathbb{N}$, it holds*

$$e_{k_1+\ldots+k_n+1-n}(T)^q \leq 4^{nq/p} \cdot 2^{-q\sum_{j=1}^{n} \frac{k_j-1}{m_j}} \|T\|^q + 4^{q/p} \sum_{k=1}^{n} 4^{q(k-1)/p} \cdot 2^{-q\sum_{j=1}^{k-1} \frac{k_j-1}{m_j}} c_{m_k+1}(T)^q.$$

Proof. For each $j \in \mathbb{N}$, let $M_j \subset X$ be a linear subspace of X with $\operatorname{codim} M_j = \dim X/M_j = m_j$. Furthermore, for $j, k \in \mathbb{N}$ we set

$$\varepsilon_j = 4^{1/p} \cdot 2^{\frac{-(k_j-1)}{m_j}}, \qquad \delta_k = \prod_{j=1}^{k} \varepsilon_j = 4^{k/p} \cdot 2^{-\sum_{j=1}^{k} \frac{k_j-1}{m_j}}.$$

Using the upper bound from Lemma 4.12 we get

$$e_{k_j}(\mathrm{id}\colon X/M_j \to X/M_j) \leq 4^{1/p} \cdot 2^{\frac{-(k_j-1)}{m_j}} = \varepsilon_j.$$

Lemma 4.20 now provides a covering of $T(B_X)$ using $2^{k_1+\ldots k_n-n}$ center points, and following iii) from this lemma we obtain the estimate

$$\begin{aligned} &e_{k_1+\ldots+k_n-n+1}(T)^q \\ &\leq (\|T\|\delta_n)^q + 4^{q/p} \sum_{k=1}^{n} \delta_{k-1}^q \|T|_{M_k}\|^q \\ &= 4^{nq/p} \cdot 2^{-q\sum_{j=1}^{n} \frac{k_j-1}{m_j}} \|T\|^q + 4^{q/p} \sum_{k=1}^{n} 4^{q(k-1)/p} \cdot 2^{-q\sum_{j=1}^{k-1} \frac{k_j-1}{m_j}} \|T|_{M_k}\|^q. \end{aligned}$$

Taking the infimum over all linear subspaces M_k with $\operatorname{codim} M_k = m_k$ yields the claim. □

With the previous lemma we are ready to prove Carl's inequality, cf. [57, Theorem 1.1], which is the main result of this chapter.

Theorem 4.22 ([57])**.** *Let $T\colon X \to Y$ be a bounded linear operator between two quasi-Banach spaces X and Y. For $\alpha > 0$ there exists a constant γ_α such that, for any $n \in \mathbb{N}$, it holds*

$$\sup_{k\in[n]} k^\alpha e_k(T) \leq \gamma_\alpha \sup_{k\in[n]} k^\alpha c_k(T). \tag{4.12}$$

Proof. Before proving the theorem, let us make the following simplifications:

i) According to the Theorem of Aoki-Rolewicz 4.1, the quasi-norms on X and Y are equivalent to a p-, respectively to a q-norm with $0 < p, q \leq 1$. Since equivalence does not change the claim of Carl's inequality, we may assume without loss of generality that X and Y are already a p- and a q-Banach space, respectively.

ii) Since both sides of Carl's inequality are homogeneous, i.e., multiplying the operator T by some positive factor does not change the inequality, we can assume that $c_k(T) \leq k^{-\alpha}$ holds for all $k \in [n]$. In particular, it follows

$$c_1(T) = \|T\| \leq 1.$$

iii) Clearly, it is enough to show that

$$n^\alpha e_n(T) \leq \gamma_\alpha \sup_{k\in[n]} k^\alpha c_k(T) \tag{4.13}$$

holds for every $n \in \mathbb{N}$. Indeed, in that case it follows

$$\sup_{k\in[n]} k^\alpha e_k(T) \leq \sup_{k\in[n]} \gamma_\alpha \sup_{j\in[k]} j^\alpha c_j(T) = \sup_{k\in[n]} \gamma_\alpha k^\alpha c_k(T).$$

iv) By monotonicity of the Gelfand and entropy numbers, it is also enough to prove Carl's inequality for $n = C2^N$, where $C \in \mathbb{N}$ denotes a universal constant and $N \in \mathbb{N}$. Indeed, assume that we have already shown (4.13) for every $n = C2^N$. For $\tilde{n} \in \mathbb{N}$ let $N \in \mathbb{N}$ be such that $C2^{N-1} \leq \tilde{n} \leq C2^N$. Then, by monotonicity, it follows

$$\begin{aligned}\tilde{n}^\alpha e_{\tilde{n}}(T) &\leq (C2^N)^\alpha e_{C2^{N-1}}(T) \leq (2C)^\alpha \gamma_\alpha \sup_{k\in[C2^{N-1}]} k^\alpha c_k(T) \\ &\leq \tilde{\gamma}_\alpha \sup_{k\in[\tilde{n}]} k^\alpha c_k(T),\end{aligned}$$

where we set $\tilde{\gamma}_\alpha := (2C)^\alpha \gamma_\alpha$.

Using those simplifications, we now proceed with the proof of the claim. For any $\alpha > 0$ and $N \in \mathbb{N}$ we choose some $\beta > \alpha$ and we set $m_j = 2^{N-j}$, $j = 1, \dots, N$. In order to apply Lemma 4.21, for $j = 1, \dots, N$ we further set

$$k_j = \lceil 2^{N-j}(2/p + \beta) + 1 \rceil, \qquad \varepsilon_j = 4^{1/p} \cdot 2^{\frac{-(k_j-1)}{m_j}}.$$

Since $(k_j - 1)/m_j \geq 2/p + \beta$, it follows $\varepsilon_j \leq 2^{-\beta}$ and Lemma 4.21 yields

$$\begin{aligned}&e_{k_1+\ldots+k_N+1-N}(T)^q \\ &\leq 4^{Nq/p} \cdot 2^{-q\sum_{j=1}^{N} \frac{k_j-1}{m_j}} \|T\|^q + 4^{q/p} \sum_{k=1}^{N} 4^{q(k-1)/p} \cdot 2^{-q\sum_{j=1}^{k-1} \frac{k_j-1}{m_j}} c_{m_k+1}(T)^q \\ &\leq 4^{Nq/p} \cdot 2^{-2Nq/p - Nq\beta} + 4^{q/p} \sum_{k=1}^{N} 4^{q(k-1)/p} \cdot 2^{-2q(k-1)/p - q(k-1)\beta} (m_k+1)^{-q\alpha} \\ &= 2^{-Nq\beta} + 4^{q/p} \sum_{k=1}^{N} 2^{-q(k-1)\beta} (2^{N-k}+1)^{-q\alpha},\end{aligned} \tag{4.14}$$

where we used the assumption $c_k(T) \leq k^{-\alpha}$. We further estimate by

$$\begin{aligned}&\sum_{k=1}^{N} 2^{-q(k-1)\beta} (2^{N-k}+1)^{-q\alpha} \leq \sum_{k=1}^{N} 2^{-q(k-1)\beta} \cdot 2^{-(N-k)q\alpha} \\ &= 2^{q\beta} \cdot 2^{-Nq\alpha} \sum_{k=1}^{N} 2^{-qk\beta} \cdot 2^{kq\alpha} = 2^{q\beta} 2^{-Nq\alpha} \sum_{k=1}^{N} 2^{kq(\alpha-\beta)} \\ &\leq C_0 2^{q\beta} 2^{-Nq\alpha} 2^{Nq(\alpha-\beta)} \leq C_1 2^{-Nq\alpha},\end{aligned} \tag{4.15}$$

where the constants C_0 and C_1 only depend on α, β, p and q. Combining (4.14) and (4.15), we arrive at

$$e_{k_1+\ldots+k_N+1-N}(T)^q \leq 2^{-Nq\beta} + C_1 4^{q/p} 2^{-Nq\alpha} \leq C_2 2^{-Nq\alpha}.$$

Now let $C \in \mathbb{N}$ be such that $1 + \beta + 2/p \leq C$. Then it follows

$$\begin{aligned} 1 - N + \sum_{j=1}^{N} k_j &= 1 - N + \sum_{j=1}^{N} \lceil 2^{N-j}(2/p + \beta) + 1 \rceil \\ &\leq 1 - N + \sum_{j=1}^{N} (2^{N-j}(2/p + \beta) + 2) \\ &= N + 1 + (2/p + \beta)(2^N - 1) \leq C 2^N \end{aligned}$$

and, finally, we obtain

$$e_{C2^N}(T)^q \leq e_{k_1+\ldots+k_N+1-N}(T)^q \leq C_2 2^{-Nq\alpha} \leq C_3 (C2^N)^{-\alpha q}$$

for a constant C_3, which is independent of N. Setting $\gamma_\alpha = c_1(T)^\alpha \cdot C_3^{1/q}$ we get

$$(C2^N)^\alpha e_{C2^N}(T) \leq C_3^{1/q} \leq \gamma_\alpha \sup_{k \in [C2^N]} k^\alpha c_k(T),$$

as claimed. □

Using Carl's inequality, we can easily compare the Lorentz quasi-norms of $(e_k(T))_{k\in\mathbb{N}}$ and $(c_k(T))_{k\in\mathbb{N}}$, cf. [57, Theorem 3.4] and [23, Theorem 3.1.2]. For this purpose we will make use of *Hardy's inequality* (cf. [23, Lemma 1.5.3]), which states that

$$\sum_{k=1}^{n} k^{t/s-1} \Big(\frac{1}{k} \sum_{j=1}^{k} \sigma_j^q \Big)^{t/q} \leq \Big(1 + \frac{s}{s-q} \Big) \sum_{k=1}^{n} k^{t/s-1} \sigma_k^t \tag{4.16}$$

holds for a sequence of nonnegative, decreasing numbers $0 \leq \sigma_n \leq \ldots \leq \sigma_1$, $0 < s, t < \infty$ and $0 < q < \min\{s, t\}$, cf. [57, Theorem 3.4].

Theorem 4.23 ([57])**.** *Let $T\colon X \to Y$ be a bounded linear operator between the quasi-Banach spaces X and Y. Then, for every $0 < s \leq \infty$ and $0 < t < \infty$, there exists a constant $C_{s,t} > 0$, such that for every $n \in \mathbb{N}$, it holds*

$$\left(\sum_{k=1}^{n} k^{t/s-1} e_k(T)^t \right)^{1/t} \leq C_{s,t} \left(\sum_{k=1}^{n} k^{t/s-1} c_k(T)^t \right)^{1/t}.$$

Proof. For $0 < s < \infty$ and $0 < t < \infty$, let $\alpha > \max\{1/s, 1/t\}$. Using Carl's

inequality, we obtain

$$\begin{aligned}
&\sum_{k=1}^{n} k^{t/s-1} e_k(T)^t = \sum_{k=1}^{n} k^{t/s-1-t\alpha} (k^\alpha e_k(T))^t \\
&\le \sum_{k=1}^{n} k^{t/s-1-t\alpha} \Big(\sup_{l\in[k]} l^\alpha e_l(T) \Big)^t \le \gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1-t\alpha} \Big(\sup_{l\in[k]} l^\alpha c_l(T) \Big)^t \\
&\le \gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1-t\alpha} \Big(\sup_{l\in[k]} l c_l(T)^{1/\alpha} \Big)^{\alpha t} \le \gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1-t\alpha} \Big(\sup_{l\in[k]} \sum_{j=1}^{l} c_j(T)^{1/\alpha} \Big)^{\alpha t} \\
&= \gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1-t\alpha} \Big(\sum_{j=1}^{k} c_j(T)^{1/\alpha} \Big)^{\alpha t} = \gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1} \Big(\frac{1}{k} \sum_{j=1}^{k} c_j(T)^{1/\alpha} \Big)^{\alpha t},
\end{aligned}$$

where we used the monotonicity of Gelfand numbers for the last inequality. Next, with the choice $q = 1/\alpha$ we apply Hardy's inequality (4.16) to end up with

$$\gamma_\alpha^t \sum_{k=1}^{n} k^{t/s-1} \Big(\frac{1}{k} \sum_{j=1}^{k} c_j(T)^{1/\alpha} \Big)^{\alpha t} \le \gamma_\alpha^t \left(1 + \frac{s}{s - 1/\alpha} \right) \sum_{k=1}^{n} k^{t/s-1} c_k(T)^t,$$

which finishes the proof for the case $s < \infty$. The claim for $s = \infty$ now easily follows by taking the limit $s \to \infty$ and noting that $s/(s - 1/\alpha) \to 1$ for $s \to \infty$. □

4.4 Encoder-Decoder Performance

Recently, Gelfand and entropy numbers were used in the area of compressed sensing [17, 35] to determine the performance of optimal encoder-decoder pairs, which we will discuss in this section.

In its basic setting, compressed sensing studies pairs (A, Δ) of linear measurement maps

$$A \colon X \to \mathbb{R}^m,$$

which is called the *encoder*, and in general nonlinear recovery maps

$$\Delta \colon \mathbb{R}^m \to X$$

for some (quasi-)normed space X, which we will call the *decoder*. The aim is to find a particular pair (A, Δ), such that for each signal $x \in K$ of a known set $K \subset X$, e.g. the set of s-sparse signals in $\mathbb{R}^d$, the error of reconstruction

$$\|x - \Delta(Ax)\|_X$$

is small. For a fixed set $K \subset X$ the performance of a particular encoder/decoder pair (A, Δ) is given as its worst case error

$$\varepsilon(A, \Delta, K, X) := \sup_{x\in K} \|x - \Delta(Ax)\|_X.$$

Since we are interested in the best possible performance, we define the *compressive n-width* of K by taking the infimum over all pairs (A, Δ), i.e., we set

$$E_n(K, X) := \inf\{\varepsilon(A, \Delta, K, X) \mid A \colon X \to \mathbb{R}^n \text{ linear}, \ \Delta \colon \mathbb{R}^n \to X\}.$$

It turns out that the compressive n-width of K is closely related to the so-called Gelfand n-width of K, which we define as follows:

Definition 4.24. The *Gelfand n-width* $d_n(K, X)$ of a subset $K \subset X$ of a (quasi-)normed space X is defined by

$$d_n(K, X) := \inf_{\operatorname{codim} M \leq n} \sup_{x \in M \cap K} \|x\|_X,$$

where the infimum is taken over all linear subspaces M of X with $\operatorname{codim}(M) = \dim(X/M) \leq n$.

The Gelfand numbers can be seen as a generalization of the Gelfand width in the following way: Let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote two quasi-norms on the vector space X and let $\overline{B_1} = \{x \in X \mid \|x\|_1 \leq 1\}$ denote the closed unit ball in X with respect to $\|\cdot\|_1$. For the identity operator $\mathrm{id}\colon (X, \|\cdot\|_1) \to (X, \|\cdot\|_2)$ and any $n \geq 2$ we obtain

$$\begin{aligned} c_n(\mathrm{id}\colon X \to X) &= \inf_{\substack{M \subset X \\ \operatorname{codim} M < n}} \sup_{\substack{x \in M \\ \|x\|_1 \leq 1}} \|x\|_2 = \inf_{\substack{M \subset X \\ \operatorname{codim} M < n}} \sup_{x \in M \cap \overline{B_1}} \|x\|_2 \\ &= d_{n-1}(\overline{B_1}, \|\cdot\|_2). \end{aligned} \tag{4.17}$$

The following well known theorem shows the equivalence of compressive- and Gelfand n-width, cf. [45, Theorem 10.4].

Theorem 4.25. *Let $K \subset X$ be a subset of a (quasi-)normed space X. If K is symmetric, i.e., if $K = -K$, it holds*

$$d_n(K, X) \leq E_n(K, X).$$

If there further exists a constant $C > 0$ such that $K + K \subset CK$, then it holds

$$E_n(K, X) \leq C d_n(K, X).$$

The measurement map $A\colon X \to \mathbb{R}^n$ is called *adaptive*, if the i-th measurement $a_i(x)$ of $Ax = (a_1(x), \ldots, a_n(x))$ is allowed to depend on the previous results $a_1(x), \ldots, a_{i-1}(x)$. In the previous theorem we restricted to nonadaptive measurement maps A, although the same result also holds in the adaptive case. Hence, adaptivity does not improve the worst case performance, which justifies our restriction to nonadaptive encoders A.

If we choose $X = \mathbb{R}^d$ and $K = \Sigma_s^d$ to be the set of all s-sparse vectors in $\mathbb{R}^d$, it holds $K = -K$ and the previous Theorem 4.25 translates any lower bound on the Gelfand n-width into a lower bound for the compressive n-width. But since $K + K$ is not a subset of CK for any $C > 0$, we cannot deduce upper bounds in the same way.

4.5 Gelfand Numbers for ℓ_p-Balls

Our main motivation for Carl's inequality is the relationship between compressive n-width of a given set $K \subset \mathbb{R}^d$ and its Gelfand width. Hence, we are in particular interested in Gelfand numbers of the open unit balls $B_p^d \subset \mathbb{R}^d$ for some $0 < p \leq 1$, since they serve as a good model for compressible signals in $\mathbb{R}^d$, as we already highlighted in Remark 3.3.

We start with the following Theorem of Kuhn and Schütt [69, 102] on entropy numbers for the identity operator between ℓ_q^d and ℓ_p^d with $0 < p \leq q \leq \infty$. There we will use the notation $A \asymp B$, meaning that the two quantities A and B are equivalent, i.e., that there exists two constant $C_0, C_1 > 0$ such that

$$C_0 A \leq B \leq C_1 A.$$

Theorem 4.26. *[69, 102] For $0 < p \leq q \leq \infty$ and $m \in \mathbb{N}$ it holds*

$$e_n(\mathrm{id}\colon \ell_p^d \to \ell_q^d) \asymp \begin{cases} 1, & \text{if } 1 \leq n \leq \log(d), \\ \left(\frac{1+\log(d/n)}{n}\right)^{1/p-1/q}, & \text{if } \log(d) \leq n \leq d, \\ 2^{-n/d} d^{1/q-1/p}, & \text{if } d \leq n, \end{cases} \tag{4.18}$$

where the constants of equivalence do not depend on d and n.

By applying Carl's inequality to the identity operator $\mathrm{id}\colon \ell_p^d \to \ell_2^d$ and using the previous result of Kuhn and Schütt, Donoho obtained a lower bound for the Gelfand number $c_n(\mathrm{id}\colon \ell_p^d \to \ell_2^d)$ with $0 < p \leq 2$ and, consequently, also for the compressive n-width $E_n(B_p^d, \ell_2^d)$ [35]. Unfortunately, this proof contained a crucial flaw, namely that Carl's inequality was only proven for Banach spaces at that time and could not be used for ℓ_p^d with $p < 1$. This gap was solved in [44], where the authors used a completely different approach avoiding the use of Carl's inequality.

Since we provided an alternate proof for Carl's inequality working also for quasi-Banach spaces, we will now reproduce the lower bound for the Gelfand numbers, cf. [57, Theorem 4.1.]. The proof follows the original proof of Carl and Pisier with only minor modifications, cf. [22, Corollary 2.6.].

Theorem 4.27 ([22, 44, 57])**.** *For $d \in \mathbb{N}$, $1 \leq n \leq d$ and $0 < p < 2$ it holds*

$$C \min\left\{1, \frac{1+\log(d/n)}{n}\right\}^{\frac{1}{p}-\frac{1}{2}} \leq c_n(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \leq C' \min\left\{1, \frac{1+\log(d/n)}{n}\right\}^{\frac{1}{p}-\frac{1}{2}}, \tag{4.19}$$

where the constants $C, C' > 0$ do not depend on d or n.

Proof. The upper bound of this inequality was already provided in [109], so we will only prove the lower bound.

For brevity let us set $\alpha = 1/p - 1/2 > 0$. For any $n \in \mathbb{N}$ such that $\log(d) \leq n \leq d$ we use Carl's inequality 4.22 and the Theorem of Kuhn and Schütt 4.26 to get

$$\begin{aligned} C_0 n^\alpha (1+\log(d/n))^\alpha &\leq n^{2\alpha} e_n(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \leq \sup_{k\in[n]} k^{2\alpha} e_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \\ &\leq \gamma_{2\alpha} \sup_{k\in[n]} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \end{aligned} \tag{4.20}$$

for some constants $C_0, \gamma_{2\alpha} > 0$. For $\lambda > 1$, which we will fix later on, we split this supremum into two parts, namely into

$$\sup_{k\in[n]} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \leq \sup_{1\leq k\leq \lfloor \frac{n}{\lambda} \rfloor} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) + \sup_{\lfloor \frac{n}{\lambda} \rfloor < k \leq n} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d), \tag{4.21}$$

where $\lfloor t \rfloor = \max\{k \in \mathbb{N} \mid k \leq t\}$ denotes the largest natural number not larger than $t > 0$. Now we estimate both summands of the right hand side of (4.21) separately. For the first summand we use the upper bound of (4.19), to arrive at

$$\begin{aligned}\sup_{1\leq k\leq \lfloor \frac{n}{\lambda}\rfloor} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) &\leq C' \sup_{1\leq k\leq \lfloor \frac{n}{\lambda}\rfloor} k^{2\alpha} \min\left\{1, \frac{1+\log(d/k)}{k}\right\}^{\alpha} \\ \leq C' \sup_{1\leq k\leq \lfloor \frac{n}{\lambda}\rfloor} k^{\alpha}(1+\log(d/k))^{\alpha} &\leq C' \left(\frac{n(1+\log(\lambda d/n))}{\lambda}\right)^{\alpha}, \end{aligned} \tag{4.22}$$

where we used the fact that the function $t \mapsto t(1+\log(d/t))$ monotonically increases on the interval $[1, d]$ (since its derivative is positive) and it holds $n/\lambda < d$ by assumption. Using $\lambda > 1$ we further estimate

$$\begin{aligned} C' \left(\frac{n(1+\log(\lambda d/n))}{\lambda}\right)^{\alpha} &= C' \left(\frac{n(1+\log(\lambda)+\log(d/n))}{\lambda}\right)^{\alpha} \\ &\leq C' \left(\frac{1+\log(\lambda)}{\lambda} \cdot n(1+\log(d/n))\right)^{\alpha}. \end{aligned} \tag{4.23}$$

Next we estimate the second summand of the right hand side of (4.21) by using the monotonicity of Gelfand numbers

$$\sup_{\lfloor \frac{n}{\lambda}\rfloor < k\leq n} k^{2\alpha} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \leq n^{2\alpha} c_{\lceil \frac{n}{\lambda}\rceil}(\mathrm{id}\colon \ell_p^d \to \ell_2^d), \tag{4.24}$$

where $\lceil t \rceil = \min\{k \in \mathbb{N} \mid k \geq t\}$ denotes the smallest natural number not smaller than $t > 0$. Combining the estimates (4.20), (4.21), (4.22), (4.23) and (4.24), we end up with

$$\gamma_{2\alpha} c_{\lceil \frac{n}{\lambda}\rceil}(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq \left(\frac{1+\log(d/n)}{n}\right)^{\alpha} \left(C_0 - C'\gamma_{2\alpha}\left(\frac{1+\log(\lambda)}{\lambda}\right)^{\alpha}\right).$$

Observing that $(1+\log(\lambda))/\lambda \to 0$ for $\lambda \to \infty$, there exists some $\lambda_0 > 1$ such that

$$c_{\lceil \frac{n}{\lambda_0}\rceil}(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq C'' \left(\frac{1+\log(d/n)}{n}\right)^{\alpha} \tag{4.25}$$

holds for all $n \in \mathbb{N}$ with $\log(d) \leq n \leq d$ and a constant $C'' > 0$. Let us note that (4.25) still remains true in the case $n < \lambda_0$ with only minor modifications on the argument. If $n < \lambda_0$, the first supremum in (4.21) becomes empty and (4.25) simply follows from (4.20) and (4.24).

Next we show the claim for all $k \in \mathbb{N}$ with $\log(d) \leq k \leq d/\lambda_0$ if such a k exists. We set $n = \lfloor \lambda_0(k-1)+1 \rfloor$ such that $\lceil n/\lambda_0 \rceil = k$ and $n \leq \lambda_0 k$. Since $\log(d) \leq n \leq d$ we can apply (4.25) to obtain

$$c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) = c_{\lceil n/\lambda_0\rceil}(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq C'' \left(\frac{1+\log(d/n)}{n}\right)^{\alpha}$$

and by using the monotonicity of the function $t \mapsto (1+\log(d/t))/t$ we arrive at

$$\begin{aligned} c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) &\geq C'' \left(\frac{1+\log(d/n)}{n}\right)^{\alpha} \geq C'' \left(\frac{1+\log(d/(\lambda_0 k))}{\lambda_0 k}\right)^{\alpha} \\ &\geq \frac{C''}{\lambda_0^{\alpha}(1+\lambda_0)^{\alpha}} \left(\frac{1+\log(d/k)}{k}\right)^{\alpha}. \end{aligned}$$

Hence, we have shown the claim for all $\log(d) \leq k \leq d/\lambda_0$ and it remains to consider the two cases $k < \log(d)$ and $d/\lambda_0 < k \leq d$. If $k < \log(d)$, we again use the monotonicity to get

$$c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq c_{\lceil \log(d) \rceil}(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq C'' \left(\frac{1 + \log(d) - \log(\log(d))}{\log(d)} \right)^{\alpha} \geq C''',$$

and if $d/\lambda_0 < k \leq d$, the claim follows from

$$c_k(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \geq c_d(\mathrm{id}\colon \ell_p^d \to \ell_2^d) = \| \mathrm{id}\colon \ell_2^d \to \ell_p^d \|^{-1} = \frac{1}{d^{\alpha}} = \left(\frac{1 + \log(d/d)}{d} \right)^{\alpha},$$

where we used the second part of Lemma 4.16. □

The previous theorem provides an upper and lower bound on the Gelfand numbers $c_n(\mathrm{id}\colon \ell_p^d \to \ell_2^d)$ of the identity operator from ℓ_p^d to ℓ_2^d. Hence, by equivalence of Gelfand and compressive n-width, from the previous Theorem 4.26 we can now deduce the result of Donoho, cf. [35, Theorem 1].

Corollary 4.28 ([35])**.** *For $R > 0$ and $0 < p < 2$ it holds*

$$E_n(R\overline{B_p^d}, \ell_2^d) \asymp R \min\left\{1, \frac{1 + \log(d/n)}{n}\right\}^{\frac{1}{p} - \frac{1}{2}},$$

where the constants of equivalence do not depend on n or d.

Proof. By (4.17) and Theorem 4.25 we get

$$E_n(R\overline{B_p^d}, \ell_2^d) \asymp d_n(R\overline{B_p^d}, \ell_2^d) = R d_n(\overline{B_p^d}, \ell_2^d) = R c_{n+1}(\mathrm{id}\colon \ell_p^d \to \ell_2^d).$$

Applying the previous Theorem 4.27, we end up with

$$\begin{aligned} E_n(R\overline{B_p^d}, \ell_2^d) &\asymp R c_{n+1}(\mathrm{id}\colon \ell_p^d \to \ell_2^d) \asymp R \min\left\{1, \frac{1 + \log(d/(n+1))}{n+1}\right\}^{\frac{1}{p} - \frac{1}{2}} \\ &\asymp R \min\left\{1, \frac{1 + \log(d/n)}{n}\right\}^{\frac{1}{p} - \frac{1}{2}}, \end{aligned}$$

as claimed. □

Remark 4.29. If we are interested in an encoder/decoder pair, which reconstructs every signal $x \in B_p^d$ with an accuracy of $\varepsilon > 0$, i.e., if we want

$$E_n(R\overline{B_p^d}, \ell_2^d) \asymp R \min\left\{1, \frac{1 + \log(d/n)}{n}\right\}^{\frac{1}{p} - \frac{1}{2}} < \varepsilon,$$

we need at least

$$n > (1 + \log(d/n))(R^2 \varepsilon^{-2})^{\frac{p}{2-p}}$$

measurements. By choosing $p = 1$ we in particular get

$$n > (1 + \log(d/n)) R^2 \varepsilon^{-2},$$

which goes in hand with the results of compressed sensing presented in Chapter 3.

Chapter 5

Sparse Recovery from Binary Measurements via ℓ_1-Support Vector Machines

In Chapter 3 we discussed the recovery of sparse signals $x \in \mathbb{R}^d$ from linear measurements $Ax = y$. But in some applications the sensing process might not follow a linear structure. For example, in analog-to-digital conversion we have to take quantization as specific nonlinearity into account. The problem of 1-*bit compressed sensing* deals with quantization in its most extreme case, that is, with the recovery of x from 1-bit measurements of the form

$$y_i = \operatorname{sign}(\langle a_i, x \rangle), \qquad i = 1, \dots, m \tag{5.1}$$

for some measurement vectors $a_i \in \mathbb{R}^d$, where $\operatorname{sign}(t) = +1$ if $t \geq 0$ and $\operatorname{sign}(t) = -1$ if $t < 0$ denotes the *sign* of $t \in \mathbb{R}$. Here the quantizer takes the form of a comparator to zero, which is a quite inexpensive and fast hardware device. Hence, it would be beneficial if we could already recover x only from its single bit information (5.1).

Before going on, let us note that the measurements (5.1) are independent of the length of x, i.e., rescaling x by a positive factor does not change the y_i. Therefore, we usually assume x to be normalized, i.e., $\|x\|_2 = 1$. Furthermore, even if we know that the signal x is s-sparse, we observe that x is usually not uniquely determined by the a_i's, y_i's and s. Indeed, typically x lies in the open set

$$\bigcap_{i=1}^{m} \{w \in \mathbb{R}^d \mid y_i \cdot \operatorname{sign}(\langle a_i, w \rangle) > 0\},$$

which then also contains a small neighborhood of x. Any other s-sparse signal $x_0 \neq x$ contained in this neighborhood then gives the same measurements (5.1) as x. Hence, from only knowing a_i, y_i and s we cannot distinguish between x and x_0, so we cannot expect to recover x exactly by any method whatsoever and the best we can hope for is to achieve at least a good approximation.

The problem of 1-bit compressed sensing was originally introduced by Boufounos and Baraniuk in 2008 [10] followed by several lines of research, for example see [2, 51, 96] and references therein. Among all algorithms proposed for the recovery

of x from (5.1) let us highlight the convex maximization problem analyzed in [96], which is given by

$$\max_{w\in\mathbb{R}^d} \sum_{i=1}^{m} y_i\langle a_i, w\rangle \quad \text{subject to} \quad \|w\|_2 \leq 1,\ \|w\|_1 \leq R \tag{5.2}$$

for some parameter $R > 0$ controlling the sparsity level of the solution.

Theorem 5.1 ([96])**.** *Let $a_1, \dots, a_m \in \mathbb{R}^d$ be i.i.d. standard Gaussian vectors, let $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$ and $\|x\|_1 \leq R$ and assume that the measurements $y_1, \dots, y_m$ are given by* (5.1)*. Then there exist constants $c, C > 0$ such that for any $\varepsilon > 0$ with*

$$m \geq C\varepsilon^{-2}R^2 \log(2d/R^2)$$

any solution $\hat{x}$ of (5.2) *satisfies*

$$\|x - \hat{x}\|_2^2 \leq \varepsilon$$

with probability at least $1 - 8\exp(-c\varepsilon^2 m)$.

Remark 5.2. The previous theorem is only a simplified version of the results presented in [96]. The authors generalized the compressibility conditions $\|x\|_2 = 1$ and $\|x\|_1 \leq R$ by assuming that x lies in a (known) set $K \subset \mathbb{R}^d$ with small *Gaussian mean width* $\omega(K)$, which can be understood as *effective dimension* of K. Furthermore, they also considered noisy measurements and they gave uniform recovery results.

Returning to the initial problem of 1-bit compressed sensing, we want to recover the sparse classifier $x \in \mathbb{R}^d$ from its nonlinear measurements (5.1). An approximation $\hat{x} \in \mathbb{R}^d$ of x should, therefore, be at least consistent with (most of) the measurements, i.e., for (almost every) $i = 1, \dots, m$, it should hold

$$\operatorname{sign}(\langle a_i, \hat{x}\rangle) = y_i = \operatorname{sign}(\langle a_i, x\rangle).$$

This means that the hyperplane $E_{\hat{x}} = \{w \in \mathbb{R}^d \mid \langle w, \hat{x}\rangle = 0\} \subset \mathbb{R}^d$ with normal vector $\hat{x}$ (almost) separates the two sets

$$C_+ = \{a_i \mid y_i = +1\} \quad \text{and} \quad C_- = \{a_i \mid y_i = -1\}$$

from each other. From this point of view, instead of searching for an approximation $\hat{x}$ of x, we could equivalently also search for a hyperplane separating C_+ and C_-. Since there are in general infinitely many of those hyperplanes, we have to choose one of them. But this is exactly the problem of the so-called *support vector machines* (SVMs), which seek for the separating hyperplane maximizing the distance (with respect to some norm) to both sets C_+ and C_-. Hence, SVMs are suitable for the problem of 1-bit compressed sensing. The aim of this chapter is to analyze the so-called ℓ_1-SVM in order to prove a similar approximation result as Theorem 5.1, which describes the performance of Algorithm (5.2).

5.1 Support Vector Machines

The aim of this section is to introduce and discuss the so-called *soft margin* and the *hard margin* SVMs. Since they seek for a hyperplane which separates two (finite) sets C_+ and C_- from another while simultaneously maximizing the distance to both of them, we start this section by calculating the distance of some point in $\mathbb{R}^d$ to a given hyperplane with respect to a general norm.

5.1.1 Distance to Hyperplanes

We start this section with the definition of the dual norm, which will turn out to be helpful to characterize the distance of some points $a_i \in \mathbb{R}^d$, $i = 1, \dots, m$ to a hyperplane $E \subset \mathbb{R}^d$. After calculating this distance we can easily formulate the optimization problem of support vector machines.

Definition 5.3 (Dual norm)**.** The *dual norm* $\|\cdot\|'$ of the norm $\|\cdot\|$ on $\mathbb{R}^d$ is defined by

$$\|\cdot\|' \colon \mathbb{R}^d \to [0,\infty), \quad w \mapsto \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} \langle v, w\rangle = \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} |\langle v, w\rangle|.$$

Example 5.4. For $1 \le p \le \infty$ let $1 \le q \le \infty$ be such that $1/p + 1/q = 1$, where we set $q = \infty$ if $p = 1$ and vice versa. Then

$$\|\cdot\|_p' = \|\cdot\|_q \quad \text{and} \quad \|\cdot\|_q' = \|\cdot\|_p.$$

Lemma 5.5. *Let $\|\cdot\|$ denote a norm on $\mathbb{R}^d$. Then we have the following:*

i) The dual norm $\|\cdot\|'$ of $\|\cdot\|$ defines a norm on $\mathbb{R}^d$.

ii) $\|\cdot\|$ and its dual norm $\|\cdot\|'$ satisfy the Cauchy-Schwarz inequality, i.e.,

$$|\langle v, w\rangle| \le \|v\| \cdot \|w\|'$$

holds for every $v, w \in \mathbb{R}^d$.

iii) It holds $(\|\cdot\|')' = \|\cdot\|$.

Proof. i) Clearly the dual norm is nonnegative and positive homogeneous, so it remains to prove the triangle inequality. For any $w_1, w_2 \in \mathbb{R}^d$ we obtain

$$\begin{aligned}\|w_1 + w_2\|' &= \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} \langle v, w_1 + w_2\rangle = \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} (\langle v, w_1\rangle + \langle v, w_2\rangle) \\ &\le \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} \langle v, w_1\rangle + \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} \langle v, w_2\rangle = \|w_1\|' + \|w_2\|'.\end{aligned}$$

ii) Let $v, w \in \mathbb{R}^d$ with $v \neq 0$, then we get

$$|\langle v, w\rangle| = \|v\| \cdot \Big|\langle \frac{v}{\|v\|}, w\rangle\Big| \le \|v\| \cdot \sup_{\substack{\tilde{v}\in\mathbb{R}^d \\ \|\tilde{v}\|=1}} |\langle \tilde{v}, w\rangle| = \|v\| \cdot \|w\|'.$$

iii) For $w \in \mathbb{R}^d$ we consider the functional $\langle \cdot, w\rangle \colon \mathbb{R}^d \to \mathbb{R}$, $v \mapsto \langle v, w\rangle$. Its operator norm with respect to $\|\cdot\|$ is given by

$$\|\langle \cdot, w\rangle\| = \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|=1}} \langle v, w\rangle = \|w\|'$$

and using the theorem of Hahn-Banach we end up with

$$\|w\| = \sup_{\substack{v\in\mathbb{R}^d \\ \|\langle\cdot, v\rangle\|=1}} \langle v, w\rangle = \sup_{\substack{v\in\mathbb{R}^d \\ \|v\|'=1}} \langle v, w\rangle = \|w\|''.$$

□

Next we want to calculate the *distance*

$$d_{\|\cdot\|}(a, E) := \inf_{v \in E} \|a - v\| \tag{5.3}$$

of some point $a \in \mathbb{R}^d$ to the hyperplane $E \subset \mathbb{R}^d$ with respect to the norm $\|\cdot\|$. Furthermore, $P_{\|\cdot\|}(a, E) \in E$ is called *proximum* of a onto E with respect to $\|\cdot\|$ if it is a minimizer of the distance, i.e., if it holds

$$\left\| a - P_{\|\cdot\|}(a, E) \right\| = \inf_{v \in E} \|a - v\|. \tag{5.4}$$

Note that we only consider the finite-dimensional vector space $\mathbb{R}^d$. Hence, by continuity arguments, there always exists at least one proximum. Although the distance of some point to the hyperplane is always unique, the proximum $P_{\|\cdot\|}(a, E_w)$ however may not. There might be more than one proximum if the unit ball with respect to the norm $\|\cdot\|'$ contains a straight line segment, such as the unit ball with respect to the ℓ_1- or the ℓ_∞-norm.

Before we will go on with the general case, i.e., with respect to some arbitrary norm on $\mathbb{R}^d$, let us first consider the particularly well known case of the Euclidean norm.

For $w \in \mathbb{R}^d \backslash \{0\}$, let $E_w = \{v \in \mathbb{R}^d \mid \langle v, w \rangle = 0\}$ denote the hyperplane with normal vector w. For a point $a \in \mathbb{R}^d$ it is well known that the proximum of a onto E_w with respect to the Euclidean norm $\|\cdot\|_2$ is given by the orthogonal projection of a onto E_w. Hence, we can represent the proximum $P_{\|\cdot\|_2}(a, E_w)$ as a linear combination of a and w, i.e., for some $\lambda \in \mathbb{R}$ we have

$$P_{\|\cdot\|_2}(a, E_w) = a + \lambda w. \tag{5.5}$$

Since $P_{\|\cdot\|_2}(a, E_w)$ is contained in E_w, we observe

$$0 = \left\langle P_{\|\cdot\|_2}(a, E_w)\,,\, w \right\rangle = \langle a, w \rangle + \lambda \|w\|_2^2.$$

Therefore, $P_{\|\cdot\|_2}(a, E_w) = a - \langle a, w \rangle / \|w\|_2^2 \cdot w$ and the distance $d_{\|\cdot\|_2}(a, E_w)$ between the hyperplane E_w and a with respect to the ℓ_2-norm is given by

$$d_{\|\cdot\|_2}(a, E_w) = \left\| P_{\|\cdot\|_2}(a, E_w) - a \right\|_2 = \|a + \lambda w - a\|_2 = \frac{|\langle a, w \rangle|}{\|w\|_2}.$$

Using the duality $\|\cdot\|_2' = \|\cdot\|_2$, the distance $d_{\|\cdot\|_2}(a, E_w)$ can also be express by

$$d_{\|\cdot\|_2}(a, E_w) = \frac{|\langle a, w \rangle|}{\|w\|_2'}.$$

The next theorem shows that this formula remains true for an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, if we replace the term $\|w\|_2'$ by $\|w\|'$, cf. [78, Theorem 2.2].

Theorem 5.6 ([78])**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and for $w \in \mathbb{R}^d \backslash \{0\}$, let $E_w = \{v \in \mathbb{R}^d \mid \langle v, w \rangle = 0\} \subset \mathbb{R}^d$ denote the hyperplane with normal vector w. Then the distance of some point $a \in \mathbb{R}^d$ to the hyperplane E_w with respect to $\|\cdot\|$ is given by*

$$d_{\|\cdot\|}(a, E_w) = \frac{|\langle a, w \rangle|}{\|w\|'}. \tag{5.6}$$

Furthermore, the proximum $P_{\|\cdot\|}(a, E_w)$ of a onto E_w is given by

$$P_{\|\cdot\|}(a, E_w) = a - \frac{\langle a, w\rangle}{\|w\|'} \cdot \arg\max_{\|v\|=1} \langle v, w\rangle. \tag{5.7}$$

Remark 5.7. The proximum $P_{\|\cdot\|}(a, E_w)$ is not always unique. However, if it is not uniquely determined, every proximum is still of the form (5.7) with only possible different choices for the maximizer of $\arg\max_{\|v\|=1}\langle v, w\rangle$.

Proof. This result is by far not new and, for instance, contained in [78, Theorem 2.2.]. But since the proof there is based on the so-called *Karush-Kuhn-Tucker saddle point sufficient optimality criterion* which we have not discussed here, we will give an alternate proof instead.

We start calculating the proximum $P_{\|\cdot\|}(a, E_w)$ first, which afterwards allows us to easily calculate the distance between a and E_w. For this purpose we assume $a \notin E_w$, i.e., $\langle a, w\rangle \neq 0$, since otherwise the statement follows trivially.

Following the calculation for the ℓ_2-norm (5.5), let us first find the right direction $e \in \mathbb{R}^d$ with $\|e\| = 1$ such that

$$P_{\|\cdot\|}(a, E_w) = a + \lambda e \tag{5.8}$$

for some $\lambda \in \mathbb{R}$. Since the proximum has to be contained in the hyperplane E_w, we obtain

$$0 = \left\langle P_{\|\cdot\|}(a, E_w)\,,\, w \right\rangle = \langle a, w\rangle + \lambda\langle e, w\rangle.$$

Further, using the assumption $\langle a, w\rangle \neq 0$, we get

$$\left\| P_{\|\cdot\|}(a, E_w) - a \right\| = |\lambda| = \left| \frac{\langle a, w\rangle}{\langle e, w\rangle} \right| = \min_{v \in E_w} \|a - v\|.$$

Minimizing over $|\lambda|$ yields

$$e = \arg\max_{\|v\|=1} \langle v, w\rangle \quad \text{and} \quad \lambda = \frac{\langle a, w\rangle}{\langle e, w\rangle} = \frac{\langle a, w\rangle}{\|w\|'}.$$

Combined with (5.8) this results into

$$P_{\|\cdot\|}(a, E_w) = a - \frac{\langle a, w\rangle}{\|w\|'} \cdot \arg\max_{\|v\|=1} \langle v, w\rangle$$

and

$$d_{\|\cdot\|}(a, E_w) = \|a - P_{\|\cdot\|}(a, E_w)\| = |\lambda| = \frac{|\langle a, w\rangle|}{\|w\|'},$$

as claimed. □

5.1.2 Hard Margin Support Vector Machines

Given a norm $\|\cdot\|$ on $\mathbb{R}^d$ and measurements

$$y_i = \operatorname{sign}(\langle a_i, x\rangle), \quad i = 1, \dots, m \tag{5.9}$$

of the signal $x \in \mathbb{R}^d$ for certain measurement vectors $a_i \in \mathbb{R}^d$, a support vector machine searches for the hyperplane $E_w \subset \mathbb{R}^d$ with normal vector $w \neq 0$, which is on the one hand consistent with the measurements, i.e., such that

$$y_i = \operatorname{sign}(\langle a_i, x\rangle) = \operatorname{sign}(\langle a_i, w\rangle), \quad i = 1, \dots, m \tag{5.10}$$

and, on the other hand, simultaneously maximizes the distance to the sample points a_i. Here we define the distance of the sample points to the hyperplane as minimal distance, i.e.,

$$d_{\|\cdot\|}(a_1, \dots, a_m, E_w) := \min_{i\in[m]} d_{\|\cdot\|}(a_i, E_w). \tag{5.11}$$

If $A \in \mathbb{R}^{m,d}$ denotes the matrix with rows a_i, to shorten notation, we set $d_{\|\cdot\|}(A, E_w) = d_{\|\cdot\|}(a_1, \dots, a_m, E_w)$. Further, by Theorem 5.6, this distance is given by

$$d_{\|\cdot\|}(A, E_w) = \min_{i\in[m]} \frac{|\langle a_i, w\rangle|}{\|w\|'}. \tag{5.12}$$

If we multiply the normal vector w by a factor $\lambda \neq 0$, the hyperplane E_w does not change, i.e., it holds $E_w = E_{\lambda w}$ for any $\lambda \neq 0$. Hence, without loss of generality we may assume that w is normalized with respect to the dual norm $\|\cdot\|'$. Combining (5.10) and (5.12), we end up with the following definition for the (hard margin) SVM.

Definition 5.8. Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and for $x, a_1, \dots, a_m \in \mathbb{R}^d$ let

$$y_i = \operatorname{sign}(\langle a_i, x\rangle), \quad i = 1, \dots, m.$$

Then, the optimization problem

$$\max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \left\{ \min_{i\in[m]} y_i\langle a_i, w\rangle \right\} \tag{5.13}$$

is called *(hard margin) support vector machine.*

Remark 5.9.

i) The support vector machine (5.13) does not always have a unique solution. But since we are maximizing the continuous function $w \mapsto \min_{i\in[m]} y_i\langle a_i, w\rangle$ over the compact set $\mathcal{S}_{\|\cdot\|'} := \{w \in \mathbb{R}^d \mid \|w\|' = 1\}$, there always exists at least one maximizer.

ii) For fixed y_i and a_i, note that $w \mapsto \min_{i\in[m]} y_i\langle a_i, w\rangle$ is a concave function. Moreover, instead of maximizing over $\{w \in \mathbb{R}^d \mid \|w\|' = 1\}$ in (5.13) we could also maximize over the convex unit ball $\{w \in \mathbb{R}^d \mid \|w\|' \leq 1\}$. Hence, the hard margin SVM (5.13) can be recast as a convex optimization problem.

Lemma 5.10. *Suppose that the SVM* (5.13) *has two different maximizers* $\hat{x}_1, \hat{x}_2 \in \mathbb{R}^d$*. Then every point in the line segment*

$$\overline{\hat{x}_1\hat{x}_2} := \{\lambda\hat{x}_1 + (1-\lambda)\hat{x}_2 \mid \lambda \in [0,1]\}$$

from $\hat{x}_1$ *to* $\hat{x}_2$ *also maximizes* (5.13) *and is normalized with respect to the dual norm* $\|\cdot\|'$*.*

Proof. Let $\hat{x}_1$ and $\hat{x}_2$ be two different maximizers of (5.13) and let $\lambda \in (0,1)$. Setting $\hat{x}_\lambda = \lambda\hat{x}_1 + (1-\lambda)\hat{x}_2$ we first observe

$$\|\hat{x}_\lambda\|' = \|\lambda\hat{x}_1 + (1-\lambda)\hat{x}_2\|' \leq \lambda\|\hat{x}_1\|' + (1-\lambda)\|\hat{x}_2\|' = 1.$$

Further, using that $\hat{x}_1$ and $\hat{x}_2$ are maximizers of (5.13), we obtain

$$\begin{aligned}
&\min_{i=1,\dots,m} y_i\langle a_i, \frac{\hat{x}_\lambda}{\|\hat{x}_\lambda\|'}\rangle = \frac{1}{\|\hat{x}_\lambda\|'}\Big(\min_{i\in[m]} \lambda y_i\langle a_i, \hat{x}_1\rangle + (1-\lambda) y_i\langle a_i, \hat{x}_2\rangle\Big) \\
&\geq \frac{1}{\|\hat{x}_\lambda\|'}\Big(\lambda \min_{i\in[m]} y_i\langle a_i, \hat{x}_1\rangle + (1-\lambda)\min_{i\in[m]} y_i\langle a_i, \hat{x}_2\rangle\Big) \\
&= \frac{1}{\|\hat{x}_\lambda\|'}\Big(\max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \Big\{\min_{i\in[m]} y_i\langle a_i, w\rangle\Big\}\Big) \geq \max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \Big\{\min_{i\in[m]} y_i\langle a_i, w\rangle\Big\}.
\end{aligned}$$

To not end up with a contradiction, it must hold equality, i.e., we get

$$\frac{1}{\|\hat{x}_\lambda\|'}\Big(\max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \Big\{\min_{i\in[m]} y_i\langle a_i, w\rangle\Big\}\Big) = \max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \Big\{\min_{i\in[m]} y_i\langle a_i, w\rangle\Big\}.$$

This implies $\|\hat{x}_\lambda\|' = 1$ and $\hat{x}_\lambda$ is another maximizer of (5.13), as claimed. □

Corollary 5.11. *If the support vector machine* (5.13) *has no unique maximizer, the unit sphere* $\mathcal{S}_{\|\cdot\|'} = \{w \in \mathbb{R}^d \mid \|w\|' = 1\}$ *contains a straight line segment.*

Assume that the hyperplane E_w does not contain any sample point a_i and is consistent with the measurements y_i, i.e., it holds $y_i\langle a_i, w\rangle > 0$ for $i = 1, \dots, m$. By rescaling the normal vector w of E_w (which does not change the hyperplane), we can always achieve $y_i\langle a_i, \tilde{w}\rangle \geq 1$, where we set $\tilde{w} = \lambda w$ for some $\lambda > 0$. This simple observation motivates the following equivalent reformulation of the SVM (5.13):

$$\min_{w\in\mathbb{R}^d} \|w\|' \quad \text{subject to} \quad y_i\langle a_i, w\rangle \geq 1 \tag{5.14}$$

The next lemma shows that the two optimization problems (5.13) and (5.14) are indeed equivalent.

Lemma 5.12. *Let* $\hat{x}$ *be a maximizer of the support vector machine* (5.13) *and let* $\tilde{x}$ *be a minimizer of* (5.14)*. Then, using the notation of Definition 5.8,* $\tilde{x}/\|\tilde{x}\|'$ *is a maximizer of* (5.13) *and* $\hat{x}/d_{\|\cdot\|}(A, E_{\hat{x}})$ *is a minimizer of* (5.14)*. Further, it holds*

$$d_{\|\cdot\|}(A, E_{\tilde{x}}) = \min_{i\in[m]} d_{\|\cdot\|}(a_i, E_{\tilde{x}}) = \min_{i\in[m]} \min_{v\in E_{\tilde{x}}} \|a_i - v\| = \frac{1}{\|\tilde{x}\|'}. \tag{5.15}$$

Proof. A proof of this well known result is contained in [29, Theorem 9.13.] with the particular choice $\|\cdot\| = \|\cdot\|' = \|\cdot\|_2$. To show that the arguments presented there also hold in the general case, we recast its proof with only minor modifications.

We start by showing the statement (5.15). Since $\tilde{x}$ is a minimizer of (5.14) it holds

$$\min_{i\in[m]} y_i\langle a_i, \tilde{x}\rangle = 1$$

and, using Theorem 5.6, we observe

$$d_{\|\cdot\|}(A, E_{\tilde{x}}) = \min_{i\in[m]} d(a_i, E_{\tilde{x}}) = \min_{i\in[m]} \frac{|\langle a_i, \tilde{x}\rangle|}{\|\tilde{x}\|'} = \frac{1}{\|\tilde{x}\|'}.$$

Next let us show that $\tilde{x}/\|\tilde{x}\|'$ is a maximizer of (5.13). Towards a contradiction, we assume that

$$\max_{\substack{w\in\mathbb{R}^d \\ \|w\|'=1}} \Big\{ \min_{i=1,\dots,d} y_i\langle a_i, w\rangle \Big\} = d_{\|\cdot\|}(A, E_{\hat{x}}) > d_{\|\cdot\|}(A, E_{\tilde{x}}) = \frac{1}{\|\tilde{x}\|'}.$$

Setting $x_0 = \hat{x}/d(A, E_{\hat{x}})$ implies $y_i\langle a_i, x_0\rangle \geq 1$ for $i = 1, \dots, m$. Hence, x_0 is feasible for (5.14) and we end up with

$$\|x_0\|' = \frac{\|\hat{x}\|'}{d_{\|\cdot\|}(A, E_{\hat{x}})} = \frac{1}{d_{\|\cdot\|}(A, E_{\hat{x}})} < \frac{1}{d_{\|\cdot\|}(A, E_{\tilde{x}})} = \|\tilde{x}\|',$$

which is a contradiction to the assumption that $\tilde{x}$ is a minimizer of (5.14).

It remains to show that $x_1 = \hat{x}/d_{\|\cdot\|}(A, E_{\hat{x}})$ is a minimizer of (5.14). We observe

$$\min_{i\in[m]} y_i\langle a_i, x_1\rangle = \min_{i\in[m]} \frac{y_i\langle a_i, \hat{x}\rangle}{d_{\|\cdot\|}(A, E_{\hat{x}})} = 1,$$

which in particular shows $y_i\langle a_i, x_1\rangle \geq 1$ for $i = 1, \dots, m$. Hence, x_1 is feasible for (5.14) and it remains to show $\|x_1\|' = \|\tilde{x}\|'$. Since $\hat{x}$ is a maximizer of (5.13), it holds $d(A, E_{\tilde{x}}) \leq d(A, E_{\hat{x}}) = d(A, E_{x_1})$ and we arrive at

$$\frac{1}{\|x_1\|'} = \min_{i\in[m]} \frac{y_i\langle a_i, x_1\rangle}{\|x_1\|'} = d(A, E_{x_1}) \geq d(A, E_{\tilde{x}}) = \frac{1}{\|\tilde{x}\|'},$$

from which the statement follows. □

For sample points $a_1, \dots, a_m$ and measurements $y_1, \dots, y_m$ let $\hat{x} \in \mathbb{R}^d$ be a minimizer of the hard margin SVM (5.14). Then there is no sample point a_i contained in the so-called *margin*

$$M := \{v \in \mathbb{R}^d \mid -1 < \langle v, \hat{x}\rangle < 1\}, \tag{5.16}$$

which explains why we denote (5.14) (and the equivalent formulation (5.13)) as *hard margin SVMs*. Furthermore, there are always sample points a_i lying on the boundary of M. This means, the set

$$SV := \{a_i \mid 1 = |\langle a_i, \hat{x}\rangle|\}$$

is not empty and we denote its elements as *support vectors*. By construction, those support vectors already determine the solution $\hat{x}$ of the hard margin SVM and the other sample points $a_i \notin SV$ are completely ignored by the SVM. Typically, there are only few support vectors. Hence, the vector $v \in \mathbb{R}^m$ with entries $v_i = 1$ if $a_i \in SV$ and $v_i = 0$ else is sparse. But note that the support vectors and, therefore, also the nonzero entries of v depend on the sample points a_i. Hence, here we have another type of sparsity compared to the usual setting of compressed sensing, where the nonzero entries of the signal are fixed in advance.

Although the sparsity of the support vectors can be used to reduce the computational complexity, this restriction also brings a major drawback. Namely, if only few vectors already determine the separating hyperplane, most of the other sample points are completely ignored. This can lead to unsatisfactory results, for instance, if few of the sample points a_i do not fit into the structure of the remaining samples, but are much closer to the origin than the other, cf. Figure 5.1. To overcome this drawback, an idea is to introduce weights $\xi_i \in [0, 1]$, which interpolate between the two states of being a support vector or not. The weight ξ_i should be close to one if the sample point a_i should have a big influence for the separating hyperplane. If those choices are done carefully, this would lead to a compressible vector $(\xi_i) \in \mathbb{R}^m$, instead of the sparse vector v.

A second major drawback of hard margin SVMs is their behavior in the presence of misclassifications as a particular kind of noise. This means, if some of the sample points a_i get misclassified such that, instead of the true label y_i, we observe the flipped value $\tilde{y}_i = -y_i$, in general, no separating hyperplane exists anymore. Hence, in that case the hard margin SVMs completely fail, since the set of hyperplanes they are optimizing over is empty.

In order to overcome both mentioned drawbacks which, in particular, makes the SVMs more robust for practical purposes, in the next subsection we will introduce the so-called *soft margin support vector machines*.

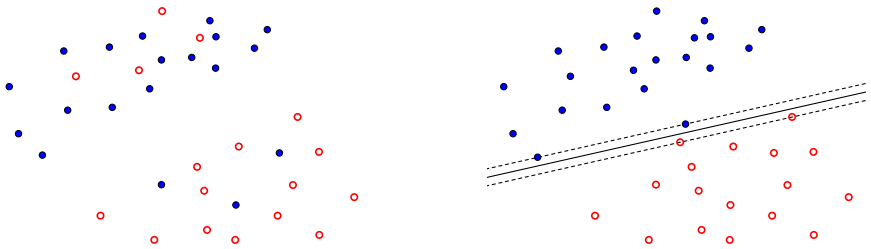

Figure 5.1: Due to misclassifications, no separating hyperplane between the blue and the red points exists (left) and a hyperplane which is already determined by few outliers, which do not exactly fit with the structure of the remaining points (right).

5.1.3 Soft Margin Support Vector Machines

Support vector machines became a standard tool in the analysis of high dimensional classification problems and are used in many different areas. However, in practice the measurement processes are usually corrupted by noise, which often leads to misclassifications. For instance, if a doctor has to distinguish whether a patient has

a certain disease, his/her guess might be wrong. In that case a separation of the two classes by a hyperplane is probably not possible anymore.

To overcome this drawback and to make the SVMs applicable for practical purposes, we introduce so-called *slack variables* $\xi_i \geq 0$. The slack variables should give the possibility that some sample point a_i lies "on the wrong side" of the separating hyperplane or, similar, in the margin M given by (5.16). That is, comparing with the hard margin SVM (5.14), the slack variables should give the possibility that it holds $y_i\langle a_i, w\rangle < 1$ for some of the sample points a_i. Nevertheless, we still want to find the hyperplane that, at least in some sense, maximizes the distance to the two sets $\{a_i \mid y_i = +1\}$ and $\{a_i \mid y_i = -1\}$. Hence, by allowing sample points a_i satisfying $y_i\langle a_i, w\rangle < 1$, we simultaneously have to penalize them. This gives the idea to define the *soft margin support vector machine* by

$$\min_{\substack{w\in\mathbb{R}^d\\ \xi\in\mathbb{R}^m}} \|\xi\|_1 + \lambda\|w\|' \quad \text{subject to} \quad y_i\langle a_i, w\rangle \geq 1 - \xi_i \tag{5.17}$$

for a trade-off parameter $\lambda > 0$.

Remark 5.13. Following our motivation for the slack variables, it is not clear why we particularly chose the ℓ_1-norm of ξ as penalizer. This freedom leads to different versions of soft margin SVMs. For instance, with the choice of the ℓ_2-norm we obtain the optimization problem

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

which is often denoted as ℓ_2-SVM. In a similar way, a huge amount of different SVMs are generated and can be found in the literature.

As for the hard margin SVM there are several equivalent reformulations of (5.17) which we will discuss in the remaining part of this subsection. We start with the following minimization problem which avoids the explicit use of the slack variables ξ_i:

$$\min_{w\in\mathbb{R}^d} \sum_{i=1}^{m} \big[1 - y_i\langle a_i, w\rangle\big]_+ + \lambda\|w\|' \tag{5.18}$$

for a trade-off parameter $\lambda > 0$, where $[t]_+ := \max\{t, 0\}$ denotes the positive part of $t \in \mathbb{R}$. Before proving the equivalence between (5.18) and (5.17), for brevity we introduce the so-called hinge loss

$$f_x \colon \mathbb{R}^d \to \mathbb{R}, \quad w \mapsto \frac{1}{m}\sum_{i=1}^{m} \big[1 - y_i\langle a_i, w\rangle\big]_+ \tag{5.19}$$

and we set $\tilde{f}_x := m \cdot f_x$. Here the subindex x indicates the dependency of f_x and $\tilde{f}_x$ on x via $y_i = \operatorname{sign}(\langle a_i, x\rangle)$.

Lemma 5.14. *Let $(w^*, \xi^*) \in \mathbb{R}^d \times \mathbb{R}^m$ be a minimizer of* (5.17) *and let $\hat{w} \in \mathbb{R}^d$ be a minimizer of* (5.18)*. Then w^* also minimizes* (5.18) *and, conversely, $(\hat{w}, \hat{\xi})$ is a minimizer of* (5.17)*, where we set $\hat{\xi}_i = \big[1 - y_i\langle a_i, \hat{w}\rangle\big]_+$.*

Proof. By definition of $\hat{\xi}_i$ we obtain $\|\hat{\xi}\|_1 = \tilde{f}_x(\hat{w})$ and

$$1 - \hat{\xi}_i = 1 - \big[1 - y_i\langle a_i, \hat{w}\rangle\big]_+ \leq 1 - (1 - y_i\langle a_i, \hat{w}\rangle) = y_i\langle a_i, \hat{w}\rangle,$$

hence, $(\hat{\xi}, \hat{w})$ is feasible for (5.17). Since (ξ^*, w^*) is a minimizer of (5.17), it follows

$$\|\xi^*\|_1 + \lambda\|w^*\| \leq \|\hat{\xi}\|_1 + \lambda\|\hat{w}\| = \tilde{f}_x(\hat{w}) + \lambda\|\hat{w}\|. \tag{5.20}$$

Further, one easily observes $\xi_i^* \geq 0$ and we get

$$1 - \xi_i^* \leq y_i\langle a_i, w^*\rangle \quad \Rightarrow \quad \xi_i^* \geq \big[1 - y_i\langle a_i, w^*\rangle\big]_+,$$

which yields $\|\xi^*\|_1 \geq \tilde{f}_x(w^*)$. Using that $\hat{w}$ minimizes (5.18) and applying (5.20) we obtain

$$\tilde{f}_x(\hat{w}) + \lambda\|\hat{w}\| \leq \tilde{f}_x(w^*) + \lambda\|w^*\| \leq \|\xi^*\|_1 + \lambda\|w^*\| \leq \tilde{f}_x(\hat{w}) + \lambda\|\hat{w}\|.$$

Hence, it must hold equality which, combined with (5.20), yields the claim. □

Next we will use duality arguments to show that (5.18) also has the two equivalent formulations

$$\min_{w\in\mathbb{R}^d} \tilde{f}_x(w) \quad \text{subject to} \quad \|w\|' \leq R \tag{5.21}$$

for a parameter $R > 0$ and

$$\min_{w\in\mathbb{R}^d} \|w\|' \quad \text{subject to} \quad \tilde{f}_x(w) \leq \tau \tag{5.22}$$

for a parameter $\tau > 0$. We obtain the following theorem:

Theorem 5.15. *i) Let $w^* \in \mathbb{R}^d$ be a minimizer of* (5.18) *for some $\lambda > 0$. Then there exists a $\tau \geq 0$ such that w^* minimizes* (5.22) *as well.*

ii) Let $\tau \geq 0$ be such that (5.22) *has the unique minimizer $w^* \in \mathbb{R}^d$. Then there exists a $R \geq 0$ such that* (5.21) *also has the unique minimizer w^*.*

Proof. i) With $\tau = \tilde{f}_x(w^*) \geq 0$ for any $w \in \mathbb{R}^d$ with $\tilde{f}_x(w) \leq \tau$ we get

$$\tau + \lambda\|w^*\|' = \tilde{f}_x(w^*) + \lambda\|w^*\|' \leq \tilde{f}_x(w) + \lambda\|w\|' \leq \tau + \lambda\|w\|',$$

where we used that $w^* \in \mathbb{R}^d$ is a minimizer of (5.18) for the second last step. Hence, $\|w^*\|' \leq \|w\|'$ and w^* also minimizes (5.22).

ii) If $w^* = 0$, the claim follows trivially, so let us assume $w^* \neq 0$. For $w \neq w^*$ with $\|w\|' \leq \|w^*\|' =: R$, it then follows

$$\tilde{f}_x(w) > \tau \geq \tilde{f}_x(w^*),$$

since otherwise w would be another minimizer of (5.22). Hence, w^* also is the unique minimizer of (5.21).

□

To show the equivalence of (5.18), (5.21) and (5.22), it remains to show that any minimizer of (5.21) can be translated into a minimizer of (5.18). For that we follow [45, Appendix B.5].

For a fixed $R > 0$ we define the *Lagrange function* $L\colon \mathbb{R}^d \times [0,\infty) \to \mathbb{R}$ by

$$L(w,\lambda) := \tilde{f}_x(w) + \lambda(\|w\|' - R). \tag{5.23}$$

The *Lagrange dual problem* is then given by

$$\max_{\lambda\geq 0} \min_{w\in\mathbb{R}^d} L(w,\lambda). \tag{5.24}$$

We are interested in interchanging the maximum with the minimum here, which we will then use to prove the desired equivalence of (5.21) and (5.18). First, for any $w \in \mathbb{R}^d$ we observe

$$\max_{\lambda\geq 0} L(w,\lambda) = \max_{\lambda\geq 0} \tilde{f}_x(w) + \lambda(\|w\|' - R) = \begin{cases} \tilde{f}_x(w), & \text{if } \|w\|' - R \leq 0, \\ \infty, & \text{else,} \end{cases}$$

which gives the identity

$$\min_{w\in\mathbb{R}^d} \max_{\lambda\geq 0} L(w,\lambda) = \min_{\substack{w\in\mathbb{R}^d \\ \|w\|'\leq R}} \tilde{f}_x(w). \tag{5.25}$$

Further, for any $\lambda \geq 0$ we get the estimate

$$\min_{\substack{w\in\mathbb{R}^d \\ \|w\|'\leq R}} \tilde{f}_x(w) \geq \min_{\substack{w\in\mathbb{R}^d \\ \|w\|'\leq R}} \tilde{f}_x(w) + \lambda(\|w\|' - R) \geq \min_{w\in\mathbb{R}^d} \tilde{f}_x(w) + \lambda(\|w\|' - R), \tag{5.26}$$

where the first inequality holds since $\|w\|' - R \leq 0$. Combining (5.25) with (5.26) and taking the maximum with respect to $\lambda \geq 0$ yields

$$\min_{w\in\mathbb{R}^d} \max_{\lambda\geq 0} L(w,\lambda) = \min_{\substack{w\in\mathbb{R}^d \\ \|w\|'\leq R}} \tilde{f}_x(w) \geq \max_{\lambda\geq 0} \min_{w\in\mathbb{R}^d} L(w,\lambda),$$

which is referred to as *weak duality*. Since $\tilde{f}_x$ is a convex function and there exists a $w \in \mathbb{R}^d$ (for instance $w = 0$) such that $\|w\|' - R < 0$, the so-called *Slater's condition* (see, for instance, [45, Theorem B.26] or [11, 5.2.3.]) states that we even have equality, i.e.,

$$\min_{w\in\mathbb{R}^d} \max_{\lambda\geq 0} L(w,\lambda) = \max_{\lambda\geq 0} \min_{w\in\mathbb{R}^d} L(w,\lambda), \tag{5.27}$$

which is referred to as *strong duality*. Now let $\lambda^* \geq 0$ be a maximizer of the right hand side and let w^* be a minimizer of the left hand side of (5.27), that is,

$$\max_{\lambda\geq 0} L(w^*,\lambda) = \min_{w\in\mathbb{R}^d} \max_{\lambda\geq 0} L(w,\lambda) = \max_{\lambda\geq 0} \min_{w\in\mathbb{R}^d} L(w,\lambda) = \min_{w\in\mathbb{R}^d} L(w,\lambda^*).$$

In particular, we get

$$\min_{w\in\mathbb{R}^d} L(w,\lambda^*) \leq L(w^*,\lambda^*) \leq \max_{\lambda\geq 0} L(w^*,\lambda) = \min_{w\in\mathbb{R}^d} L(w,\lambda^*),$$

yielding

$$L(w^*, \lambda^*) = \max_{\lambda \geq 0} L(w^*, \lambda) = \min_{w \in \mathbb{R}^d} L(w, \lambda^*).$$

From this, for any $\lambda \geq 0$ and $w \in \mathbb{R}^d$ we deduce the so-called *saddle point inequality*

$$L(w^*, \lambda) \leq L(w^*, \lambda^*) \leq L(w, \lambda^*). \tag{5.28}$$

Equipped with the saddle point inequality we can now show the missing part of the proof of the equivalences between (5.18), (5.21) and (5.22), namely that any minimizer of (5.21) can be translated into a minimizer of (5.18).

Theorem 5.16. *Let w^* be a minimizer of* (5.21) *for some $R > 0$. Then there exists a $\lambda \geq 0$ such that w^* minimizes* (5.18).

Proof. By continuity of the function $\lambda \mapsto \min_{w \in \mathbb{R}^d} \tilde{f}_x(w) + \lambda(\|w\|' - R)$ and

$$\lim_{\lambda \to \infty} \min_{w \in \mathbb{R}^d} \tilde{f}_x(w) + \lambda(\|w\|' - R) = -\infty,$$

there exists a maximizer $\lambda^* \geq 0$ of the Lagrange dual problem (5.24). Applying the saddle point inequality (5.28), we end up with

$$\tilde{f}_x(w^*) + \lambda^* \|w^*\|' \leq \tilde{f}_x(w) + \lambda^* \|w\|'$$

for every $w \in \mathbb{R}^d$, hence, w^* minimizes (5.18). □

5.2 Recovery via ℓ_1-Support Vector Machines

In 1-bit compressed sensing [10, 96] we aim for the reconstruction of a sparse or compressible classifier $x \in \mathbb{R}^d$ from its 1-bit measurements

$$y_i = \operatorname{sign}(\langle a_i, x \rangle), \quad i = 1, \ldots, m \tag{5.29}$$

for some measurement vectors $a_i \in \mathbb{R}^d$, as we have already discussed in the beginning of this chapter. Following the theory of compressed sensing, to encourage sparsity in the reconstruction we may incorporate the ℓ_1-norm into the recovery algorithm. This idea leads to the so-called ℓ_1-SVM

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{m} \left[1 - y_i \langle a_i, w \rangle\right]_+ \quad \text{subject to} \quad \|w\|_1 \leq R \tag{5.30}$$

for some parameter $R > 0$ [12, 116]. Compared to the soft margin SVM (5.21), we have replaced the dual norm $\|w\|'$ with the particular choice $\|w\|_1$, meaning that the ℓ_1-SVM (5.30) maximizes the distance of the sample points a_i to the separating hyperplane with respect to the ℓ_∞-norm, cf. Lemma 5.5. Note that this is in particular an example where we cannot guarantee the existence of unique minimizers.

The main goal of this section is to analyze the performance of (5.30) in the non-asymptotic regime, i.e., for fixed d and m we want to get error bounds on the difference between the true classifier x and its approximation $\hat{x}$. For this let us first fix the model setup for our analysis.

Using the ℓ_1-SVM (5.30) we aim to recover the true classifier $x \in \mathbb{R}^d$ from the 1-bit measurements (5.29) with some i.i.d. Gaussian measurement vectors

$$a_i = r\tilde{a}_i \in \mathbb{R}^d, \quad \tilde{a}_i \sim \mathcal{N}(0, \mathrm{id}) \tag{5.31}$$

for a fixed scaling parameter $r > 0$. We assume x to be ℓ_2-normalized and compressible in the form

$$\|x\|_2 = 1, \quad \|x\|_1 \leq R \tag{5.32}$$

for some $R > 0$. Note that if x is s-sparse with $\|x\|_2 = 1$, it follows from the Cauchy-Schwarz inequality that

$$\|x\|_1 \leq \sqrt{s}.$$

Hence, we interpret (5.32) in the way that x is effectively $s = R^2$-sparse. Further, for brevity, we set

$$K := R \cdot \overline{B_1^d} = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R\}. \tag{5.33}$$

A minimizer of the ℓ_1-SVM (5.30) will be denoted by $\hat{x}$, i.e.,

$$\hat{x} := \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{i=1}^m \left[1 - y_i \langle a_i, w\rangle\right]_+ \quad \text{subject to} \quad \|w\|_1 \leq R.$$

Before we will go on let us summarize the assumptions we have made.

Standing Assumptions I

i) The true classifier $x \in \mathbb{R}^d$, which we want to approximate, is compressible in the way that $\|x\|_2 = 1$ and $\|x\|_1 \leq R$ for some $R > 0$.

ii) For a scaling parameter $r > 0$ we take the measurement vectors $a_i = r \cdot \tilde{a}_i \in \mathbb{R}^d$ for some i.i.d. $\tilde{a}_i \sim \mathcal{N}(0, \mathrm{id})$.

iii) The measurements are given by $y_i = \mathrm{sign}(\langle a_i, x\rangle)$.

iv) We denote $f_x(w) = \sum_{i=1}^m [1 - y_i\langle a_i, w\rangle]_+$ and $K = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R\}$.

v) $\hat{x}$ denotes a minimizer of the ℓ_1-SVM, i.e., $\hat{x} = \arg\min_{w \in K} f_x(w)$.

Remark 5.17. We introduced the additional scaling parameter $r > 0$ for the Gaussian measurement vectors. This additional parameter is not needed for the 1-bit compressed sensing algorithm (5.2), since the objective function there is actually linear in the a_i's. Hence, multiplying a_i by a positive factor does not change the minimizer of (5.2).

However, this situation completely changes for the ℓ_1-SVM. Assume that we choose $r \ll 1$ sufficiently small, then it follows

$$[1 - y_i\langle a_i, w\rangle]_+ = [1 - ry_i\langle \tilde{a}_i, w\rangle]_+ = 1 - ry_i\langle \tilde{a}_i, w\rangle,$$

so the optimization problem (5.30) can actually be reformulated as

$$\min_{w\in K} \sum_{i=1}^m \left[1 - y_i\langle a_i, w\rangle\right]_+ = \min_{w\in K} \sum_{i=1}^m 1 - y_i\langle a_i, w\rangle = m - \max_{w\in K} \sum_{i=1}^m y_i\langle a_i, w\rangle.$$

Using the duality $\|\cdot\|_1' = \|\cdot\|_\infty$ we can further simplify to

$$\max_{w\in K} \sum_{i=1}^m y_i\langle a_i, w\rangle = R\left\| \sum_{i=1}^m y_i a_i \right\|_\infty.$$

Furthermore, this maximum is achieved if we choose a 1-sparse $w \in K$ with nonzero entry at the position of the largest entry of $\sum_{i=1}^m y_i a_i$.

Hence, in that case the ℓ_1-SVM will always return a 1-sparse solution, no matter what the ground truth signal x is. We conclude that the parameter r might play an important role and should not be chosen too small. This we will also demonstrate in the numerical experiments at the end of this chapter.

The aim of the remainder of this section is to show that $\hat{x}$ (or actually $\hat{x}/\|\hat{x}\|_2$) is a good approximation of x, i.e., that their difference is small. For this proof we adapt the ideas of [96].

Remembering the definition (5.19) of the function f_x, we first observe

$$\begin{aligned} 0 &\le f_x(x) - f_x(\hat{x}) \\ &= \Big(\mathbb{E}\, f_x(x) - \mathbb{E}\, f_x(\hat{x})\Big) + \Big(f_x(x) - \mathbb{E}\, f_x(x)\Big) + \Big(\mathbb{E}\, f_x(\hat{x}) - f_x(\hat{x})\Big) \\ &\le \mathbb{E}\left(f_x(x) - f_x(\hat{x})\right) + 2\sup_{w\in K} |f_x(w) - \mathbb{E}\, f_x(w)|. \end{aligned}$$

Rearranging the terms yields

$$\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) \le 2\sup_{w\in K} |f_x(w) - \mathbb{E}\, f_x(w)|. \tag{5.34}$$

Here the expectation has to be understood as follows: For independent copies a', a_i' of a_i we define

$$\mathbb{E}\, f_x(w) := \mathbb{E}_{a'}\left[\frac{1}{m}\sum_{i=1}^m \left[1 - y_i'\langle a_i', w\rangle\right]_+\right] = \mathbb{E}_{a'}\left[1 - y'\langle a', w\rangle\right]_+, \tag{5.35}$$

where we set $y_i' = \operatorname{sign}(\langle a_i', x\rangle)$. With this interpretation of the expected value, we indeed get

$$\mathbb{E}\, f_x(\hat{x}) \le \sup_{w\in K} \mathbb{E}\, f_x(w),$$

since $\hat{x}$ is independent of the a_i''s, although it clearly depends on the variables a_i. For brevity, in the following we will not use the additional variables a_i', but the expected values always have to be understood as described above.

Following (5.34), it remains to bound the right hand side of this inequality from above and the left hand side from below by terms of the distance between x and $\hat{x}$. With our Standing Assumptions I we obtain the following result, whose proof is given in section 5.2.3, cf. [67, Theorem II.3]:

Theorem 5.18 ([67]). *Let $d \geq 2$, $0 < \varepsilon < 0.18$, $r > \sqrt{2\pi}(0.57 - \varepsilon\pi)^{-1}$ and*

$$m \geq C\varepsilon^{-2} r^2 R^2 \log(d)$$

for a constant $C > 0$. With our Standing Assumptions I it then holds

$$\left\| x - \frac{\hat{x}}{\|\hat{x}\|_2} \right\|_2 \leq \frac{\left\| x - \frac{\hat{x}}{\|\hat{x}\|_2} \right\|_2}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2} \rangle} \leq C' \left(\varepsilon + \frac{1}{r} \right) \tag{5.36}$$

with probability at least $1 - \gamma \exp(-C'' \log(d))$ for some positive constants γ, C', C''.

Remark 5.19. i) The first inequality of (5.36) implies that $\langle x, \hat{x} \rangle > 0$.

ii) In the previous Theorem 5.18 we use the constants γ, C, C' and C'' only for simplicity. More explicit, by taking

$$m \geq 4\varepsilon^{-2} \left(16\sqrt{2\pi} + 19 r R \sqrt{2\log(2d)} \right)^2,$$

it holds

$$\frac{\|x - \hat{x}/\|\hat{x}\|_2\|_2}{\langle x, \hat{x}/\|\hat{x}\|_2 \rangle} \leq 2e^{1/2} \left(\pi\varepsilon + \frac{\sqrt{2\pi}}{r} \right)$$

with probability at least

$$1 - 8 \left(2\exp\left(\frac{-r^2 R^2 \log(2d)}{16} \right) + \exp\left(\frac{-\log(2d)}{16} \right) \right).$$

iii) If we introduce an additional parameter $t > 0$ in Theorem 5.18 and choose

$$m \geq 4\varepsilon^{-2} \left(16\sqrt{2\pi} + (18 + t) r R \sqrt{2\log(2d)} \right)^2,$$

nothing but the probability changes to

$$1 - 8 \left(2\exp\left(\frac{-t^2 r^2 R^2 \log(2d)}{16} \right) + \exp\left(\frac{-t^2 \log(2d)}{16} \right) \right).$$

The two main ingredients for the proof of Theorem 5.18 are given by Theorem 5.20, which estimates the right hand side of (5.34) and is the main result of section 5.2.1, and Theorem 5.28, which estimates the left hand side of (5.34) and is the main result of section 5.2.2.

5.2.1 Estimate of the Right Hand Side of (5.34)

We want to show that

$$\sup_{w \in K} |f_x(w) - \mathbb{E}\, f_x(w)|$$

is small with high probability. Therefore, we will first estimate its mean

$$\mu := \mathbb{E} \left(\sup_{w \in K} |f_x(w) - \mathbb{E}\, f_x(w)| \right) \tag{5.37}$$

and afterwards use concentration inequalities to bound the probability

$$\mathbb{P}\left(\sup_{w\in K}|f_x(w)-\mathbb{E}\,f_x(w)|\geq\mu+t\right)$$

for $t>0$. This approach is inspired by the proof of [96, Theorem 1.1] and relies on standard techniques from [74, 75]. We obtain the following result (cf. [67, Theorem II.1]), whose proof is given at the end of this subsection.

Theorem 5.20 ([67])**.** *For any $u>0$, it holds*

$$\sup_{w\in K}|f_x(w)-\mathbb{E}\,f_x(w)|\leq\frac{16\sqrt{2\pi}+18rR\sqrt{2\log(2d)}}{\sqrt{m}}+u$$

with probability at least

$$1-8\left(2\exp\left(\frac{-mu^2}{32}\right)+\exp\left(\frac{-mu^2}{32r^2R^2}\right)\right).$$

1. Step: Estimate of the Mean μ

To estimate the mean μ we will make use of the following three lemmas, cf. [67, Lemma III.1-III.3].

Lemma 5.21 ([67])**.** *For independent Bernoulli variables $\xi_1,\dots,\xi_m$ and any $\lambda_1,\dots,\lambda_m\in\mathbb{R}$ and $t\in\mathbb{R}$, it holds*

$$\mathbb{P}\left(\sum_{i=1}^m\xi_i[\lambda_i]_+\geq t\right)\leq 2\,\mathbb{P}\left(\sum_{i=1}^m\xi_i\lambda_i\geq t\right).$$

Proof. Plugging in the definition of $[\cdot]_+$ we first observe

$$\mathbb{P}\left(\sum_{i=1}^m\xi_i[\lambda_i]_+\geq t\right)=\mathbb{P}\left(\sum_{\lambda_i\geq 0}\xi_i\lambda_i\geq t\right).$$

If every λ_i is positive, the claim follows trivially. So assume that $\lambda_j<0$ holds for at least one $j\in[m]$. It follows

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^m\xi_i[\lambda_i]_+\geq t\Big)\\
&=\mathbb{P}\Big(\sum_{\lambda_i\geq 0}\xi_i[\lambda_i]_+\geq t\Big)\cdot\Big[\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i\geq 0\Big)+\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i<0\Big)\Big]\\
&=\mathbb{P}\Big(\sum_{i=1}^m\xi_i[\lambda_i]_+\geq t\Big)\,\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i\geq 0\Big)+\mathbb{P}\Big(\sum_{i=1}^m\xi_i[\lambda_i]_+\geq t\Big)\,\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i<0\Big).
\end{aligned}$$

By symmetry of the Bernoulli variables ξ_i, it holds

$$\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i<0\Big)\leq\mathbb{P}\Big(\sum_{\lambda_i<0}\xi_i\lambda_i\geq 0\Big).$$

We finish the proof by

$$\mathbb{P}\left(\sum_{i=1}^m \xi_i [\lambda_i]_+ \geq t\right) \leq 2\,\mathbb{P}\left(\sum_{i=1}^m \xi_i [\lambda_i]_+ \geq t\right) \mathbb{P}\left(\sum_{\lambda_i<0} \xi_i \lambda_i \geq 0\right) \leq 2\,\mathbb{P}\left(\sum_{i=1}^m \xi_i \lambda_i \geq t\right),$$

where we used the independency of the ξ_i's for the last inequality. □

Lemma 5.22 ([67]). *For $a_1, \ldots, a_m \in \mathbb{R}^d$ and $K \subset \mathbb{R}^d$ according to* (5.31) *and* (5.33) *let*

$$\tilde{\mu} := \mathbb{E}\left(\sup_{w\in K} \left\langle \frac{1}{m}\sum_{i=1}^m a_i, w\right\rangle\right). \tag{5.38}$$

Then, for any $u > 0$, it holds

$$\mathbb{P}\left(\sup_{w\in K} \left\langle \frac{1}{m}\sum_{i=1}^m a_i, w\right\rangle \geq \tilde{\mu} + u\right) \leq \exp\left(\frac{-mu^2}{2r^2R^2}\right). \tag{5.39}$$

Proof. Using the 2-stability of Gaussian variables (3.17) we get

$$\tilde{\sigma}^2 := \sup_{w\in K} \mathbb{E}\left(\left\langle \frac{1}{m}\sum_{i=1}^m a_i, w\right\rangle^2\right) = \sup_{w\in K} \frac{r^2\|w\|_2^2}{m} \cdot \mathbb{E}\, g^2 = \frac{r^2R^2}{m} \tag{5.40}$$

for $g \sim \mathcal{N}(0,1)$. Applying Theorem 3.26 we end up with

$$\mathbb{P}\left(\sup_{w\in K} \left\langle \frac{1}{m}\sum_{i=1}^m a_i, w\right\rangle \geq \tilde{\mu} + u\right) \leq \exp\left(\frac{-u^2}{2\tilde{\sigma}^2}\right) = \exp\left(\frac{-mu^2}{2r^2R^2}\right)$$

as claimed. □

Lemma 5.23 ([67]). *Let the Standing Assumptions I be fulfilled and let $\xi_1, \ldots, \xi_m$ be i.i.d. Bernoulli variables, which are independent of the a_i's. For $\mu \in \mathbb{R}$ defined by* (5.37) *it then holds*

$$\mu \leq 2\,\mathbb{E}\sup_{w\in K} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \big[1 - y_i\langle a_i, w\rangle\big]_+\right|. \tag{5.41}$$

Proof. For brevity we set

$$\mathcal{A}_i(w) := \big[1 - y_i\langle a_i, w\rangle\big]_+, \quad \mathcal{A}_i'(w) := \big[1 - y_i'\langle a_i', w\rangle\big]_+,$$

where a_i' denote independent copies of a_i and $y_i' = \operatorname{sign}(\langle a_i', x\rangle)$. In the following, we will denote the expected value with respect to a_i' by $\mathbb{E}'$ and the expected value with respect to a_i by $\mathbb{E}$. Using $\mathbb{E}'\left(\mathcal{A}_i'(w) - \mathbb{E}'\,\mathcal{A}_i'(w)\right) = 0$ for $i = 1, \ldots, m$ we first observe

$$\begin{aligned}
\mu &= \mathbb{E}\sup_{w\in K} \left|\frac{1}{m}\sum_{i=1}^m \left(\mathcal{A}_i(w) - \mathbb{E}\,\mathcal{A}_i(w)\right)\right| \\
&= \mathbb{E}\sup_{w\in K} \left|\frac{1}{m}\sum_{i=1}^m \left(\mathcal{A}_i(w) - \mathbb{E}\,\mathcal{A}_i(w)\right) - \mathbb{E}'\left(\mathcal{A}_i'(w) - \mathbb{E}'\,\mathcal{A}_i'(w)\right)\right| \\
&= \mathbb{E}\sup_{w\in K} \left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}'\left(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\right)\right|.
\end{aligned}$$

Next we apply Jensen's inequality (cf. Theorem 3.17) to conclude

$$\mu = \mathbb{E} \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}'\left(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\right)\Big| \leq \mathbb{E}\,\mathbb{E}' \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \left(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\right)\Big|.$$

Since $\mathcal{A}_i'(w)$ is an independent copy of $\mathcal{A}_i(w)$, their difference $\mathcal{A}_i(w) - \mathcal{A}_i'(w)$ is equally likely positive or negative. Hence, we can multiply $\mathcal{A}_i(w) - \mathcal{A}_i'(w)$ by some Bernoulli variable ξ_i independent of a_i and a_i', without changing its distribution. It follows

$$\begin{aligned}
\mu &\leq \mathbb{E}\,\mathbb{E}' \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \left(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\right)\Big| = \mathbb{E}\,\mathbb{E}' \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\left(\mathcal{A}_i(w) - \mathcal{A}_i'(w)\right)\Big| \\
&\leq \mathbb{E}\,\mathbb{E}' \sup_{w\in K} \left(\Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\mathcal{A}_i(w)\Big| + \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\mathcal{A}_i'(w)\Big|\right) \leq 2\,\mathbb{E} \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\mathcal{A}_i(w)\Big| \\
&= 2\,\mathbb{E} \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\big[1 - y_i\langle a_i, w\rangle\big]_+\Big|,
\end{aligned}$$

as claimed. □

Equipped with the Lemmas 5.21, 5.22 and 5.23 we can now deduce the following estimate for μ, cf. [67, Lemma III.4].

Lemma 5.24 ([67])**.** *Let the Standing Assumptions I be fulfilled. Then it holds*

$$\mu = \mathbb{E} \sup_{w\in K} \left(f_x(w) - \mathbb{E}\, f_x(w)\right) \leq \frac{8\sqrt{2\pi} + 8rR\sqrt{2\log(2d)}}{\sqrt{m}}. \tag{5.42}$$

Proof. Let $\xi_1, \ldots, \xi_m$ be i.i.d. Bernoulli variables. Using Lemma 5.23 and the identity (3.8) for the expected value in terms of probabilities, we obtain

$$\begin{aligned}
\mu &= \mathbb{E} \sup_{w\in K} \left(f_x(w) - \mathbb{E}\, f_x(w)\right) \leq 2\,\mathbb{E} \sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\big[1 - y_i\langle a_i, w\rangle\big]_+\Big| \\
&= 2\int_0^\infty \mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\big[1 - y_i\langle a_i, w\rangle\big]_+\Big| \geq t\right) \mathrm{d}\,t.
\end{aligned}$$

Note that we cannot apply Lemma 5.21 directly to this expression. However, by using similar arguments as in the proof of Lemma 5.21, we obtain

$$\begin{aligned}
\mu &\leq 4\int_0^\infty \mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i(1 - y_i\langle a_i, w\rangle)\Big| \geq t\right) \mathrm{d}\,t \\
&\leq 4\int_0^\infty \mathbb{P}\left(\sup_{w\in K} \left(\Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\Big| + \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i y_i\langle a_i, w\rangle\Big|\right) \geq t\right) \mathrm{d}\,t \qquad (5.43) \\
&\leq 4\int_0^\infty \mathbb{P}\left(\Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\Big| \geq t/2\right) + \mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i y_i\langle a_i, w\rangle\Big| \geq t/2\right) \mathrm{d}\,t.
\end{aligned}$$

Using Corollary 3.22 the first summand can be further estimated by

$$\int_0^\infty \mathbb{P}\left(\Big|\frac{1}{m}\sum_{i=1}^{m} \xi_i\Big| \geq t/2\right) \mathrm{d}\,t \leq 2\int_0^\infty \exp\left(\frac{-mt^2}{8}\right) \mathrm{d}\,t = \frac{2\sqrt{2\pi}}{\sqrt{m}}. \tag{5.44}$$

To estimate the second summand, we observe that $\xi_i y_i$ is a Bernoulli variable independent of a_i, hence $\langle a_i, w\rangle$ and $\xi_i y_i \langle a_i, w\rangle$ are identically distributed. We get

$$\begin{aligned}
&\int_0^\infty \mathbb{P}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^m \xi_i y_i\langle a_i, w\rangle\Big| \geq t/2\right)\mathrm{d}\,t \\
&= 2\int_0^\infty \mathbb{P}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^m \langle a_i, w\rangle\Big| \geq t\right)\mathrm{d}\,t = 2\,\mathbb{E}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^m \langle a_i, w\rangle\Big|\right) \\
&= 2R\,\mathbb{E}\,\Big\|\frac{1}{m}\sum_{i=1}^m a_i\Big\|_\infty,
\end{aligned}$$

where we used the duality $\|\cdot\|_1' = \|\cdot\|_\infty$ for the last equality. Using again the 2-stability of Gaussian variables (3.16) we get $\frac{1}{\sqrt{m}}\sum_{i=1}^m a_i \sim \mathcal{N}(0, \mathrm{id})$ and estimate (3.19) yields

$$2R\,\mathbb{E}\,\Big\|\frac{1}{m}\sum_{i=1}^m a_i\Big\|_\infty \leq \frac{2R\sqrt{2\log(2d)}}{\sqrt{m}}. \tag{5.45}$$

Putting (5.43), (5.44) and (5.45) together, we get

$$\mu \leq 4\left(\frac{2\sqrt{2\pi}}{\sqrt{m}} + \frac{2R\sqrt{2\log(2d)}}{\sqrt{m}}\right) = \frac{8\sqrt{2\pi} + 8R\sqrt{2\log(2d)}}{\sqrt{m}},$$

as claimed. □

2. Step: Using Concentration Inequalities

In this step we want to show that $f_x(w)$ uniformly concentrates around its mean, i.e., we want to estimate the probability that $f_x(w)$ deviates anywhere on K far from its mean. We start with [67, Lemma III.5], which is a modified version of the second part of [96, Lemma 5.1], cf. also [75, Chapter 6.1].

Lemma 5.25 ([67])**.** *Let $\xi_1, \dots, \xi_m$ be i.i.d. Bernoulli variables and let the Standing Assumptions I be fulfilled. Then, for $\mu \in \mathbb{R}$ defined by* (5.37) *and any $t > 0$, it holds*

$$\mathbb{P}\left(\sup_{w\in K}|f_x(w) - \mathbb{E}\,f_x(w)| \geq 2\mu + t\right) \leq 4\,\mathbb{P}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^m \xi_i\big[1 - y_i\langle a_i, w\rangle\big]_+\Big| \geq t/2\right). \tag{5.46}$$

Proof. Using Markov's inequality 3.16, let us first note

$$\begin{aligned}
\frac{1}{2} &= 1 - \frac{\mathbb{E}\sup_{w\in K}|f_x(w) - \mathbb{E}\,f_x(w)|}{2\mu} \leq 1 - \mathbb{P}\left(\sup_{w\in K}|f_x(w) - \mathbb{E}\,f_x(w)| \geq 2\mu\right) \\
&= \mathbb{P}\Big(|f_x(w) - \mathbb{E}\,f_x(w)| < 2\mu \text{ for all } w \in K\Big),
\end{aligned}$$

which yields

$$\begin{aligned}
&\frac{1}{2}\,\mathbb{P}\left(\sup_{w\in K}|f_x(w) - \mathbb{E}\,f_x(w)| \geq 2\mu + t\right) \\
&\leq \mathbb{P}\Big(|f_x(w) - \mathbb{E}\,f_x(w)| < 2\mu \text{ for all } w \in K\Big)\cdot \mathbb{P}\left(\sup_{w\in K}|f_x(w) - \mathbb{E}\,f_x(w)| \geq 2\mu + t\right)
\end{aligned}$$

for any $t \geq 0$. Next let a_i', y_i' and f_x' be independent copies of a_i, y_i and f_x, respectively. It follows

$$\begin{aligned}
&\mathbb{P}\left(|f_x(w) - \mathbb{E}\, f_x(w)| < 2\mu \text{ for all } w \in K\right) \cdot \mathbb{P}\left(\sup_{w\in K} |f_x'(w) - \mathbb{E}'\, f_x'(w)| \geq 2\mu + t\right) \\
&\leq \mathbb{P}\left(\sup_{w\in K} \left|(f_x(w) - \mathbb{E}\, f_x(w)) - (f_x'(w) - \mathbb{E}'\, f_x'(w))\right| \geq t\right) \\
&= \mathbb{P}\left(\sup_{w\in K} |f_x(w) - f_x'(w)| \geq t\right)
\end{aligned}$$

and since $[1 - y_i\langle a_i, w\rangle]_+ - [1 - y_i'\langle a_i', w\rangle]_+$ is equally likely positive or negative, we can multiply it with some Bernoulli variable ξ_i without changing its distribution and we end up with

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in K} |f_x(w) - \mathbb{E}\, f_x(w)| \geq 2\mu + t\right) \leq 2\,\mathbb{P}\left(\sup_{w\in K} |f_x(w) - f_x'(w)| \geq t\right) \\
&= 2\,\mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^m \xi_i\big([1 - y_i\langle a_i, w\rangle]_+ - [1 - y_i'\langle a_i', w\rangle]_+\big)\Big| \geq t\right) \\
&\leq 4\,\mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^m \xi_i[1 - y_i\langle a_i, w\rangle]_+\Big| \geq t/2\right)
\end{aligned}$$

as claimed. □

Combining Lemma 5.21 with the previous Lemma 5.25 we deduce the following result on the uniform concentration of f_x around its mean, cf. [67, Lemma III.6].

Lemma 5.26 ([67])**.** *Let the Standing Assumptions I be fulfilled and let μ and $\tilde{\mu}$ be defined by* (5.37) *and* (5.38)*. Then, for any $u > 0$, it holds*

$$\mathbb{P}\left(\sup_{w\in K} |f_x(w) - \mathbb{E}\, f_x(w)| \geq 2\mu + 2\tilde{\mu} + u\right) \leq 8\left(2\exp\left(\frac{-mu^2}{32}\right) + \exp\left(\frac{-mu^2}{32r^2R^2}\right)\right). \tag{5.47}$$

Proof. Let $\xi_1, \ldots, \xi_m$ be i.i.d. Bernoulli variables. Applying first Lemma 5.25 with $t = 2\tilde{\mu} + u$ and afterwards using Lemma 5.21 we get

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in K} |f_x(w) - \mathbb{E}\, f_x(w)| \geq 2\mu + 2\tilde{\mu} + u\right) \\
&\leq 4\,\mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^m \xi_i[1 - y_i\langle a_i, w\rangle]_+\Big| \geq \tilde{\mu} + u/2\right) \\
&\leq 8\,\mathbb{P}\left(\sup_{w\in K} \Big|\frac{1}{m}\sum_{i=1}^m \xi_i(1 - y_i\langle a_i, w\rangle)\Big| \geq \tilde{\mu} + u/2\right).
\end{aligned}$$

Using Corollary 3.22 and Lemma 5.22 we end up with

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^{m}\xi_i(1-y_i\langle a_i,w\rangle)\Big|\geq\tilde{\mu}+u/2\right)\\
&\leq\mathbb{P}\left(\Big|\frac{1}{m}\sum_{i=1}^{m}\xi_i\Big|\geq u/4\right)+\mathbb{P}\left(\sup_{w\in K}\Big|\frac{1}{m}\sum_{i=1}^{m}\langle a_i,w\rangle\Big|\geq\tilde{\mu}+u/4\right)\\
&\leq 2\exp\left(\frac{-mu^2}{32}\right)+\exp\left(\frac{-mu^2}{32r^2R^2}\right)
\end{aligned}$$

which finishes the proof. □

Combining the two results of Lemma 5.24 and Lemma 5.26 we can finally give a proof of Theorem 5.20:

Proof of Theorem 5.20. Using Lemma 5.26 it holds

$$\mathbb{P}\left(\sup_{w\in K}|f_x(w)-\mathbb{E}\,f_x(w)|\geq 2\mu+2\tilde{\mu}+u\right)\leq 8\left(2\exp\left(\frac{-mu^2}{32}\right)+\exp\left(\frac{-mu^2}{32r^2R^2}\right)\right),$$

so it remains to estimate μ and $\tilde{\mu}$.

By invoking the duality $\|\cdot\|_1'=\|\cdot\|_\infty$ from (3.19), which estimates the ℓ_∞-norm of a Gaussian vector, we obtain

$$\tilde{\mu}=\mathbb{E}\left(\sup_{w\in K}\Big\langle\frac{1}{m}\sum_{i=1}^{m}a_i,w\Big\rangle\right)=R\,\mathbb{E}\Big\|\frac{1}{m}\sum_{i=1}^{m}a_i\Big\|_\infty\leq\frac{rR\sqrt{2\log(2d)}}{\sqrt{m}}.$$

Furthermore, Lemma 5.24 yields

$$\mu\leq\frac{8\sqrt{2\pi}+8rR\sqrt{2\log(2d)}}{\sqrt{m}},$$

which finishes the proof. □

5.2.2 Estimate of the Left Hand Side of (5.34)

The aim of this section is to give a lower bound for the expected value

$$\mathbb{E}\left(f_x(w)-f_x(x)\right)=\mathbb{E}\left[1-y\langle a,w\rangle\right]_+-\mathbb{E}\left[1-y\langle a,x\rangle\right]_+ \tag{5.48}$$

for $w\in\mathbb{R}^d\backslash\{0\}$ with $\|w\|_1\leq R$. We will first estimate both expected values of the right hand side of (5.48) separately and combine them later for an estimate of their difference. We start with the following characterization of the expected values, cf. [67, Lemma III.11]:

Lemma 5.27 ([67])**.** *Let $x\in\mathbb{R}^d$ and $f_x\colon\mathbb{R}^d\to\mathbb{R}$ be according to* (5.32) *and* (5.19)*. For $w\in\mathbb{R}^d\backslash\{0\}$, we set $c=\langle x,w\rangle$ and $c'=\sqrt{\|w\|_2^2-c^2}$. Then it holds*

$$i)\qquad \mathbb{E}\,f_x(x)=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\left[1-r|t|\right]_+e^{-t^2/2}\,\mathrm{d}\,t,$$

$$ii)\qquad \mathbb{E}\,f_x(w)=\frac{1}{2\pi}\int_{\mathbb{R}^2}\left[1-cr|t_1|-c'rt_2\right]_+\exp\left(\frac{-t_1^2-t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2.$$

Proof. i) Since $\|x\|_2 = 1$ for any $a = r\tilde{a}$ with $\tilde{a} \sim \mathcal{N}(0, \mathrm{id})$ it holds $\frac{1}{r}\langle x, a\rangle \sim \mathcal{N}(0,1)$. Using the density function (3.14) of Gaussian variables and the identity (3.10) we get

$$\mathbb{E}\, f_x(x) = \mathbb{E}\left[1 - |\langle a, x\rangle|\right]_+ = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \left[1 - r|t|\right]_+ e^{\frac{-t^2}{2}}\, \mathrm{d}\, t,$$

as claimed.

ii) If $c' = 0$ the claim follows from part i), so let us assume that it holds $c' \neq 0$. For some Gaussian variable $a = r\tilde{a} \in \mathbb{R}^d$ for some $r > 0$ and $\tilde{a} \sim \mathcal{N}(0, \mathrm{id})$ we set

$$g := \frac{1}{c'}(\langle a, w\rangle - c\langle a, x\rangle) = \frac{1}{c'} \sum_{i=1}^{d} a_i(w_i - cx_i).$$

From the 2-stability of Gaussian variables (3.16) we observe that g is normally distributed with $\mathbb{E}\, g = 0$. To calculate the variance of g we first calculate the covariance of $\langle a, x\rangle$ and $\langle a, w\rangle$ by

$$\begin{aligned}\mathrm{Cov}\left(\langle a, x\rangle, \langle a, w\rangle\right) = \mathbb{E}\left(\langle a, x\rangle\langle a, w\rangle\right) &= \mathbb{E}\left(\sum_{i=1}^{d} a_i x_i \cdot \sum_{j=1}^{d} w_j x_j\right) \\ &= \sum_{i,j=1}^{d} x_i w_j\, \mathbb{E}(a_i a_j) = r^2\langle x, w\rangle,\end{aligned}$$

where we used $\mathbb{E}(a_i a_j) = 0$ if $i \neq j$ and $\mathbb{E}(a_i^2) = r^2$. With $\mathbb{E}\langle a, x\rangle^2 = r^2$ and $\mathbb{E}\langle a, w\rangle^2 = r^2\|w\|_2^2$ we arrive at

$$\begin{aligned}\mathrm{Var}(g) = \mathbb{E}\, g^2 = \frac{\mathbb{E}\left(\langle a, w\rangle - c\langle a, x\rangle\right)^2}{c'^2} &= \frac{\mathbb{E}\left(\langle a, w\rangle^2 - 2c\langle a, w\rangle\langle a, x\rangle + c^2\langle a, x\rangle^2\right)}{c'^2} \\ = \frac{r^2\|w\|_2^2 - 2r^2c^2 + r^2c^2}{c'^2} &= r^2.\end{aligned}$$

With our choice of c it holds $\langle x, w - cx\rangle = 0$, so the second part of Lemma 3.20 yields that $\langle a, x\rangle$ and g are independent. And since Gaussian variables are symmetric, g and $y \cdot g$ are identically distributed. Using (3.11), which describes the joint density of a function of independent random variables, we end up with

$$\begin{aligned}\mathbb{E}\, f_x(w) = \mathbb{E}\left[1 - y\langle a, w\rangle\right]_+ &= \mathbb{E}\left[1 - c|\langle a, x\rangle| - c'g\right]_+ \\ &= \frac{1}{2\pi} \int_{\mathbb{R}^2} \left[1 - cr|t_1| - c't_2\right]_+ \exp\left(\frac{-t_1^2 - t_2^2}{2}\right) \mathrm{d}\, t_1\, \mathrm{d}\, t_2.\end{aligned}$$

□

Using Lemma 5.27 we can now deduce the following result, cf. [67, Theorem II.2]:

Theorem 5.28 ([67]). *For $w \in K$ let $c = \langle x, w\rangle$ and $c' = \sqrt{\|w\|_2^2 - c^2}$. Assume that $c' > 0$ holds. If $c \leq 0$, it then follows*

$$\pi\,\mathbb{E}\left(f_x(w) - f_x(x)\right) \geq \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r}$$

and if $c > 0$, it holds

$$\pi\,\mathbb{E}\left(f_x(w) - f_x(x)\right) \geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1-crt)e^{-t^2/2}\,\mathrm{d}\,t + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r}.$$

Remark 5.29. From the Cauchy-Schwartz inequality it follows $\|w\|_2^2 - c^2 \geq 0$, hence c' is well defined.

Proof. Using the first part of Lemma 5.27 we get

$$\begin{aligned}
-\pi\,\mathbb{E}\,f_x(x) &= \frac{-\sqrt{\pi}}{\sqrt{2}}\int_{\mathbb{R}}\left[1-r|t|\right]_+ e^{-t^2/2}\,\mathrm{d}\,t = -\sqrt{2\pi}\int_0^{1/r}(1-rt)e^{-t^2/2}\,\mathrm{d}\,t \\
&\geq -\sqrt{2\pi}\int_0^{1/r} e^{-t^2/2}\,\mathrm{d}\,t \geq -\sqrt{2\pi}\int_0^{1/r} 1\,\mathrm{d}\,t \qquad (5.49)\\
&= -\frac{\sqrt{2\pi}}{r}.
\end{aligned}$$

It remains to give a lower bound for the expected value

$$\begin{aligned}
\pi\,\mathbb{E}\,f_x(w) &= \frac{1}{2}\int_{\mathbb{R}^2}\left[1-cr|t_1|-c'rt_2\right]_+\exp\left(\frac{-t_1^2-t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2 \\
&= \int_{\mathbb{R}}\int_0^{\infty}\left[1-crt_1-c'rt_2\right]_+\exp\left(\frac{-t_1^2-t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2
\end{aligned}$$

for $w \in \mathbb{R}^d\backslash\{0\}$. We distinguish the two cases $c = \langle x, w\rangle \leq 0$ and $c > 0$.
1. Case: $c \leq 0$. It holds $-crt_1 \geq 0$ for $0 \leq t_1$, so we get

$$\begin{aligned}
\pi\,\mathbb{E}\,f_x(w) &= \int_{\mathbb{R}}\int_0^{\infty}\left[1-crt_1-c'rt_2\right]_+\exp\left(\frac{-t_1^2-t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2 \\
&\geq \int_{\mathbb{R}}\int_0^{\infty}\left[1-c'rt_2\right]_+\exp\left(\frac{-t_1^2-t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2 \qquad (5.50)\\
&= \frac{\sqrt{\pi}}{\sqrt{2}}\int_{\mathbb{R}}\left[1-c'rt_2\right]_+ e^{-t_2^2/2}\,\mathrm{d}\,t_2 \geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_{-\infty}^{0}(1-c'rt_2)e^{-t_2^2/2}\,\mathrm{d}\,t_2 \\
&= \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}}.
\end{aligned}$$

As claimed, putting both estimates (5.49) and (5.50) together, we get

$$\pi\,\mathbb{E}\left(f_x(w) - f_x(x)\right) \geq \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r}.$$

2. Case: $c > 0$. In that case it holds $1 - crt_1 - c'rt_2 \geq 0$ for $(t_1, t_2) \in [0, \frac{1}{cr}] \times (-\infty, 0]$, hence

$$\begin{aligned}
\pi \, \mathbb{E} \, f_x(w) &= \int_{\mathbb{R}} \int_0^\infty \left[1 - crt_1 - c'rt_2\right]_+ \exp\left(\frac{-t_1^2 - t_2^2}{2}\right) \mathrm{d}\, t_1 \, \mathrm{d}\, t_2 \\
&\geq \int_0^{\frac{1}{cr}} \int_{-\infty}^0 (1 - crt_1 - c'rt_2) \exp\left(\frac{-t_1^2 - t_2^2}{2}\right) \mathrm{d}\, t_2 \, \mathrm{d}\, t_1 \\
&= \frac{\sqrt{\pi}}{\sqrt{2}} \int_0^{\frac{1}{cr}} (1 - crt_1) e^{-t_1^2/2} \, \mathrm{d}\, t_1 + c'r \int_0^{\frac{1}{cr}} e^{-t_1^2/2} \, \mathrm{d}\, t_1 \\
&\geq \frac{\sqrt{\pi}}{\sqrt{2}} \int_0^{\frac{1}{cr}} (1 - crt) e^{-t^2/2} \, \mathrm{d}\, t + \frac{c'}{c} \exp\left(\frac{-1}{2c^2r^2}\right)
\end{aligned}$$

which, combined with (5.49), yields the claim. □

5.2.3 Proof of Theorem 5.18

First we intend to apply Theorem 5.20 to bound the right hand side of (5.34). By choosing

$$u = \frac{rR\sqrt{2\log(2d)}}{\sqrt{m}} \quad \text{and} \quad m \geq 4\varepsilon^{-2}\Big(16\sqrt{2\pi} + 19rR\sqrt{2\log(2d)}\Big)^2$$

we obtain the estimate

$$\sup_{w \in K} |f_x(w) - \mathbb{E}\, f_x(w)| \leq \frac{8\sqrt{8\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u \leq \frac{\varepsilon}{2}$$

with probability at least

$$1 - 8\left(2\exp\left(\frac{-r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-\log(2d)}{16}\right)\right). \tag{5.51}$$

Using (5.34) this already implies

$$\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) \leq \varepsilon \tag{5.52}$$

with at least the same probability (5.51). Next we want to apply Theorem 5.28 with $w = \hat{x}$, but for this we first have to show $c' = \sqrt{\|\hat{x}\|_2^2 - \langle x, \hat{x}\rangle} \neq 0$. In order to get a contradiction, let us assume $c' = 0$, which only holds if $\hat{x} = \lambda x$ for some $\lambda \in \mathbb{R}$. If $\lambda > 0$, we get $\hat{x}/\|\hat{x}\|_2 = x$ and the claim of the theorem follows trivially. And if $\lambda \leq 0$, we use $f_x(x) \geq f_x(\hat{x})$ to get

$$f_x(x) = \sum_{i=1}^m \left[1 - |\langle a_i, x\rangle|\right]_+ \geq \sum_{i=1}^m \left[1 + |\lambda\langle a_i, x\rangle|\right]_+ = f_x(\hat{x}),$$

which implies $\langle a_i, x\rangle = 0$ for $i = 1, \ldots, m$. Since this only happens with probability zero, it almost surely holds $c' > 0$ and we can apply Theorem 5.28.

Here we distinguish three cases $c = \langle x, \hat{x}\rangle \leq 0$, $0 < c < 1/r$ and $c > 1/r$. First we will show that the first two cases lead to a contradiction and afterwards we will

use the third case to prove our claim.

1. Case: $c \leq 0$. With our choices of ε and r such that $0 < \varepsilon < 0.18$, $r > \sqrt{2\pi}(0.57 - \varepsilon\pi)^{-1}$, we combine Theorem 5.28 and the upper bound (5.52) to end up with the contradiction

$$\frac{\pi}{2} - \frac{\sqrt{2\pi}}{r} \leq \frac{\pi}{2} + c'r\frac{\sqrt{\pi}}{\sqrt{2}} - \frac{\sqrt{2\pi}}{r} \leq \pi\,\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) \leq \pi\varepsilon.$$

2. Case: $0 < c < 1/r$. Again we want to use the upper bound (5.52) to end up with a contradiction. First, Theorem 5.28 yields the estimate

$$\begin{aligned}\pi\,\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1 - crt)e^{-t^2/2}\,\mathrm{d}\,t + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r}\\ &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1 - crt)e^{-t^2/2}\,\mathrm{d}\,t - \frac{\sqrt{2\pi}}{r}.\end{aligned}$$

To estimate this integral we consider the function

$$g\colon (0,\infty) \to \mathbb{R}, \quad z \mapsto \int_0^{1/z}(1 - zt)e^{-t^2/2}\,\mathrm{d}\,t.$$

Since

$$g'(z) = -\int_0^{1/z} te^{-t^2/2}\,\mathrm{d}\,t < 0,$$

the function g is monotonicly decreasing, yielding $g(cr) > g(1)$. Again, with our choices of ε and r we end up with the contradiction

$$\begin{aligned}\pi\varepsilon \geq \pi\,\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1 - crt)e^{-t^2/2}\,\mathrm{d}\,t - \frac{\sqrt{2\pi}}{r}\\ &= \frac{\sqrt{\pi}}{\sqrt{2}}g(cr) - \frac{\sqrt{2\pi}}{r} \geq \frac{\sqrt{\pi}}{\sqrt{2}}g(1) - \frac{\sqrt{2\pi}}{r} = \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^1(1-t)e^{-t^2/2}\,\mathrm{d}\,t - \frac{\sqrt{2\pi}}{r}\\ &\geq 0.57 - \frac{\sqrt{2\pi}}{r}.\end{aligned}$$

3. Case: $c > 1/r$. In this case Theorem 5.28 gives the estimate

$$\begin{aligned}\pi\,\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) &\geq \frac{\sqrt{\pi}}{\sqrt{2}}\int_0^{\frac{1}{cr}}(1 - crt)e^{-t^2/2}\,\mathrm{d}\,t + \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r}\\ &\geq \frac{c'}{c}\exp\left(\frac{-1}{2c^2r^2}\right) - \frac{\sqrt{2\pi}}{r} \geq \frac{c'}{c}e^{-1/2} - \frac{\sqrt{2\pi}}{r},\end{aligned} \tag{5.53}$$

where we used $cr > 1$, hence $\exp(-1/(2c^2r^2)) \geq e^{-1/2}$ for the last inequality. Further, plugging in the definitions of c' and c we get

$$\begin{aligned}\frac{c'}{c} &= \frac{\sqrt{\|\hat{x}\|_2^2 - \langle x,\hat{x}\rangle^2}}{\langle x,\hat{x}\rangle} = \sqrt{\frac{\Big(\|\hat{x}\|_2 - \langle x,\hat{x}\rangle\Big)\Big(\|\hat{x}\|_2 + \langle x,\hat{x}\rangle\Big)}{\langle x,\hat{x}\rangle^2}}\\ &= \sqrt{\frac{\Big(1 - \langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle\Big)\cdot\Big(1 + \langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle\Big)}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle^2}}.\end{aligned}$$

Using the assumption $\|x\|_2 = 1$, we simplify this expression further by

$$\frac{c'}{c} = \sqrt{\frac{\left(\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\left\|\frac{\hat{x}}{\|\hat{x}\|_2}\right\|^2 - \langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle\right) \cdot \left(\frac{1}{2}\|x\|_2^2 + \frac{1}{2}\left\|\frac{\hat{x}}{\|\hat{x}\|_2}\right\|^2 + \langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle\right)}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle^2}}$$

$$= \frac{1}{2}\sqrt{\frac{\left\|x - \frac{\hat{x}}{\|\hat{x}\|}\right\|_2^2 \cdot \left\|x + \frac{\hat{x}}{\|\hat{x}\|}\right\|_2^2}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle^2}} \geq \frac{1}{2}\frac{\left\|x - \frac{\hat{x}}{\|\hat{x}\|_2}\right\|_2}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle}, \tag{5.54}$$

where we used $\langle x, \hat{x}\rangle \geq 0$ for the last inequality. Combining (5.52), (5.53) and (5.54), we finally arrive at

$$\frac{1}{\pi}\left(\frac{1}{2}\frac{\left\|x - \frac{\hat{x}}{\|\hat{x}\|_2}\right\|_2}{\langle x, \frac{\hat{x}}{\|\hat{x}\|_2}\rangle}e^{-1/2} - \frac{\sqrt{2\pi}}{r}\right) \leq \mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) \leq \varepsilon,$$

which finishes the proof. □

5.3 Recovery with $\ell_{1,2}$-Support Vector Machines

By construction, the ℓ_1-SVM tends to pick a sparse classifier. In the presence of highly correlated variables a_i, this may lead to the following issue. The ℓ_1-SVM tends to pick only few support vectors, i.e., only few of the a_i's and removes the rest. But this selection is quite unstable and, in particular, heavily depends on the actual measurement vectors.

Hence, in order to get a stable selection, one prefers to select or remove highly correlated variables together. In the theory of elastic nets a combination of ℓ_1- and ℓ_2-constraints has been proposed [11, 54, 117], which already has been successfully used in gene selection and microarray analysis [71, 115, 117]. Support vector machines which combine both the ℓ_1- and the ℓ_2-penalty are called *doubly regularized* SVMs [112].

However, our motivation for introducing an additional ℓ_2-penalty is rather different, since we assume the sample points $\tilde{a}_i \sim \mathcal{N}(0, \mathrm{id})$ to be independent. But a detailed inspection of our analysis so far shows that it would be convenient if the convex body $K = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R\}$ would not include vectors having a large ℓ_2-norm. For instance, in (5.40), we used $R^2 = \sup_{w \in K} \|w\|_2^2$ to prove Lemma 5.22, although the set of vectors in K with large ℓ_2-norm is very small.

Therefore, we suggest to modify the ℓ_1-SVM (5.30) by introducing an additional ℓ_2-constraint

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{m} \left[1 - y_i \langle a_i, w\rangle\right]_+ \quad \text{subject to} \quad \|w\|_2 \leq 1,\ \|w\|_1 \leq R, \tag{5.55}$$

which we will denote as $\ell_{1,2}$-SVM. It turns out that for our data model this approach in some sense will outperform the ℓ_1-SVM, which we will also discuss in the numerical examples. Another motivation for the additional ℓ_2-constraint is that the true signal $x \in \mathbb{R}^d$ is contained in the set

$$\tilde{K} := \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1,\ \|w\|_1 \leq R\}, \tag{5.56}$$

which also plays a crucial role in the analysis of [96], where the algorithm (5.2) was proposed. Before going on with the analysis let us fix the standing assumptions of this section.

Standing Assumptions II

i) The true classifier $x \in \mathbb{R}^d$, which we want to approximate, is compressible in the way that $\|x\|_2 = 1$ and $\|x\|_1 \leq R$ for some $R > 0$.

ii) For a scaling parameter $r > 0$ we take the measurement vectors $a_i = r \cdot \tilde{a}_i \in \mathbb{R}^d$ for some i.i.d. $\tilde{a}_i \sim \mathcal{N}(0, \mathrm{id})$.

iii) The measurements are given by $y_i = \mathrm{sign}(\langle a_i, x \rangle)$.

iv) We denote $f_x(w) = \sum_{i=1}^m [1 - y_i \langle a_i, w \rangle]_+$ and $\tilde{K} = \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1,\ \|w\|_1 \leq R\}$.

v) $\hat{x}$ denotes a minimizer of the $\ell_{1,2}$-SVM, i.e., $\hat{x} = \arg\min_{w \in \tilde{K}} f_x(w)$.

Let $\hat{x}$ denote a minimizer of the $\ell_{1,2}$-SVM (5.55). As seen before in (5.34) we observe

$$\mathbb{E}\left(f_x(\hat{x}) - f_x(x)\right) \leq 2 \sup_{w \in \tilde{K}} |f_x(w) - f_x(x)|, \tag{5.57}$$

and since $\tilde{K} \subset K$, the estimate of the right hand side from Theorem 5.20 remains true if we replace K by $\tilde{K}$, i.e., it holds

$$\sup_{w \in \tilde{K}} |f_x(w) - f_x(x)| \leq \frac{16\sqrt{2\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} \tag{5.58}$$

with high probability. In order to formulate an analogue of Theorem 5.18 it remains to estimate the expected value $\mathbb{E}(f_x(w) - f_x(x))$ for some $w \in \mathbb{R}^d$. We obtain the following theorem:

Lemma 5.30 ([67])**.** *Let the Standing Assumptions II be fulfilled. For any $w \in \tilde{K}$, it then holds*

$$\mathbb{E}\left(f_x(w) - f_x(x)\right) \geq \frac{r\|x - w\|_2^2}{\sqrt{2\pi}} \left(1 - \exp\left(\frac{-1}{2r^2}\right)\right).$$

Proof. For $w \in \tilde{K}$ we set $c = \langle x, w \rangle$ and $c' = \sqrt{\|w\|_2^2 - c^2}$. Using Lemma 5.27 we

then obtain

$$\begin{aligned}
&\mathbb{E}\left(f_x(w) - f_x(x)\right) \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^2}\left(\left[1 - cr|t_1| - c'rt_2\right]_+ - \left[1 - r|t_1|\right]_+\right)\exp\left(\frac{-t_1^2 - t_2^2}{2}\right)\mathrm{d}\,t_1\,\mathrm{d}\,t_2 \\
&\geq \frac{1}{\pi}\int_0^{1/r}\int_{\mathbb{R}}\left((1 - crt_1 - c'rt_2) - (1 - rt_1)\right)\exp\left(\frac{-t_1^2 - t_2^2}{2}\right)\mathrm{d}\,t_2\,\mathrm{d}\,t_1 \\
&= r(1-c)\frac{\sqrt{2}}{\sqrt{\pi}}\int_0^{1/r} te^{-t^2/2}\,\mathrm{d}\,t = r(1-c)\frac{\sqrt{2}}{\sqrt{\pi}}\left(1 - \exp\left(\frac{-1}{2r^2}\right)\right).
\end{aligned}$$

The claim now follows from $\|w\|_2 \leq 1$ and

$$1 - c \geq \frac{1}{2}\left(\|x\|_2^2 + \|w\|_2^2 - 2\langle x, w\rangle\right) = \frac{\|x - w\|_2^2}{2}.$$

□

Combining (5.57) and (5.58) with the previous Lemma 5.30, we obtain the following main result of this section, cf. [67, Theorem IV.1].

Theorem 5.31 ([67])**.** *Let $d \geq 2$, $0 < \varepsilon < 1/2$, $r > \sqrt{2\pi}(1-2\varepsilon)^{-1}$ and*

$$m \geq C\varepsilon^{-2}r^2R^2\log(d)$$

for a constant $C > 0$ and let the Standing Assumptions II be fulfilled. Then it holds

$$\|x - \hat{x}\|_2^2 \leq \frac{C'\varepsilon}{r\left(1 - \exp\left(\frac{-1}{2r^2}\right)\right)}$$

with probability at least $1 - \gamma\exp(-C''R^2\log(d))$ for some positive constants γ, C', C''.

Remark 5.32. i) As in Theorem 5.18 we use the constants γ, C, C', C'' only for simplicity. More explicit, taking

$$m \geq \frac{4\left(16\sqrt{2\pi} + (18+t)rR\sqrt{2\log(2d)}\right)^2}{\varepsilon^2}$$

for some $t > 0$, we get

$$\|x - \hat{x}\|_2^2 \leq \frac{\varepsilon\sqrt{\pi/2}}{r - \left(1 - \exp\left(\frac{-1}{2r^2}\right)\right)}$$

with probability at least

$$1 - 8\left(2\exp\left(\frac{-t^2r^2R^2\log(2d)}{16}\right) + \exp\left(\frac{-t^2R^2\log(2d)}{16}\right)\right).$$

ii) The main advantage of Theorem 5.31 compared to Theorem 5.18 is the dependency on r. In Theorem 5.18 we needed r to grow to infinity for the error $\|x - \hat{x}/\|\hat{x}\|_2\|_2$ to go down to zero. In Theorem 5.31 we instead can choose r to be fixed. Indeed, if we for example take $\varepsilon < 0.2$, we can choose $r = 10$ and $m \geq \tilde{C}\varepsilon^{-2}R^2\log(d)$ and obtain

$$\|x - \hat{x}\|_2^2 \leq \tilde{C}'\varepsilon$$

with high probability. In that sense, the $\ell_{1,2}$-SVM outperforms the ℓ_1-SVM which we will also demonstrate in the numerical examples.

5.4 Recovery from Noisy Measurements

In practice, we usually do not have access to exact measurements, since they are disturbed by different kinds of noise such as rounding errors, limitations of the measurement devices, human failures, etc.

In the setting of 1-bit compressed sensing, each sample point a_i gets classified by its corresponding label $y_i = \operatorname{sign}(\langle a_i, x\rangle)$, so the measurements only take the values ± 1. Hence, a reasonable model for noisy measurements is given by

$$\hat{y}_i = z_i \cdot \operatorname{sign}\left(\langle a_i, x\rangle\right) = z_i \cdot y_i \tag{5.59}$$

for some random variables z_i, which takes the value $z_i = -1$ if the sample point a_i gets wrongly classified and it holds $z_i = +1$ in the case of correct classification. The only parameters left over are the probabilities of the events $z_i = +1$ and $z_i = -1$, which even may or may not depend on the true signal x and the particular instances of a_i, y_i.

Before analyzing the performance of the ℓ_1-SVM with noisy measurements, to avoid confusion let us summarize the assumptions.

Standing Assumptions III

i) The true classifier $x \in \mathbb{R}^d$, which we want to approximate, is compressible in the way that $\|x\|_2 = 1$ and $\|x\|_1 \leq R$ for some $R > 0$.

ii) For a scaling parameter $r > 0$ we take the measurement vectors $a_i = r \cdot \tilde{a}_i \in \mathbb{R}^d$ for some i.i.d. $\tilde{a}_i \sim \mathcal{N}(0, \mathrm{id})$.

iii) The measurements are given by $\hat{y}_i = z_i y_i = \operatorname{sign}(\langle a_i, x\rangle)$ for some random variable $z_i \in \{\pm 1\}$.

iv) We denote $\hat{f}_x(w) = \sum_{i=1}^m [1 - \hat{y}_i \langle a_i, w\rangle]_+$ and $K = \{w \in \mathbb{R}^d \mid \|w\|_1 \leq R\}$.

v) $\hat{x}$ denotes a minimizer of the ℓ_1-SVM, i.e., $\hat{x} = \arg\min_{w \in K} \hat{f}_x(w)$.

We follow the idea of the noiseless case, so we obtain the estimate

$$\mathbb{E}\left(\hat{f}_x(\hat{x}) - \hat{f}_x(x)\right) \leq 2 \sup_{w \in K} |\hat{f}_x(w) - \mathbb{E}\, \hat{f}_x(w)|. \tag{5.60}$$

It remains to bound the right hand side from above and to find a lower estimate of the left hand side by terms of the difference between x and $\hat{x}$.

For the noiseless case in section 5.2.1, we have multiplied each y_i with an independent Bernoulli variables ξ_i in order to obtain an estimate of the right hand side of (5.34). Hence, replacing y_i by $\hat{y}_i = z_i y_i$ does not affect the analysis and the result of Theorem 5.20, since $\xi_i y_i$ and $\xi_i z_i y_i$ are identically distributed.

Corollary 5.33. *For any $u > 0$, it holds*

$$\sup_{w \in K} |\hat{f}_x(w) - \mathbb{E}\, \hat{f}_x(w)| \leq \frac{16\sqrt{2\pi} + 18rR\sqrt{2\log(2d)}}{\sqrt{m}} + u$$

with probability at least

$$1 - 8\left(2\exp\left(\frac{-mu^2}{32}\right) + \exp\left(\frac{-mu^2}{32r^2R^2}\right)\right).$$

In order to get a similar recovery result as Theorem 5.28 for the noiseless case, it remains to bound the left hand side of (5.60). Here we have to explicitly use the underlying measurement rule, i.e., we have to take the probabilities $\mathbb{P}(z_i = +1)$ and $\mathbb{P}(z_i = -1)$ into account.

Let a, y, z be independent copies of the a_i's, y_i's and z_i's. By conditioning the expected value of $\hat{f}_x(w)$ on a we obtain

$$\begin{aligned}\mathbb{E}\, \hat{f}_x(w) &= \mathbb{E}\left[1 - zy\langle a, w\rangle\right]_+ = \mathbb{E}_a\left(\mathbb{E}_z\left[1 - zy\langle a, w\rangle\right]_+ \mid a\right) \\ &= \mathbb{E}\left(\left[1 - y\langle a, w\rangle\right]_+ \cdot \mathbb{P}(z = 1 \mid a)\right) + \mathbb{E}\left(\left[1 + y\langle a, w\rangle\right]_+ \cdot \mathbb{P}(z = -1 \mid a)\right)\end{aligned} \tag{5.61}$$

and, analogously,

$$\begin{aligned}\mathbb{E}\, \hat{f}_x(x) &= \mathbb{E}\left(\left[1 - y\langle a, x\rangle\right]_+ \cdot \mathbb{P}(z = 1 \mid a)\right) + \mathbb{E}\left(\left[1 + y\langle a, x\rangle\right]_+ \cdot \mathbb{P}(z = -1 \mid a)\right) \\ &= \mathbb{E}\left(\left[1 - |\langle a, x\rangle|\right]_+ \cdot \mathbb{P}(z = 1 \mid a)\right) + \mathbb{E}\left(\left[1 + |\langle a, x\rangle|\right]_+ \cdot \mathbb{P}(z = -1 \mid a)\right).\end{aligned} \tag{5.62}$$

Combining (5.61) and (5.62), we obtain the following corollary:

Corollary 5.34. *Let the Standing Assumptions III be fulfilled. For any $w \in \mathbb{R}^d$, it then holds*

$$\begin{aligned}\mathbb{E}\left(\hat{f}_x(w) - \hat{f}_x(x)\right) &= \mathbb{E}\left(\left(f_x(w) - f_x(x)\right) \cdot \mathbb{P}(z = 1 \mid a)\right) \\ &\quad + \mathbb{E}\left(\left(\left[1 + y\langle a, w\rangle\right]_+ - \left[1 + |\langle a, x\rangle|\right]_+\right) \cdot \mathbb{P}(z = -1 \mid a)\right).\end{aligned} \tag{5.63}$$

In order to obtain an appropriate recovery result, it remains to fix the model of the noise and to estimate the expected values in (5.63) of the previous corollary, which we will leave out for further research.

5.5 Numerical Experiments

We performed several numerical tests to demonstrate different aspects of the recovery from binary measurements using the ℓ_1-SVM and the $\ell_{1,2}$-SVM. We considered the ℓ_1-SVM as given by (5.30), the $\ell_{1,2}$-SVM as given by (5.55) and the 1-bit compressed sensing algorithm (5.2) from [96].

The ℓ_1-SVM can be recast as a linear program, similar to the basis pursuit which we have discussed in Lemma 3.4. Hence, for the implementation of the ℓ_1-SVM, we used the MATLAB command `linprog`. For the implementation of the $\ell_{1,2}$-SVM, as

well as for the 1-bit compressed sensing algorithm, we used **cvx**, which is a MATLAB based toolbox for convex optimization. Further, the Gaussian measurement vectors were generated using the MATLAB command **randn**, throughout all experiments we set $R = \|x\|_1$. Furthermore, we did each calculation $n = 100$ times and plotted the average.

5.5.1 Dependency on the Scaling Parameter r

The ℓ_1-SVM and the $\ell_{1,2}$-SVM depend on the scaling parameter r of the Gaussian measurement vectors. However, their dependency is quite different, as we have already slightly discussed in Remark 5.32. If the scaling parameter r tends to 0, we have $y_i\langle a_i, w\rangle \leq 1$ with high probability, hence $[1 - y_i\langle a_i, w\rangle]_+ = 1 - y_i\langle a_i, w\rangle$ and the ℓ_1-SVM becomes equivalent to the optimization problem

$$\max_{w\in\mathbb{R}^d} \sum_{i=1}^{m} y_i\langle a_i, w\rangle \quad \text{subject to} \quad \|w\|_1 \leq R,$$

and will give us 1-sparse solutions, as we have discussed in 5.17. In contrary, if r tends to 0, the $\ell_{1,2}$-SVM becomes equivalent to the optimization problem

$$\max_{w\in\mathbb{R}^d} \sum_{i=1}^{m} y_i\langle a_i, w\rangle \quad \text{subject to} \quad \|w\|_1 \leq R,\ \|w\|_2 \leq 1,$$

which is exactly the 1-bit compressed sensing algorithm (5.2). If, on the other hand, r tends to infinity, the hinge loss $[1-y_i\langle a_i, w\rangle]_+$ becomes very large if $\operatorname{sign}(\langle a_i, w\rangle) \neq y_i$, so in that case w is enforced to be consistent with the measurements. Moreover, for large values of r we expect a similar behaviour of the ℓ_1-SVM and the $\ell_{1,2}$-SVM.

Figure 5.2 demonstrates the influence of r on the approximation error of the ℓ_1-SVM and the $\ell_{1,2}$-SVM. Note that the 1-bit compressed sensing algorithm (5.2) is linear in the measurement vectors a_i, hence the solution is independent of scaling and is therefore left out.

The left image of Figure 5.2 shows the dependency of the ℓ_1-SVM on the scaling parameter r. Here we can observe another aspect, namely that the error of reconstruction does not converge to zero if the value of $r = 0.5$ or $r = 1$ is fixed, which is in a good agreement with the error estimate (5.36) of Theorem 5.18. But if we choose $r = \sqrt{m}/20$ or $r = \sqrt{m}/30$, i.e., if r grows with m, the error decreases with the number of measurements.

In the right image of Figure 5.2 we compare the dependency on r of the ℓ_1-SVM and of the $\ell_{1,2}$-SVM. As already suggested above, for small values of r the $\ell_{1,2}$-SVM outperforms the ℓ_1-SVM. But if r is sufficiently large, their error of reconstruction coincides. Beside this comparison, we also observe that the error of the $\ell_{1,2}$-SVM remains almost constant if the scaling parameter r grows. This shows in a way the robustness of the $\ell_{1,2}$-SVM on r and demonstrates that we can choose r to be fixed for the $\ell_{1,2}$-SVM, where the particular choice of r might not be that important.

5.5.2 Comparison of the ℓ_1-SVM, the $\ell_{1,2}$-SVM and the 1-Bit Compressed Sensing Algorithm

In Figure 5.3 we compare the approximation errors of the ℓ_1-SVM, the $\ell_{1,2}$-SVM and the 1-bit compressed sensing algorithm for different values of m, where the sparsity

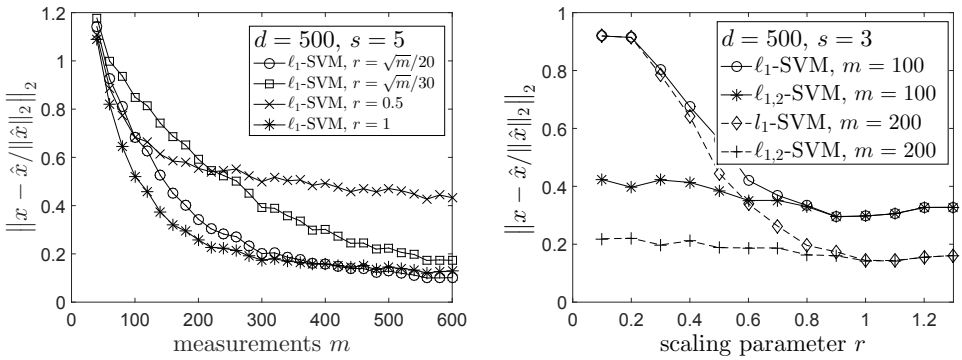

Figure 5.2: Dependency of the ℓ_1-SVM on r (left) and comparison of the dependency on r of the ℓ_1-SVM and the $\ell_{1,2}$-SVM (right). For the left image, for each value of m from 0 to 600, we generated a vector $\tilde{x} \in \mathbb{R}^d$ with exactly $s = 5$ nonzero entries ± 1, taken each with probability $1/2$ and their positions uniformly distributed. Afterwards, we set $x = \tilde{x}/\|\tilde{x}\|_2$ and run the ℓ_1-SVM with different values of $r \in \{0.5, 1, \sqrt{m}/20, \sqrt{m}/30\}$ to get, for each value of r, the recovered vector $\hat{x}$. For each pair (r, m) we then plotted the value of $\|x - \hat{x}/\|\hat{x}\|_2\|_2$. For the right image, we let r ranging from 0 to 1.3. For each value of r, we generated a signal $\tilde{x} \in \mathbb{R}^d$, $d = 500$, with exactly $s = 3$ nonzero entries ± 1 as described above. Afterwards, we again set $x = \tilde{x}/\|\tilde{x}\|_2$ and ran the ℓ_1-SVM and the $\ell_{1,2}$-SVM for different values of $m \in \{100, 200\}$. For each pair of (r, m) we then plotted the distance between x and $\hat{x}/\|\hat{x}\|_2$.

level s and the dimension d are fixed. We again observe that the error of reconstruction does not converge to zero, if we set $r = 1$ for the ℓ_1-SVM. Furthermore, we observe that the error decay of the ℓ_1-SVM with $r = \sqrt{m}/20$, of the $\ell_{1,2}$-SVM with $r = 1$ and of the 1-bit compressed sensing algorithm are similar, with slight advantages for the SVMs.

5.5.3 Dependency on the Number of Measurements m

Figure 5.4 studies the dependency of the approximation error and the classification success of unseen data on the number of measurements m and the dependency d. We observe that the number of measurements m indeed only has to grow logarithmically in the dimension d if we want to guarantee a good approximation result with high probability. Furthermore, the phase transition between the two regions with good and bad performance of the ℓ_1-SVM and the $\ell_{1,2}$-SVM is not quite sharp, as it is, for instance, for the basis pursuit in the usual compressed sensing setting, cf. Figure 3.2. Furthermore, as already observed in Figure 5.3, we observe a slightly better performance of the $\ell_{1,2}$-SVM compared to the ℓ_1-SVM.

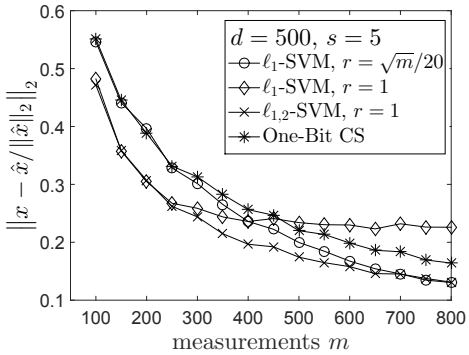

Figure 5.3: Comparison of the ℓ_1-SVM, the $\ell_{1,2}$-SVM and the 1-bit compressed sensing algorithm (5.2). For $d = 500$ fixed and each value of m ranging from 0 to 800, we generated a signal $\tilde{x} \in \mathbb{R}^d$ with entries $\{1, -1, 0.5, -0.5, 0.3\}$, whose locations where chosen uniformly at random. Afterwards we set $x = \tilde{x}/\|\tilde{x}\|_2$ and we then calculated the reconstructions $\hat{x}$ with the ℓ_1-SVM for different values of $r \in \{1, \sqrt{m}/20\}$, the $\ell_{1,2}$-SVM with $r = 1$ and the 1-bit compressed sensing algotihm. Then we plotted the distance between x and $\hat{x}/\|\hat{x}\|_2$.

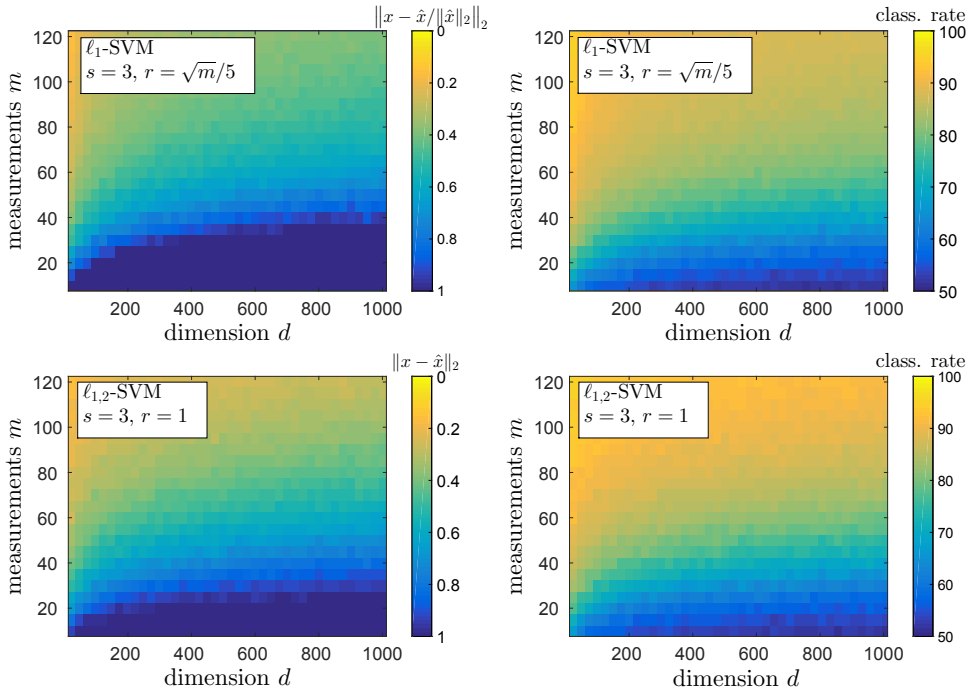

Figure 5.4: Dependency of the ℓ_1-SVM (top) and of the $\ell_{1,2}$-SVM (bottom) on the number of measurements m and the dimension d. We fixed the sparsity $s = 3$, the scaling parameter $r = \sqrt{m}/5$ for the ℓ_1-SVM and $r = 1$ for the $\ell_{1,2}$-SVM. Then, we let the values of m range from 0 to 120 and the values for d range from 0 to 1000. For each fixed pair of (m, d), we then draw a signal $x \in \mathbb{R}^d$ with exactly $s = 5$ nonzero entries ± 1, each with probability $1/2$ and uniformly distributed locations. Afterwards, we calculated the approximation $\hat{x}$ of x using the ℓ_1-SVM and the $\ell_{1,2}$-SVM. For the left images, we plotted their distance $\|x - \hat{x}/\|\hat{x}\|_2\|_2$ for the ℓ_1-SVM and $\|x - \hat{x}\|_2$ for the $\ell_{1,2}$-SVM. For the right image, we drew $n_{test} = 1000$ Gaussian data points $\tilde{a}_i$ using the MATLAB command **`randn`**. Then we counted the amount of data points, which got correctly classified by the recovered $\hat{x}$, i.e., we calculated the amount of data points $\tilde{a}_i$ satisfying $\text{sign}(\langle x, \tilde{a}_i \rangle) = \text{sign}(\langle \hat{x}, \tilde{a}_i \rangle)$.

Chapter 6

Ridge Functions

We follow the study of *ridge functions*, which are multivariate functions of the form

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto g(\langle a, x\rangle)$$

for a univariate function $g\colon \mathbb{R} \to \mathbb{R}$, called the *ridge profile*, and a constant vector $a \in \mathbb{R}^d$, called the *ridge vector*. Here we used the standard notation of the study of ridge functions, which means that we changed the roles of a and x compared to the setting of 1-bit compressed sensing of the previous chapter.

A ridge function f is constant along the hyperplanes perpendicular to a. More precisely, if $w \in \mathbb{R}^d$ is orthogonal to a, for any $x \in \mathbb{R}^d$ we obtain

$$f(x + w) = g(\langle a, x + w\rangle) = g(\langle a, x\rangle) = f(x).$$

In the study of ridge functions we assume g and a to be unknown, and we want to recover both, g and a, in order to get an approximation of f. In the framework of this thesis we interpret g as nonlinearity, so the approximation of ridge functions can be seen as a generalization of the 1-bit compressed sensing problem, where we have the particular choice $g = \operatorname{sign}$.

Many algorithms in the recovery of ridge functions are based on the simple observation that

$$\nabla f(x) = g'(\langle a, x\rangle) \cdot a, \tag{6.1}$$

for $x \in \mathbb{R}^d$ and where ∇f denotes the gradient of f. The idea is then to find an approximation $\tilde{a}$ of the gradient of f in some fixed point x_0 with non-vanishing derivative $g'(\langle a, x_0\rangle)$. Then, since $\nabla f(x)$ and a are collinear, we can use $\tilde{a}$ to find an approximation $\hat{a}$ of a. Once having this approximation, the problem is reduced to find an approximation $\hat{g}$ of the univariate function $g\colon \mathbb{R} \to \mathbb{R}$. Here we can apply standard numerical algorithms as the *spline interpolation*, so in the following we will concentrate on effective algorithms for the recovery of a, whereas the recovery of g and f will always be only given implicitly.

Instead of the approach (6.1) using the gradient of f let us also make the following similar observation: For $\varphi \in \mathbb{R}^d$, the directional derivative of f into the direction φ is given by

$$\frac{\partial f}{\partial \varphi}(x) = \langle \nabla f(x), \varphi\rangle = g'(\langle a, x\rangle) \cdot \langle \varphi, a\rangle. \tag{6.2}$$

If we know in advance that the signal a is sparse, we can choose directions $\varphi_1, \dots, \varphi_m \in \mathbb{R}^d$ at random. In this case (6.2) directly leads us to use techniques from compressed sensing for the recovery of a.

In this chapter we mainly present the results of [68] which was joint work with Jan Vybìral.

6.1 Ridge Functions on Cubes

The authors of [42] studied the uniform approximation of ridge functions $f\colon B_2^d \to \mathbb{R}$ defined on the unit ball $B_2^d \subset \mathbb{R}^d$. Although the possibility of extending this analysis to ridge functions defined on other domains was already mentioned there, no further steps in this direction were done.

In this section we study the uniform approximation of ridge functions defined on the unit cube $[-1,1]^d \subset \mathbb{R}^d$ in detail, i.e., ridge functions of the form

$$f\colon [-1,1]^d \to \mathbb{R}, \quad x \mapsto f(x) = g(\langle a, x\rangle). \tag{6.3}$$

Note that a and g are not uniquely determined by (6.3). Namely, if we choose some $\lambda \neq 0$ and set $\tilde{a} = \lambda a$ and $\tilde{g}(x) = g(x/\lambda)$, we obtain another representation of f of the form (6.3):

$$f(x) = g(\langle a, x\rangle) = g\left(\frac{\langle \lambda a, x\rangle}{\lambda}\right) = \tilde{g}(\langle \tilde{a}, x\rangle). \tag{6.4}$$

Thus, without loss of generality, we can pose a scaling condition on a. In [42] the authors posed the condition $\|a\|_2 = 1$, which fitted together with both the scalar product structure used in the definition of f, as well as the geometry of the domain of f, namely the open Euclidean unit ball.

In our case, it is easy to observe that it will be more convenient to work with the ℓ_1-norm instead. Indeed, assume that the ridge profile $g(t) = t$ is known by an oracle in advance, i.e., for $x \in [-1,1]^d$ we have $f(x) = \langle a, x\rangle$ for an unknown ridge vector $a \in \mathbb{R}^d$. Further, let us assume that we already have some approximation $\hat{a} \in \mathbb{R}^d$ of a with $\|\hat{a} - a\|_1 \leq \varepsilon$ for some $\varepsilon > 0$. Then, using Hölder's inequality, we get

$$\|\hat{f} - f\|_\infty := \sup_{x\in[-1,1]^d} |\hat{f}(x) - f(x)| = \sup_{x\in[-1,1]^d} |\langle \hat{a} - a, x\rangle| \leq \|\hat{a} - a\|_1 \|x\|_\infty \leq \varepsilon.$$

In what follows, we therefore assume

$$\|a\|_1 = 1, \tag{6.5}$$

so that $|\langle a, x\rangle| \leq 1$ holds for all $x \in [-1,1]^d$ and $g\colon [-1,1] \to \mathbb{R}$ is a univariate function defined on the interval $[-1,1]$.

The main idea (6.1) for the approximation of the ridge vector a uses the derivative of g, so g has to be differentiable. Furthermore, we assume that g and g' are Lipschitz continuous with Lipschitz constants $c_0, c_1 > 0$, i.e., that it holds

$$|g(t_0) - g(t_1)| \leq c_0 |t_0 - t_1|, \tag{6.6}$$

$$|g'(t_0) - g'(t_1)| \leq c_1 |t_0 - t_1| \tag{6.7}$$

for every $t_0, t_1 \in [-1,1]$. Last, we assume $g'(0) \neq 0$, since it is known that the approximation of ridge functions may be intractable if this condition is left out [80]. If $g'(0) \neq 0$, by switching from a to $-a$, we can also assume without loss of generality

$$g'(0) > 0. \tag{6.8}$$

Let us summarize the assumptions we made so far:

Standing Assumptions IV

i) $f\colon [-1,1]^d \to \mathbb{R}$ is a ridge function with ridge profile $g\colon [-1,1] \to \mathbb{R}$ and ridge vector $a \in \mathbb{R}^d$, i.e., for $x \in [-1,1]^d$, it holds $f(x) = g(\langle a, x\rangle)$.

ii) The ridge vector is normalized: $\|a\|_1 = 1$.

iii) The ridge profile g is differentiable with $g'(0) > 0$.

iv) g and g' are Lipschitz continuous with Lipschitz constants c_0, c_1, respectively.

6.1.1 Approximation Scheme without Sparsity

Motivated by the formula (6.1), for a small constant $h > 0$ and $j = 1, \dots, d$, we set

$$\tilde{a}_j := \frac{f(he_j) - f(0)}{h}, \tag{6.9}$$

where $e_1, \dots, e_d \in \mathbb{R}^d$ denote the canonical basis vectors in $\mathbb{R}^d$. Note that

$$\lim_{h \to 0} \frac{f(he_j) - f(0)}{h} = g'(0)a_j,$$

so if we choose $h > 0$ small enough, we expect $\tilde{a}_j$ to be a good approximation of $g'(0)a_j$.

Lemma 6.1 ([68])**.** *Let $f\colon [-1,1]^d \to \mathbb{R}$ be a ridge function satisfying the Standing Assumptions IV. For $h > 0$ define $\tilde{a}$ by (6.9). Then it holds*

$$\|\tilde{a} - g'(0)a\|_1 \leq c_1 h. \tag{6.10}$$

Proof. For $j = 1, \dots, d$ the mean value theorem gives the existence of some $\xi_{h,j} \in \mathbb{R}$ with $|\xi_{h,j}| \leq |ha_j|$ and

$$\tilde{a}_j = \frac{f(he_j) - f(0)}{h} = \frac{g(\langle a, he_j\rangle) - g(0)}{h} = \frac{g(ha_j) - g(0)}{h} = g'(\xi_{h,j})a_j.$$

Hence, using the Lipschitz continuity of g', it follows

$$\begin{aligned}\|\tilde{a} - g'(0)a\|_1 &= \sum_{j=1}^{d} |\tilde{a}_j - g'(0)a_j| = \sum_{j=1}^{d} |g'(\xi_{h,j})a_j - g'(0)a_j| \leq \sum_{j=1}^{d} c_1 |\xi_{h,j}||a_j| \\ &\leq c_1 h \sum_{j=1}^{d} |a_j|^2 \leq c_1 h,\end{aligned}$$

where we used $\|a\|_2^2 \leq \|a\|_1^2 = 1$ for the last inequality. □

Defining $\tilde{a}$ by (6.9), the previous lemma states that we already found a good approximation of $g'(0)a$. And since we are interested in finding an approximation of the ℓ_1-normalized ridge vector a, we set

$$\hat{a} = \frac{\tilde{a}}{\|\tilde{a}\|_1}. \tag{6.11}$$

Here we use the assumption $g'(0) > 0$. If otherwise $g'(0) < 0$, we would have to choose $\hat{a} = -\tilde{a}/\|\tilde{a}\|_1$ instead. To estimate the difference between a and $\hat{a}$ we use the following variant of [42, Lemma 3.4], cf. [68, Lemma 3.1].

Lemma 6.2 ([68])**.** *Let $x \in \mathbb{R}^d$ with $\|x\|_1 = 1$. For any $\tilde{x} \in \mathbb{R}^d \backslash \{0\}$ and any $\lambda \in \mathbb{R}$, it holds*

$$\left\| \operatorname{sign}(\lambda) \frac{\tilde{x}}{\|\tilde{x}\|_1} - x \right\|_1 \leq \frac{2\|\tilde{x} - \lambda x\|_1}{\|\tilde{x}\|_1}.$$

Proof. Using the triangle inequality we obtain

$$\begin{aligned} \left\| \operatorname{sign}(\lambda) \frac{\tilde{x}}{\|\tilde{x}\|_1} - x \right\|_1 &\leq \left\| \frac{\operatorname{sign}(\lambda)\tilde{x} - \operatorname{sign}(\lambda)\lambda x}{\|\tilde{x}\|_1} \right\|_1 + \left\| \frac{\operatorname{sign}(\lambda)\lambda x - x\|\tilde{x}\|_1}{\|\tilde{x}\|_1} \right\|_1 \\ &= \frac{\|\tilde{x} - \lambda x\|_1}{\|\tilde{x}\|_1} + \frac{\big||\lambda| - \|\tilde{x}\|_1\big|}{\|\tilde{x}\|_1} \leq \frac{2\|\tilde{x} - \lambda x\|_1}{\|\tilde{x}\|_1}, \end{aligned}$$

which proves the claim. □

Remark 6.3. Since the proof only relies on the triangle inequality, it remains true for any normed space.

Combining Lemma 6.1 with Lemma 6.2 with the particular choices $\lambda = g'(0)$ and $\tilde{x} = \tilde{a}$, we deduce the following bound for the difference between a and $\hat{a}$:

Corollary 6.4 ([68])**.** *Using the notation of Lemma* (6.1) *and defining $\hat{a} \in \mathbb{R}^d$ by* (6.11)*, it holds*

$$\|\hat{a} - a\|_1 \leq \frac{2c_1 h}{\|\tilde{a}\|_1}.$$

Although we now found a good approximation $\hat{a}$ to a, it is not clear how to define the uniform approximation $\hat{f}$ to f. In [42] the authors successfully used the naive approach to sample f along $\hat{a}$, i.e., they set

$$\hat{g}(t) := f(t \cdot \hat{a}) \quad \text{and} \quad \hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle).$$

Using this approach, when trying to show that $\hat{f}$ is uniformly close to f, we have to ensure that $\langle a, \hat{a} \rangle$ is close to 1, which fails in our case. In order to get a new approach, using $\|a\|_1 = 1$, we observe

$$\langle a, \operatorname{sign}(a) \rangle = \|a\|_1 = 1,$$

where we define the sign of a vector entrywise, i.e., for any $x \in \mathbb{R}^d$ we set

$$\operatorname{sign}(x)_j := \operatorname{sign}(x_j). \tag{6.12}$$

Note that the sign-function is discontinuous, hence, $\operatorname{sign}(a)$ and $\operatorname{sign}(\hat{a})$ may be far from each other, even if a and $\hat{a}$ are close. Nevertheless, comparing their scalar product with a, we get

$$\begin{aligned} |\langle a, \operatorname{sign}(a) - \operatorname{sign}(\hat{a})\rangle| &= |\langle a, \operatorname{sign}(a)\rangle - \langle \hat{a}, \operatorname{sign}(\hat{a})\rangle + \langle \hat{a} - a, \operatorname{sign}(\hat{a})\rangle| \\ &= |\langle \hat{a} - a, \operatorname{sign}(\hat{a})\rangle| \leq \|\hat{a} - a\|_1 \cdot \|\operatorname{sign}(\hat{a})\|_\infty = \|\hat{a} - a\|_1. \end{aligned} \tag{6.13}$$

Thus, instead of sampling f along $\hat{a}$, we sample f along $\operatorname{sign}(\hat{a})$. For any $t \in [-1, 1]$ and $x \in [-1, 1]^d$ we set

$$\hat{g}(t) := f(t \cdot \operatorname{sign}(\hat{a})) \quad \text{and} \quad \hat{f}(x) := \hat{g}(\langle \hat{a}, x\rangle). \tag{6.14}$$

Let us summarize this approximation scheme as follows, cf. [68, Algorithm A]:

Algorithm A

Input: Ridge function $f\colon [-1, 1]^d \to \mathbb{R}$ satisfying the Standing Assumptions IV and $h > 0$ small.

- For $j = 1, \dots, d$ set $\tilde{a}_j := \frac{f(he_j) - f(0)}{h}$.
- Set $\hat{a} := \tilde{a}/\|\tilde{a}\|_1$.
- Set $\hat{g}(t) := f(t \cdot \operatorname{sign}(\hat{a}))$ and $\hat{f}(x) := \hat{g}(\langle \hat{a}, x\rangle)$ for $t \in [-1, 1]$ and $x \in [-1, 1]^d$.

Output: Approximation $\hat{f}$.

The next theorem shows that the approximation scheme described in Algorithm A indeed gives a suitable approximation of the ridge function f, cf. [68, Theorem 3.3].

Theorem 6.5 ([68])**.** *Using the notation of Algorithm A, it holds*

$$\|f - \hat{f}\|_\infty \leq 2c_0\|a - \hat{a}\|_1 \leq \frac{4c_0c_1h}{g'(0) - c_1h}, \tag{6.15}$$

where the last inequality only holds if the denominator $g'(0) - c_1h$ is positive.

Proof. Using Lemma 6.1 we first get

$$\|\tilde{a}\|_1 = \|\tilde{a} - g'(0)a + g'(0)a\|_1 \geq \|g'(0)a\|_1 - \|\tilde{a} - g'(0)a\|_1 \geq g'(0) - c_1h$$

and applying Corollary 6.4 yields

$$\|a - \hat{a}\|_1 \leq \frac{2c_1h}{\|\tilde{a}\|_1} \leq \frac{2c_1h}{g'(0) - c_1h},$$

which shows the second inequality in (6.15). To prove the first inequality, we use the Lipschitz continuity of g, $\langle a, \operatorname{sign}(a)\rangle = \|a\|_1 = 1$ and (6.13), to get

$$\begin{aligned}|g(t) - \hat{g}(t)| &= |g(t) - f(t \cdot \operatorname{sign}(\hat{a}))| = |g(t) - g(\langle a, t\operatorname{sign}(\hat{a})\rangle)| \le c_0|t - \langle a, t\operatorname{sign}(\hat{a})\rangle| \\ &= c_0|t| \cdot |\langle a, \operatorname{sign}(a) - \operatorname{sign}(\hat{a})\rangle| \le c_0|t| \cdot \|a - \hat{a}\|_1\end{aligned}$$

for every $t \in [-1, 1]$. Combining this estimate with the definition of $\hat{f}$, for every $x \in [-1, 1]^d$ we end up with

$$\begin{aligned}|f(x) - \hat{f}(x)| &= |g(\langle a, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| \\ &\le |g(\langle a, x\rangle) - g(\langle \hat{a}, x\rangle)| + |g(\langle \hat{a}, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| \\ &\le c_0|\langle a - \hat{a}, x\rangle| + c_0|\langle \hat{a}, x\rangle| \cdot \|a - \hat{a}\|_1 \le 2c_0\|a - \hat{a}\|_1.\end{aligned} \tag{6.16}$$

□

Remark 6.6. i) To find the approximation $\hat{a}$ of a, Algorithm A needs $d + 1$ pointwise evaluations of the function f, namely at the sampling points he_j, $j = 1, \dots, d$ and at 0. Afterwards, any one-dimensional sampling method can be used to get the approximation $\hat{f}$. Since this is a well studied problem, we do not go into detail here and instead refer to [32] and references therein.

ii) The estimate (6.15) heavily depends on the value of $g'(0)$. Especially, the approximation becomes difficult if $g'(0)$ is very small and the estimate (6.15) becomes void if $g'(0) = 0$. This is a very well known aspect in the theory of approximation of ridge function and was studied in great detail in [80] and briefly discussed in [42].

iii) If the ℓ_2-norm $\|a\|_2$ of a is small (note that $\|a\|_1 = 1$ implies $\|a\|_2 \le 1$), the following observation may be of interest: We can improve the estimate (6.10) easily by

$$\|\tilde{a} - g'(0)a\|_1 \le c_1 h\|a\|_2,$$

which results in the enhanced estimate

$$\|f - \hat{f}\|_\infty \le 2c_0\|a - \hat{a}\|_1 \le \frac{4c_0c_1h\|a\|_2^2}{g'(0) - c_1h\|a\|_2^2}. \tag{6.17}$$

6.1.2 Approximation Scheme with Sparsity

In this subsection we assume that the ridge vector $a \in \mathbb{R}^d$ satisfies an additional sparsity constraint, i.e., that most of its coefficients are zero or at least very small. We will use techniques from compressed sensing to use this additional structure for reducing the needed pointwise evaluations of f.

For $m \in \mathbb{N}$, let $\Phi \in \mathbb{R}^{m,d}$ be a normalized Bernoulli matrix with rows $\varphi_1, \dots, \varphi_m \in \mathbb{R}^d$ according to Remark 3.19, i.e., with i.i.d. entries $\varphi_{ij} = \pm 1/\sqrt{m}$. Taking the directional derivative of f at 0 into the directions φ_i, $i = 1, \dots, m$, we get

$$\frac{\partial f}{\partial \varphi_i}(0) = \langle \nabla f(0), \varphi_i\rangle = g'(0)\langle a, \varphi_i\rangle.$$

As in the previous approximation scheme described in Algorithm A we estimate this derivative by finite differences, that is, for a parameter $h > 0$ small and $i = 1, \dots, m$ we set

$$\tilde{y}_i := \frac{f(h\varphi_i) - f(0)}{h}. \tag{6.18}$$

The mean value theorem gives the existence of some $\xi_{h,i} \in \mathbb{R}$ with $|\xi_{h,i}| \leq |h \cdot \langle a, \varphi_i \rangle|$ and

$$\tilde{y}_i = g'(\xi_{h,i}) \langle a, \varphi_i \rangle, \tag{6.19}$$

which leads to the estimate

$$\begin{aligned} \|\tilde{y} - g'(0)\Phi a\|_1 &= \sum_{i=1}^{m} |\tilde{y}_i - g'(0)\langle a, \varphi_i \rangle| = \sum_{i=1}^{m} |g'(\xi_{h,i}) - g'(0)| \cdot |\langle a, \varphi_i \rangle| \\ &\leq c_1 \sum_{i=1}^{m} |\xi_{h,i}| \cdot |\langle a, \varphi_i \rangle| \leq c_1 h \sum_{i=1}^{m} |\langle a, \varphi_i \rangle|^2 \leq c_1 h \sum_{i=1}^{m} \|a\|_1^2 \|\varphi_i\|_\infty^2 \\ &= c_1 h \sum_{i=1}^{m} \frac{1}{m} = c_1 h. \end{aligned} \tag{6.20}$$

Next we can apply the ℓ_1-minimizer (P_1) to recover $g'(0)a$ from $\tilde{y}$. Afterwards we can proceed as in Algorithm A. Let us summarize this procedure as follows, cf. [68, Algorithm B]

Algorithm B

Input: Ridge function $f \colon [-1,1]^d \to \mathbb{R}$ satisfying the Standing Assumptions IV, $m \in \mathbb{N}$ with $d \geq \log(6)^2 m$ and $h > 0$ small.

- Draw a normalized Bernoulli matrix $\Phi \in \mathbb{R}^{m,d}$ with rows $\varphi_1, \dots, \varphi_m$.
- For $i = 1, \dots, m$ set $\tilde{y}_i := \frac{f(h\varphi_i) - f(0)}{h}$.
- Set $\tilde{a} := \Delta_1(\tilde{y})$ with Δ_1 from (P_1).
- Set $\hat{a} := \tilde{a} / \|\tilde{a}\|_1$.
- Set $\hat{g}(t) := f(t \cdot \operatorname{sign}(\hat{a}))$ and $\hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle)$ for $t \in [-1,1]$ and $x \in [-1,1]^d$.

Output: Approximation $\hat{f}$.

The following theorem shows that the approximation scheme described in Algorithm B indeed recovers the ridge function f if the ridge profile a is sufficiently sparse or compressible in the sense that its best s-term approximation is small, cf. [68, Theorem 3.5].

Theorem 6.7 ([68])**.** *Using the notation of Algorithm B, there exist constants $C, C', C'', C''' > 0$ such that for every integer $s \in \mathbb{N}$ with $2s \leq C''m/\log(d/m)$, it holds*

$$\|f - \hat{f}\|_\infty \leq 2c_0\|a - \hat{a}\|_1 \leq 2c_0 \operatorname{err}(a, \hat{a}) \tag{6.21}$$

with probability at least $1 - e^{-\sqrt{md}} - e^{-C'''m}$, provided that $g'(0)(1 - \sigma_s^1(a)) - 2C'h$ is positive and where we set

$$\operatorname{err}(a, \hat{a}) := C \cdot \frac{g'(0) \cdot \sigma_s^1(a) + C'h}{g'(0)(1 - \sigma_s^1(a)) - 2C'h}.$$

Remark 6.8. i) If the ridge vector a is s-sparse, we get

$$\sigma_s^1(a) = \inf_{z \in \Sigma_s^d} \|a - z\|_1 = 0$$

so the estimate (6.21) is simplified to

$$\|f - \hat{f}\|_\infty \leq C \cdot C' \cdot \frac{2c_0 h}{g'(0) - 2C'h}.$$

Hence, in that case Theorem 6.7 yields the same order of approximation as Theorem 6.5, which estimates the performance of Algorithm A.

ii) If the ridge profile a is s-sparse, the condition $2s \leq C''m/\log(d/m)$ implies

$$m \geq 2s \log(d/m)/C''.$$

Hence, using Algorithm B we only need $m = O(s \log d)$ pointwise evaluations of f to reconstruct a, which is an improvement compared to $m = d+1$ samples needed in Algorithm A.

Proof. The first inequality in (6.21) is again provided by (6.16), so it only remains to prove the second inequality. Setting

$$\eta := \tilde{y} - g'(0)\Phi a \in \mathbb{R}^m, \tag{6.22}$$

from (6.20) we get $\|\eta\|_1 \leq c_1 h$ and, similarly, using (6.18) we obtain

$$\begin{aligned} \|\eta\|_\infty &= \max_{i \in [m]} |\tilde{y}_i - g'(0)\langle a, \varphi_i\rangle| = \max_{i \in [m]} |g'(\xi_{h,i}) - g'(0)| \cdot |\langle a, \varphi_i\rangle| \\ &\leq \max_{i \in [m]} c_0 h \cdot |\langle a, \varphi_i\rangle|^2 \leq \frac{c_0 h}{m}. \end{aligned} \tag{6.23}$$

It follows $\|\eta\|_2 \leq c_0 h/\sqrt{m}$, which allows us to estimate the J-norm (3.27) of η by

$$\begin{aligned} \|\eta\|_J &= \max\left\{ \sqrt{m}\|\eta\|_\infty \,,\, \sqrt{\frac{m}{\log(d/m)}}\|\eta\|_2 \right\} \leq \max\left\{ \frac{c_0 h}{\sqrt{m}} \,,\, \frac{c_1 h}{\sqrt{\log(d/m)}} \right\} \\ &\leq \max\left\{ c_1 h \,,\, \frac{c_1 h}{\sqrt{2\log(\log 6)}} \right\} = c_1 h, \end{aligned}$$

where we used $d \geq \log(6)^2 m$ for the last inequality. Hence, by applying Theorem 3.37 to $\tilde{\eta} = \eta/(c_1 h) \in \mathbb{R}^m$, we get the existence of some vector $u \in \mathbb{R}^d$ with

$$\Phi u = \eta \quad \text{and} \quad \|u\|_1 \leq \tilde{C} c_1 h$$

for a constant $\tilde{C} > 0$. Next, let us choose some $0 < \delta < 1/3$ to apply Theorem 3.12 to $g'(0)a + u$. We arrive at

$$\begin{aligned}
\|\Delta_1(\tilde{y}) - g'(0)a\|_1 &= \Big\|\Delta_1\Big(g'(0)\Phi a + \eta\Big) - g'(0)a\Big\|_1 = \Big\|\Delta_1\Big(\Phi\big(g'(0)a + u\big)\Big) - g'(0)a\Big\|_1 \\
&\leq \Big\|\Delta_1\Big(\Phi\big(g'(0)a + u\big)\Big) - \Big(g'(0)a + u\Big)\Big\|_1 + \|u\|_1 \\
&\leq C\sigma_s^1(g'(0)a + u) + \|u\|_1 \leq C\Big(g'(0)\sigma_s^1(a) + \|u\|_1\Big) + \|u\|_1 \\
&\leq (1 + C)\Big(g'(0)\sigma_s^1(a) + \|u\|_1\Big) \leq (1 + C)\Big(g'(0)\sigma_s^1(a) + \tilde{C} c_1 h\Big).
\end{aligned}$$

Finally, applying the ℓ_1-minimizer given by (P_1), we set $\tilde{a} = \Delta_1(\tilde{y})$ and $\hat{a} = \tilde{a}/\|\tilde{a}\|_1$. Using the assumption $g'(0) > 0$, Lemma 6.2 now provides

$$\|a - \hat{a}\|_1 = \left\|\operatorname{sign}(g'(0))\frac{\tilde{a}}{\|\tilde{a}\|_1} - a\right\|_1 \leq \frac{2\|\tilde{a} - g'(0)a\|_1}{\|\tilde{a}\|_1} \leq \frac{2(1 + C)\Big(g'(0)\sigma_s^1(a) + \tilde{C} c_1 h\Big)}{\|\tilde{a}\|_1}.$$

Now we can proceed as in the proof of Theorem 6.5. By estimating the ℓ_1-norm of $\tilde{a}$ from below by

$$\|\tilde{a}\|_1 \geq \|g'(0)a\|_1 - \|\Delta_1(\tilde{y}) - g'(0)a\|_1 \geq g'(0) - (1 + C)\Big(g'(0)\sigma_s^1(a) + \tilde{C} c_1 h\Big),$$

we end up with

$$\|a - \hat{a}\|_1 \leq \frac{2(1 + C)\Big(g'(0)\sigma_s^1(a) + \tilde{C} c_1 h\Big)}{g'(0) - (1 + C)\Big(g'(0)\sigma_s^1(a) + \tilde{C} c_1 h\Big)}$$

yielding the second inequality in (6.21). □

Compared to Algorithm A, Algorithm B uses the prior knowledge that the ridge vector a is sparse to reduce the amount of pointwise evaluations of f.

For the reconstruction of the ridge vector a in Algorithm B we used the basis pursuit, although we have noisy measurements $\tilde{y}$ with bounded noise $e = \tilde{y} - g'(0)\Phi a$. But if we also have prior knowledge on the norm of e, for instance, by knowing the Lipschitz constant c_0 of g, instead of the ℓ_1-minimizer Δ_1 given by (P_1), we could also use the $\ell_{1,\varepsilon}$-minimizer $\Delta_{1,\varepsilon}$ given by $(\mathrm{P}_{1,\varepsilon})$.

Algorithm B'

Input: Ridge function $f\colon [-1,1]^d \to \mathbb{R}$ satisfying the Standing Assumptions IV, $m \in \mathbb{N}$ with $d \geq \log(6)^2 m$, $h > 0$ small and $\varepsilon \geq c_0 h/\sqrt{m}$, where c_0 denotes the Lipschitz constant of g.

- Draw a normalized Bernoulli matrix $\Phi \in \mathbb{R}^{m,d}$ with rows $\varphi_1, \dots, \varphi_m$.
- For $i = 1, \dots, m$ set $\tilde{y}_i := \frac{f(h\varphi_i) - f(0)}{h}$.
- Set $\tilde{a} := \Delta_{1,\varepsilon}(\tilde{y})$ with $\Delta_{1,\varepsilon}$ from $(\mathrm{P}_{1,\varepsilon})$.
- Set $\hat{a} := \tilde{a}/\|\tilde{a}\|_1$.
- Set $\hat{g}(t) := f(t \cdot \operatorname{sign}(\hat{a}))$ and $\hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle)$ for $t \in [-1, 1]$ and $x \in [-1, 1]^d$.

Output: Approximation $\hat{f}$.

Theorem 6.9. *Using the notation of Algorithm B', there exist some constants $C, C', C'', C''' > 0$ such that for every integer s with $2s \le C''m/\log(d/m)$, it holds*

$$\|f - \hat{f}\|_\infty \le 2c_0 \|a - \hat{a}\|_1 \le 2c_0 \operatorname{err}(a, \hat{a}) \tag{6.24}$$

with probability at least $1 - e^{-\sqrt{md}} - e^{-C'''m}$, provided that $g'(0)(1 - \sigma_s^1(a)) - C'\sqrt{s}\varepsilon$ is positive and where we set

$$\operatorname{err}(a, \hat{a}) := C \cdot \frac{g'(0) \cdot \sigma_s^1(a) + C'\sqrt{s}\varepsilon}{g'(0)(1 - \sigma_s^1(a)) - C'\sqrt{s}\varepsilon}.$$

Remark 6.10. If we choose $\varepsilon = c_0 h/\sqrt{m}$, we obtain

$$\sqrt{s}\varepsilon = c_0 h \frac{\sqrt{s}}{\sqrt{m}} \le c_0 h.$$

Hence, in that case (6.24) yields the same approximation rate as (6.21), where we used the basis pursuit instead of the $\ell_{1,\varepsilon}$-minimizer.

Proof. From (6.22) we obtain

$$\tilde{y} = g'(0)\Phi a + \eta$$

for some $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \le c_0 h/\sqrt{m} \le \varepsilon$. Applying Theorem 3.35 we deduce

$$\|\Delta_{1,\varepsilon}(\tilde{y}) - g'(0)a\|_1 = \|\tilde{a} - g'(0)a\|_1 \le C g'(0) \sigma_s^1(a) + c'\sqrt{s}\varepsilon.$$

Now we proceed as in the proof of Theorem 6.7. □

6.2 Approximation of Ridge Functions from Noisy Measurements

In this section we study the approximation of ridge functions defined on the open unit ball $B_2^d = \{x \in \mathbb{R}^d \mid \|x\|_2 < 1\}$ from a limited number of point samples, which are affected by random Gaussian noise. The idea to use the Dantzig selector Δ_{DS} given by (3.29) for the recovery of the ridge vector a from noisy measurements was already proposed in [42], but no further study was done there.

To be more precise, in this section we consider ridge functions

$$f\colon B_2^d \to \mathbb{R}, \quad x \mapsto f(x) = g(\langle a, x\rangle) \tag{6.25}$$

for some compressible ridge vector $a \in \mathbb{R}^d$ satisfying

$$\|a\|_2 = 1, \quad \|a\|_1 \leq R \tag{6.26}$$

for some $R > 0$ and some differentiable ridge profile $g\colon [-1,1] \to \mathbb{R}$ with $g'(0) > 0$ such that g and g' are Lipschitz continuous with constants c_0, c_1 as in (6.6), (6.7). Let us summarize the assumptions for this section as follows:

Standing Assumptions V

i) $f\colon B_2^d \to \mathbb{R}$ is a ridge function with ridge profile $g\colon [-1,1] \to \mathbb{R}$ and ridge vector $a \in \mathbb{R}^d$, i.e., for $x \in B_2^d$, it holds $f(x) = g(\langle a, x\rangle)$.

ii) The ridge vector is compressible in the sense that $\|a\|_2 = 1$ and $\|a\|_1 \leq R$ for some $R > 0$.

iii) The ridge profile g is differentiable with $g'(0) > 0$.

iv) g and g' are Lipschitz continuous with Lipschitz constants c_0, c_1, respectively.

Since we assume the ridge vector a to be compressible, we shall again follow Algorithm B for the reconstruction of a. Hence, for $d \geq \log(6)^2 m$, let $\Phi \in \mathbb{R}^{m,d}$ be a normalized Bernoulli matrix according to Remark 3.19 with rows $\varphi_1, \ldots, \varphi_m \in \mathbb{R}^d$. To make the presentation technically simpler, we assume the value $f(0)$ to be given precisely, which can be achieved by resampling $f(0)$ several times and taking the mean. For a small constant $h > 0$ and $i = 1, \ldots, m$, we then set

$$\tilde{y}_i = \frac{f(h\varphi_i) + \tilde{z}_i - f(0)}{h} = \frac{f(h\varphi_i) - f(0)}{h} + \frac{\tilde{z}_i}{h}, \tag{6.27}$$

where we assume the pointwise evaluations $f(h\varphi_i)$ to be perturbed by random noise $\tilde{z}_i$. Here we assume $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_m)$ to have i.i.d. entries $\tilde{z}_i \sim \mathcal{N}(0, \sigma^2)$ for some small $\sigma > 0$ and we set

$$z_i := \frac{\tilde{z}_i}{h} \sim \mathcal{N}\left(0, \frac{\sigma^2}{h^2}\right). \tag{6.28}$$

To recover a from $\tilde{y}$ we use the Dantzig selector (3.29) instead of ℓ_1-minimization. After we found an approximation $\hat{a}$ of a, we can use the approach of [42] for the construction of $\hat{g}$ and $\hat{f}$. Let us summarize this procedure as the following algorithm, cf. [68, Algorithm C]:

Algorithm C

Input: Ridge function $f\colon B_2^d \to \mathbb{R}$ satisfying the Standing Assumptions V, $m \in \mathbb{N}$ with $m \leq \log(6)^2 d$ and $h, \sigma > 0$.

- Draw a normalized Bernoulli matrix $\Phi \in \mathbb{R}^{m,d}$ with rows $\varphi_1, \dots, \varphi_m$.
- For $i = 1, \dots, m$ set $\tilde{y}_i := \frac{f(h\varphi_i) - f(0)}{h} + z_i$ for some i.i.d. $z_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{h^2}\right)$.
- Set $\tilde{a} := \Delta_{DS}(\tilde{y})$ with Δ_{DS} from (3.29) and $\lambda_d = \sqrt{2\log(d)}$.
- Set $\hat{a} := \tilde{a}/\|\tilde{a}\|_2$.
- Set $\hat{g}(t) := f(t\hat{a})$ and $\hat{f}(x) := \hat{g}(\langle \hat{a}, x\rangle)$ for $t \in [-1, 1]$ and $x \in B_2^d$.

Output: Approximation $\hat{f}$.

The next theorem estimates the performance of Algorithm C, cf. [68, Theorem 4.1].

Theorem 6.11 ([68])**.** *Using the notation of Algorithm C, there exists a constant $C > 0$ such that for every integer s with $3s \leq Cm/\log(d/m)$, with high probability it holds*

$$\|f - \hat{f}\|_\infty \leq 2c_0\|a - \hat{a}\|_2 \leq \frac{4c_0 \cdot \mathrm{err}(a, \hat{a})}{g'(0) - \mathrm{err}(a, \hat{a})}, \tag{6.29}$$

where we set

$$\mathrm{err}(a, \hat{a}) := C' c_1 \left(\min_{1 \leq t \leq s} \left(\log(d)\Big(t\frac{\sigma^2}{h^2} + \frac{\tilde{R}^2}{t}\Big)\right)^{1/2} + hR^2\sqrt{\frac{\log(d/m)}{m}} \right), \tag{6.30}$$

$$\tilde{R} := R(1 + C'' hR) \tag{6.31}$$

for constants $C', C'' > 0$.

Remark 6.12. i) The input parameter $h > 0$ of Algorithm C has to be chosen carefully. If we choose $h > 0$ very small, the variance of the Gaussian noise (6.28) increases, leading to a higher noise level and a larger error bound in (6.29). On the other hand, if we choose $h > 0$ very large, we get a worse approximation of the directional derivative of f which leads to a worse approximation of a. Concluding, these simple arguments lead to the intuition that there might be an optimal trade-off value for h which we will also discuss in the numerical examples. Unfortunately, the optimal value for h might depend on the a-priori unknown function g and can, therefore, hardly be identified. We refer also to [55] for a brief discussion of this phenomenon.

ii) In this section we study ridge functions defined on the unit ball to use the estimates on the Dantzig selector available in the literature. To adapt this approach to ridge functions defined on the unit cube, it would be necessary to introduce an ℓ_1-version of Theorem 3.39 first.

Proof. As in the noiseless case, from the mean value theorem we get

$$y_i := \frac{f(h\varphi_i) - f(0)}{h} = g'(\xi_{h,i})\langle a, \varphi_i\rangle \tag{6.32}$$

for some $\xi_{h,i}$ between 0 and $h\langle a, \varphi_i\rangle$, $i = 1, \dots, m$. We set

$$\eta = y - g'(0)\Phi a \in \mathbb{R}^m \tag{6.33}$$

and in order to apply Theorem 3.37 we estimate the entries of η by

$$\begin{aligned}|\eta_i| &= |y_i - g'(0)\langle a, \varphi_i\rangle| = |g'(\xi_{h,i}) - g'(0)| \cdot |\langle a, \varphi_i\rangle| \le c_1 h \langle a, \varphi_i\rangle^2 \\ &\le c_1 h \|a\|_1^2 \cdot \|\varphi_i\|_\infty^2 \le \frac{c_1 h R^2}{m},\end{aligned}$$

yielding the estimates

$$\begin{aligned}\|\eta\|_2 &= \left(\sum_{i=1}^m \eta_i^2\right)^{1/2} \le \left(\sum_{i=1}^m \frac{c_1^2 h^2 R^4}{m^2}\right)^{1/2} = \frac{c_1 h R^2}{\sqrt{m}}, \\ \|\eta\|_\infty &= \max_{i=1,\dots,m} |\eta_i| \le \frac{c_1 h R^2}{m}.\end{aligned}$$

Applying Theorem 3.37 to $\eta/(c_1 h R^2)$, with high probability there exists some $u \in \mathbb{R}^d$ with $\Phi(u) = \eta$ and

$$\|u\|_1 \le C_3 c_1 h R^2, \qquad \|u\|_2 \le C_4 c_1 h R^2 \sqrt{\frac{\log(d/m)}{m}}$$

for some constants $C_1, C_3, C_4 > 0$ from Theorem 3.37. Further, we get

$$\begin{aligned}\|g'(0)a + u\|_{1,\infty} &\le \|g'(0)a\|_1 + \|u\|_1 \le g'(0)R + C_3 c_1 h R^2 \\ &\le c_1 R(1 + C_3 h R) =: c_1 \tilde{R}.\end{aligned}$$

Now we can apply Theorem 3.39 to $\tilde{a} = \Delta_{DS}(\tilde{y})$ to get with high probability

$$\begin{aligned}\|\tilde{a} - g'(0)a\|_2 &= \|\Delta_{DS}(\tilde{y}) - g'(0)a\|_2 = \|\Delta_{DS}(g'(0)\Phi a + \eta + z) - g'(0)a\|_2 \\ &\le \|\Delta_{DS}(\Phi(g'(0)a + u) + z) - g'(0)a - u\|_2 + \|u\|_2 \\ &\le \min_{1\le t\le s}\left(2C_2 \log(d)\Big(t\frac{\sigma^2}{h^2} + \frac{c_1^2\tilde{R}^2}{t}\Big)\right)^{1/2} + C_4 c_1 h R^2\sqrt{\frac{\log(d/m)}{m}} \\ &\le C c_1 \left(\min_{1\le t\le s}\left(\log(d)\Big(t\frac{\sigma^2}{h^2} + \frac{\tilde{R}^2}{t}\Big)\right)^{1/2} + hR^2\sqrt{\frac{\log(d/m)}{m}}\right) \\ &=: \mathrm{err}(a, \hat{a})\end{aligned}$$

for some constants C'', C_2. We set $\hat{a} := \tilde{a}/\|\tilde{a}\|_2$ and Lemma 6.2 gives

$$\begin{aligned}\|\hat{a} - a\|_2 &= \left\|\mathrm{sign}(g'(0))\frac{\tilde{a}}{\|\tilde{a}\|_2} - a\right\|_2 \le \frac{2\|\tilde{a} - g'(0)a\|_2}{\|\tilde{a}\|_2} \le \frac{2\,\mathrm{err}(a, \hat{a})}{\|\tilde{a}\|_2} \\ &\le \frac{2\,\mathrm{err}(a, \hat{a})}{g'(0)\|a\|_2 - \|\tilde{a} - g'(0)a\|_2} = \frac{2\,\mathrm{err}(a, \hat{a})}{g'(0) - \mathrm{err}(a, \hat{a})},\end{aligned}$$

which proves the second inequality in (6.29). Finally, for $t \in [-1,1]$ and $x \in B_2^d$ we define

$$\hat{g}(t) := f(t\hat{a}) \quad \text{and} \quad \hat{f}(x) := \hat{g}(\langle \hat{a}, x \rangle).$$

It remains to show that $\hat{f}$ indeed is a good approximation of f. For any $t \in [-1,1]$ we first observe

$$\begin{aligned} |\hat{g}(t) - g(t)| &= |g(t\langle a, \hat{a}\rangle) - g(t)| \leq c_0 |t(\langle a, \hat{a}\rangle - 1)| \leq c_0 |\langle a, \hat{a} - a\rangle| \\ &\leq c_0 \|a - \hat{a}\|_2, \end{aligned}$$

so that we can bound the pointwise difference of f and $\hat{f}$ for any $x \in B_2^d$ by

$$\begin{aligned} |f(x) - \hat{f}(x)| &= |g(\langle a, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| \leq |g(\langle a, x\rangle) - g(\langle \hat{a}, x\rangle)| + |g(\langle \hat{a}, x\rangle) - \hat{g}(\langle \hat{a}, x\rangle)| \\ &\leq c_0 |\langle a, x\rangle - \langle \hat{a}, x\rangle| + c_0 \|a - \hat{a}\|_2 \leq 2c_0 \|\hat{a} - a\|_2, \end{aligned}$$

which finishes the proof. □

6.3 Approximation of Translated Radial Functions

The Algorithms A-C presented in this chapter, as well as the methods proposed in [42], were developed for the quite restrictive setting of ridge functions. However, the aim of this section is to demonstrate that we can use the same tools for the approximation of functions of a different, but similar type.

We consider translated radial functions defined on the unit ball, i.e., functions of the form

$$f \colon B_2^d \to \mathbb{R}, \quad x \mapsto g(\|a - x\|_2^2) \tag{6.34}$$

for a fixed normalized center point $a \in \mathbb{R}^d$ with

$$\|a\|_2 = 1 \tag{6.35}$$

and an unknown function $g \colon [0,4] \to \mathbb{R}$. In contrary to ridge functions, which are constant along the hyperplanes perpendicular to the ridge vector, translated radial functions are constant on the spheres centered in a.

As in the setting of ridge functions, we again assume g to be differentiable and g and g' to be Lipschitz continuous, i.e., that there exist constants $c_0, c_1 > 0$ such that

$$|g(t_1) - g(t_2)| \leq c_0 |t_1 - t_2|, \tag{6.36}$$

$$|g'(t_1) - g'(t_2)| \leq c_1 |t_1 - t_2| \tag{6.37}$$

holds for all $t_1, t_2 \in [0,4]$.

In order to develop an approximation scheme for translated radial functions, we want to adapt the approach for ridge functions, which we described in Algorithm A. The main observation for the approximation of a ridge function was given by (6.1), which told us that the gradient of a ridge function is collinear with the ridge vector a. In contrary, for a translated radial function f we observe

$$\nabla f(x) = 2g'(\|a - x\|_2^2)(x - a). \tag{6.38}$$

Hence, the gradient of f at 0 is given by $-2g'(1)a$. Therefore we replace the condition $g'(0) \neq 0$ (6.8), which we used for the approximation of ridge functions, by the condition $g'(1) \neq 0$.

Note that here we do not have a scaling freedom as in the case of ridge functions (cf. (6.4)), so we cannot fix the sign of $g'(1)$ without loss of generality. Nevertheless, for brevity we assume that

$$g'(1) > 0, \tag{6.39}$$

which we will also motivate later on. Before, let us fix the assumptions we made so far:

Standing Assumptions VI

i) $f\colon B_2^d \to \mathbb{R}$ is a translated radial function with center point $a \in \mathbb{R}^d$, i.e., for some $g\colon [0,4] \to \mathbb{R}$ and any $x \in B_2^d$, it holds $f(x) = g(\|a - x\|_2^2)$.

ii) The center point is normalized: $\|a\|_2 = 1$.

iii) g is differentiable with $g'(1) > 0$.

iv) g and g' are Lipschitz continuous with Lipschitz constants c_0, c_1, respectively.

6.3.1 Approximation Scheme without Sparsity

In order to develop an approximation scheme for translated radial functions, we want to adapt the approach for ridge functions, which we have described in Algorithm A.

The main motivation for the approximation of a ridge function f was given by (6.1), which told us that the gradient of f is collinear with the ridge vector a. This simple observation gave the idea to approximate the gradient of f by finite differences (6.9)

$$\tilde{a}_i = \frac{f(0) - f(he_i)}{h}$$

for some $h > 0$ and where $e_1, \dots, e_d$ denote the canonical basis vectors of $\mathbb{R}^d$. For the approximation of translated radial functions we should not directly translate this approach. However, using other sample points than 0 and he_i, we can at least transfer the idea. So let us first discuss how we should choose the sample points.

For $h > 0$ small and fixed $x_1, \dots, x_d \in \mathbb{R}^d$, which we will determine later on, similar to (6.9) for the case of ridge functions we set

$$\tilde{a}_j = \frac{f(he_j + x_j) - f(x_j)}{h}, \qquad j = 1, \dots, d. \tag{6.40}$$

With help of the mean value theorem we obtain

$$\begin{aligned}\tilde{a}_j &= \frac{f(he_j + x_j) - f(x_j)}{h} = \frac{g(\|a - he_j - x_j\|_2^2) - g(\|a - x_j\|_2^2)}{h} \\ &= g'(\xi_{h,j}) \frac{\|a - he_j - x_j\|_2^2 - \|a - x_j\|_2^2}{h} = g'(\xi_{h,j}) \frac{-2h\langle a, e_j\rangle + h^2 + 2h\langle x_j, e_j\rangle}{h}\end{aligned}$$

for some $\xi_{h,j}$ between $\|a - he_j - x_j\|_2^2$ and $\|a - x_j\|_2^2$. By choosing $x_j = -(h/2)e_j$ we get

$$\tilde{a}_j = \frac{f(\frac{h}{2}e_j) - f(-\frac{h}{2}e_j)}{h} = -2g'(\xi_{h,j})a_j \tag{6.41}$$

for some $\xi_{h,j}$ between $\|a - (h/2)e_j\|_2^2$ and $\|a + (h/2)e_j\|_2^2$. Since $h > 0$ is very small, with this choice of x_j we obtain that $\xi_{h,j}$ is close to $\|a\|_2 = 1$, since

$$\begin{aligned}|\xi_{h,j} - 1| &\leq \max\left\{\left|\|a - (h/2)e_j\|_2^2 - 1\right|, \left|\|a + (h/2)e_j\|_2^2 - 1\right|\right\} \\ &= \max\left\{\left|-ha_j + h^2/4\right|, \left|ha_j + h^2/4\right|\right\} \leq h + h^2/4. \end{aligned} \tag{6.42}$$

Using the Lipschitz continuity of g' we conclude that $\tilde{a}_j = -2g'(\xi_{h,j})a_j$ is close to $-2g'(1)a_j = -2g'(\|a\|_2^2)a_j$, since it holds

$$\begin{aligned}\|\tilde{a} + 2g'(1)a\|_2^2 &= \sum_{j=1}^{d} \left(-2g'(\xi_{h,j})a_j + 2g'(1)a_j\right)^2 = 4\sum_{j=1}^{d} |g'(\xi_{h,j}) - g'(1)|^2 a_j^2 \\ &\leq 4c_1^2(h + h^2/4)^2. \end{aligned} \tag{6.43}$$

Due to this construction and the assumption $g'(1) \neq 0$, the normalized vector $\hat{a} := -\tilde{a}/\|\tilde{a}\|_2$ approximates a, possibly up to the sign of $g'(1)$. For brevity, we assume $g'(1) > 0$ (6.39), although this assumption cannot be made without loss of generality, as we have already mentioned above. However, if the sign of $g'(1)$ is unknown, we can sample f along any vector $\hat{a}' \in \mathbb{R}^d$ orthogonal to $\hat{a}$ to identify the sign of $g'(1)$ and afterwards assign it to $\hat{a}$.

Once we have found the approximation $\hat{a}$ to a, the problem is again reduced to finding an approximation of the univariate function g and finally of f. Let us summarize this approximation scheme as Algorithm D, cf. [68, Algorithm D]:

Algorithm D

Input: Translated radial function f satisfying the Standing Assumptions VI and $h > 0$.

- Set $\tilde{a}_j = \frac{f(he_j/2) - f(-he_j/2)}{h}$, $j = 1 \ldots, d$.
- Set $\hat{a} := -\tilde{a}/\|\tilde{a}\|_2$.
- Set $\hat{g}(t) := f(\hat{a}(1 - \sqrt{t}))$ and $\hat{f}(x) := \hat{g}(\|\hat{a} - x\|_2^2)$ for $t \in [0, 4]$ and $x \in B_2^d$.

Output: Approximation $\hat{f}$.

The performance of Algorithm D is estimated by the following theorem, cf. [68, Theorem 5.1]:

Theorem 6.13 ([68]). *Using the notation of Algorithm D, it holds*

$$\|f - \hat{f}\|_\infty \le c_0 \left(2\|\hat{a} - a\|_2 + \|\hat{a} - a\|_2^2\right) \tag{6.44}$$

and, if $g'(1) - c_1(h + h^2/4)$ is positive, it further holds

$$\|\hat{a} - a\|_2 \le \frac{2c_1(h + h^2/4)}{g'(1) - c_1(h + h^2/4)}. \tag{6.45}$$

Proof. We start estimating the difference between a and $\hat{a}$. By (6.42) and the Lipschitz continuity of g', for $j = 1, \dots, d$ we get

$$g'(1) - |g'(\xi_{h,j})| \le |g'(1) - g'(\xi_{h,j})| \le c_1|1 - \xi_{h,j}| \le c_1(h + h^2/4),$$

hence

$$|g'(\xi_{h,j})| \ge g'(1) - c_1(h + h^2/4). \tag{6.46}$$

If the right hand side of this inequality is positive, using (6.41) we get

$$\begin{aligned}\|\tilde{a}\|_2^2 &= \sum_{j=1}^d |\tilde{a}_j|^2 = 4\sum_{j=1}^d |g'(\xi_{h,j})a_j|^2 \ge 4\sum_{j=1}^d \left(g'(1) - c_1(h + h^2/4)\right)^2 |a_j|^2 \\ &= 4\left(g'(1) - c_1(h + h^2/4)\right)^2.\end{aligned}$$

Now we can apply Lemma 6.2 to bound the difference between a and $\hat{a}$ by

$$\|a - \hat{a}\|_2 \le \frac{4c_1(h + h^2/4)}{\|\tilde{a}\|_2} \le \frac{2c_1(h + h^2/4)}{g'(1) - c_1(h + h^2/4)}, \tag{6.47}$$

provided that the denominator is positive. Given the approximation $\hat{a}$ of a, for any $t \in [0, 4]$ and $x \in B_2^d$ we define

$$\hat{g}(t) := f\left(\hat{a}(1 - \sqrt{t})\right) \quad \text{and} \quad \hat{f}(x) := \hat{g}(\|\hat{a} - x\|_2^2).$$

Using the Lipschitz continuity of g we indeed obtain that $\hat{g}$ is a suitable uniform approximation of g, since for any $t \in [0, 4]$ we get

$$\begin{aligned}|g(t) - \hat{g}(t)| &= \left|g(t) - g\left(\left\|a - \hat{a}\left(1 - \sqrt{t}\right)\right\|_2^2\right)\right| \le c_0 \left|t - \left\|a - \hat{a} + \sqrt{t}\hat{a}\right\|_2^2\right| \\ &= c_0 \left|2\sqrt{t}\langle a - \hat{a}, \hat{a}\rangle + \|a - \hat{a}\|_2^2\right| = c_0 \left|(2\sqrt{t} - 1)\langle a - \hat{a}, \hat{a}\rangle + \langle a - \hat{a}, a\rangle\right| \\ &= 2c_0 \left|\left(1 - \sqrt{t}\right)(1 - \langle a, \hat{a}\rangle)\right| = c_0 \left|1 - \sqrt{t}\right| \cdot \|a - \hat{a}\|_2^2 \le c_0\|a - \hat{a}\|_2^2,\end{aligned}$$

where we used the fact that $\|a\|_2 = \|\hat{a}\|_2 = 1$ several times. With this estimate for the difference between g and $\hat{g}$ we finally get for any $x \in B_2^d$

$$\begin{aligned}|f(x) - \hat{f}(x)| &= \left|g\left(\|a - x\|_2^2\right) - \hat{g}\left(\|\hat{a} - x\|_2^2\right)\right| \\ &\le \left|g\left(\|a - x\|_2^2\right) - g\left(\|\hat{a} - x\|_2^2\right)\right| + \left|g\left(\|\hat{a} - x\|_2^2\right) - \hat{g}\left(\|\hat{a} - x\|_2^2\right)\right| \\ &\le c_0 \left|\|a - x\|_2^2 - \|\hat{a} - x\|_2^2\right| + c_0\|a - \hat{a}\|_2^2 = 2c_0|\langle a - \hat{a}, x\rangle| + c_0\|a - \hat{a}\|_2^2 \\ &\le c_0 \left(2\|a - \hat{a}\|_2 + \|a - \hat{a}\|_2^2\right). \end{aligned} \tag{6.48}$$

□

Remark 6.14. In Algorithm D, and therefore also in Theorem 6.13, we assume g' to be Lipschitz continuous on its whole domain $[0,4]$. However, if we only assume the Lipschitz continuity on the open interval between $1-(h+h^2/4)$ and $1+(h+h^2/4)$, the estimate (6.47) still remains true and we can still recover the centerpoint $a \in \mathbb{R}^d$.

In particular, this even applies to the case when g and g' are unbounded near the origin, i.e., if f has a singularity in a. In that case, a uniform approximation of f is out of reach, but we are still able to recover the position of the singularity.

6.3.2 Approximation Scheme with Sparsity

As in the approximation scheme for ridge functions we can use techniques from compressed sensing to reduce the amount of measurements if $a \in \mathbb{R}^d$ is compressible. To be more precise, if $a \in \mathbb{R}^d$ satisfies the compressibility conditions

$$\|a\|_2 = 1, \qquad \|a\|_1 \leq R \tag{6.49}$$

for some $R > 0$. For a normalized Bernoulli matrix $\Phi \in \mathbb{R}^{m,d}$ with rows $\varphi_1, \dots, \varphi_m \in \mathbb{R}^m$, we set

$$\tilde{y}_i := \frac{f(h\varphi_i/2) - f(-h\varphi_i/2)}{h}, \qquad i = 1, \dots, m. \tag{6.50}$$

Note that f is only defined on the unit ball B_2^d. Since $\|\varphi_i\|_2 = \sqrt{d/m}$, we must always have at least $h < 2\sqrt{m/d}$ to ensure that $f(h\varphi_i/2)$ is well defined. To allow for comparison with the previous, non compressible case, we set

$$\tilde{h} = h/2 \cdot \sqrt{d/m}, \tag{6.51}$$

which leads to

$$\tilde{y}_i = \frac{f\left(\tilde{h}\frac{\varphi_i}{\|\varphi_i\|_2}\right) - f\left(-\tilde{h}\frac{\varphi_i}{\|\varphi_i\|_2}\right)}{h}. \tag{6.52}$$

By defining the deterministic noise $\eta \in \mathbb{R}^m$ by

$$\eta = \tilde{y} + 2g'(1)\Phi a \tag{6.53}$$

and using the mean value theorem, we get

$$\begin{aligned}
|\eta_i| &= \left|\frac{1}{h}\left(g(\|a - h\varphi_i/2\|_2^2) - g(\|a + h\varphi_i/2\|_2^2)\right) + 2g'(1)\langle\varphi_i, a\rangle\right| \\
&= \left|\frac{1}{h}g'(\xi_{h,i})\left(\|a - h\varphi_i/2\|_2^2 - \|a + h\varphi_i/2\|_2^2\right) + 2g'(1)\langle\varphi_i, a\rangle\right| \\
&= \left|2\langle\varphi_i, a\rangle\left(g'(\xi_{h,i}) - g'(1)\right)\right| \leq 2c_1\|a\|_1 \cdot \|\varphi_i\|_\infty |\xi_{h,i} - 1| \\
&\leq \frac{2c_1 R}{\sqrt{m}} \cdot |\xi_{h,i} - 1|
\end{aligned}$$

for some unknown $\xi_{h,i}$ between $\|a - h\varphi_i/2\|_2^2$ and $\|a + h\varphi_i/2\|_2^2$. Similar to (6.42) we further estimate

$$\begin{aligned}
|\xi_{h,i} - 1| &\leq \max\left\{\left|\|a - h\varphi_i/2\|_2^2 - 1\right|, \left|\|a + h\varphi_i/2\|_2^2 - 1\right|\right\} = \frac{h^2}{4}\|\varphi_i\|_2^2 + h|\langle a, \varphi_i\rangle| \\
&\leq \frac{h^2 d}{4m} + \frac{hR}{\sqrt{m}} = \tilde{h}^2 + \frac{2R\tilde{h}}{\sqrt{d}}
\end{aligned}$$

and we obtain

$$\|\eta\|_2 \leq 2c_1 R\left(\tilde{h}^2 + \frac{2R\tilde{h}}{\sqrt{d}}\right). \tag{6.54}$$

To recover a from (6.52) we use the $\ell_{1,\varepsilon}$-minimizer ($\mathrm{P}_{1,\varepsilon}$). For $\varepsilon \geq \|\eta\|_2$ we set

$$\tilde{a} := \Delta_{1,\varepsilon}(\tilde{y}) \quad \text{and} \quad \hat{a} := -\tilde{a}/\|\tilde{a}\|_2.$$

Using Theorem 3.35 we obtain with high probability

$$\|\tilde{a} + 2g'(1)a\|_2 \leq Cg'(1)\frac{\sigma_s^1(a)}{\sqrt{s}} + c'h =: \mathrm{err}(a, \hat{a})$$

for some constants C, c', provided that $m \geq C''m/\log(d/m)$. Applying Lemma 6.2 gives the estimate

$$\|\hat{a} - a\|_2 \leq \frac{2\,\mathrm{err}(a, \hat{a})}{\|\tilde{a}\|_2},$$

where we can bound $\|\tilde{a}\|_2$ from above by

$$\|\tilde{a}\|_2 \geq 2g'(1)\|a\|_2 - \|\tilde{a} - 2g'(1)a\|_2 \geq 2g'(1) - \mathrm{err}(a, \hat{a}).$$

Provided that $2g'(1) - \mathrm{err}(a, \hat{a})$ is positive, we end up with

$$\|\hat{a} - a\|_2 \leq \frac{2\,\mathrm{err}(a, \hat{a})}{2g'(1) - \mathrm{err}(a, \hat{a})}. \tag{6.55}$$

Afterwards we can proceed as in Algorithm D, which we summarize as follows:

Algorithm E

Input: Translated radial function f satisfying the Standing Assumptions VI such that the additional compressibility condition (6.49) holds, $h > 0$, $m \in \mathbb{N}$ and $\varepsilon \geq 2c_1 R\left(\tilde{h}^2 + \frac{2R\tilde{h}}{\sqrt{d}}\right)$.

- Draw a normalized Bernoulli matrix $\Phi \in \mathbb{R}^{m,d}$ with rows $\varphi_1, \ldots, \varphi_m$.
- For $i = 1, \ldots, m$ set $\tilde{y}_i := \frac{f(h\varphi_i/2) - f(-h\varphi_i/2)}{h}$.
- Set $\tilde{a} := \Delta_{1,\varepsilon}(\tilde{y})$ with $\Delta_{1,\varepsilon}$ from ($\mathrm{P}_{1,\varepsilon}$).
- Set $\hat{a} := -\tilde{a}/\|\tilde{a}\|_2$.
- Set $\hat{g}(t) := f(\hat{a}(1 - \sqrt{t}))$ and $\hat{f}(x) := \hat{g}(\|\hat{a} - x\|_2^2)$ for $t \in [0, 4]$ and $x \in B_2^d$.

Output: Approximation $\hat{f}$.

The following theorem summarizes the performance of Algorithm E:

Theorem 6.15. *Using the notation of Algorithm E, for some constants c', C, C'' and any integer $s \geq C'' m/\log(d/m)$ with high probability, it holds*

$$\|f - \hat{f}\|_\infty \leq c_0(2\|a - \hat{a}\|_2 + \|a - \hat{a}\|_2^2),$$

where

$$\|a - \hat{a}\|_2 \leq \frac{2\,\mathrm{err}(a, \hat{a})}{2g'(1) - \mathrm{err}(a, \hat{a})} \quad \textit{and} \quad \mathrm{err}(a, \hat{a}) := Cg'(1)\frac{\sigma_s^1(a)}{\sqrt{s}} + c'h,$$

provided that $2g'(1) - \mathrm{err}(a, \hat{a})$ is positive.

Proof. The statement follows as a combination of (6.55) with (6.48). □

Remark 6.16. Once we have the approximation scheme described in Algorithm E using techniques from compressed sensing, we can easily extend it to an approximation scheme with noisy measurements with i.i.d. Gaussian noise by replacing the $\ell_{1,\varepsilon}$-minimizer by the Dantzig selector.

6.4 Numerical Experiments

In this section we test the numerical performance of the Algorithms A-E presented in this chapter. All the approximation schemes A-E start looking for a good recovery $\hat{a}$ of the ridge vector a and, afterwards, the problem is reduced to finding an approximation $\hat{g}$ of the univariate ridge profile g to finally obtaining a uniform approximation $\hat{f}$ of f. Furthermore, the difference between $\hat{f}$ and f is bounded by the corresponding difference between $\hat{a}$ and a, so, consequently, in the numerical examples we also only focus on the recovery of $\hat{a}$.

The basis pursuit (P_1) can be recast as a linear program, cf. Lemma 3.4, so for the implementation we have used the MATLAB command `linprog`. For the implementation of the $\ell_{1,\varepsilon}$-minimizer ($\mathrm{P}_{1,\varepsilon}$) we have used the MATLAB based toolbox `cvx`, and for an implementation of the Dantzig selector (3.29) we have used the so-called ℓ_1-MAGIC implementation, which is available at the web page `https://statweb.stanford.edu/~candes/l1magic/`.

For the numerical test we have done each calculation $n = 100$ times and then we have taken the average.

6.4.1 Ridge Functions on Cubes

In this section we test the numerical performance of the Algorithms A, B and B'.

Algorithm A

Figure 6.1 shows the dependency of Algorithm A on the step size h for different ridge profiles. First, let us observe that the error of reconstruction converges to zero if h tends to zero, which is in a good agreement with Theorem 6.5, describing the performance of Algorithm A. But also reasonable step sizes of h, e.g., $h = 0.2$, imply relatively small errors.

Furthermore, we observe that the approximation rapidly improves with growing dimension, which is explained by the concentration of measure phenomenon.

Finally, we observe that the order of the error depends on the ridge profile g. This can be explained by the derivatives of the particular ridge profiles. The second derivative of the first profile $g(t) = \tanh(t)$ vanishes at 0, so the first order differences (6.9) approximate the first order derivative quite accurately. The second derivative of the other two ridge profiles $g(t) = \tanh(t-1)$ and $g(t) = (1+\exp(-t))^{-1}$ does not vanish at 0, which leads to worse, but still surprisingly small approximation errors.

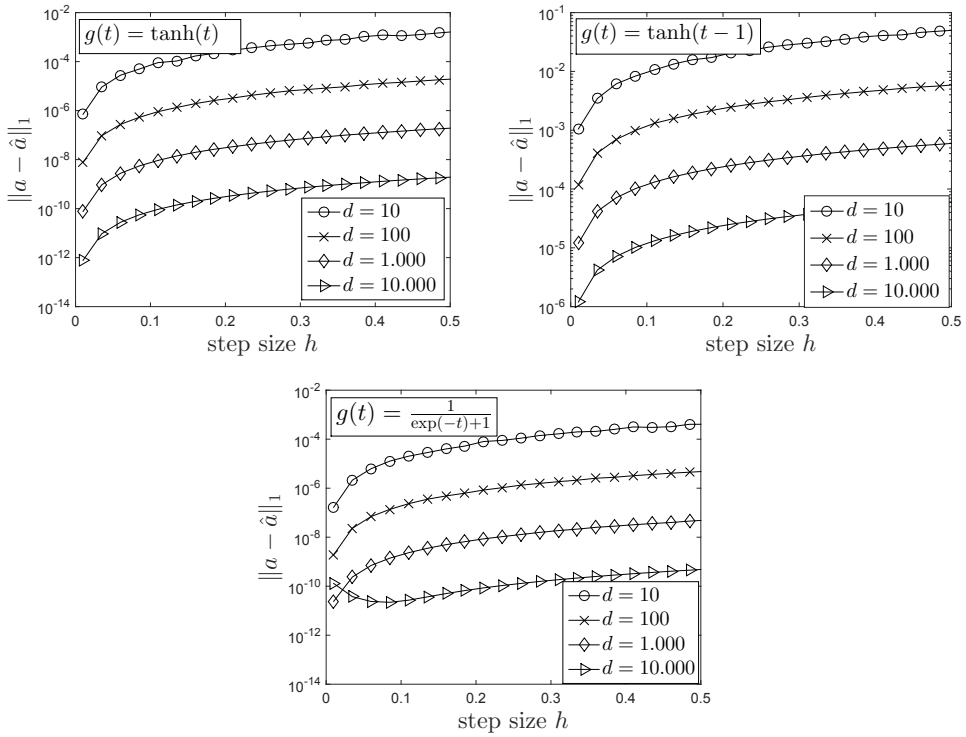

Figure 6.1: Dependency of Algorithm A on the step size h with respect to different ridge profiles $g(t) \in \{\tan(t), \tan(t-1), (1+\exp(-t))^{-1}\}$. For different values of $d \in \{10, 100, 1000, 10000\}$ and values of h ranging from 0 to 0.5, we generated a signal $\tilde{a} \in \mathbb{R}^d$ using the MATLAB command **randn**. Afterwards, we set $a = \tilde{a}/\|\tilde{a}\|_1$ and we run Algorithm A to obtain the approximation $\hat{a}$ of a. Then we calculated their difference $\|a - \hat{a}\|_1$.

Algorithms B and B'

1) Dependency on the Step Size h.

Figure 6.2 shows the dependency of the Algorithms B and B' on the step size h for two different ridge profiles $g(t) = \tanh(t-1)$ and $g(t) = (1+\exp(-t))^{-1}$. Here the sparsity level $s = 5$ is fixed and we considered the three different pairs $(d, m) \in \{(100, 40), (1000, 60), (10000, 80)\}$. Note that the amount of measurements m is quite small compared to the underlying dimension d. This means, with the additional assumption of sparsity we have reduced the number of measurements quite heavily, compared to Algorithm A, which needs $m = d$. Nevertheless, we still obtain reasonable approximation errors for small values of h. For the second profile

$g(t) = (1+\exp(-t))^{-1}$ the Algorithm B outperforms the Algorithm B', but this can be explained by the non-optimal choice of $\varepsilon = h/\sqrt{m}$.

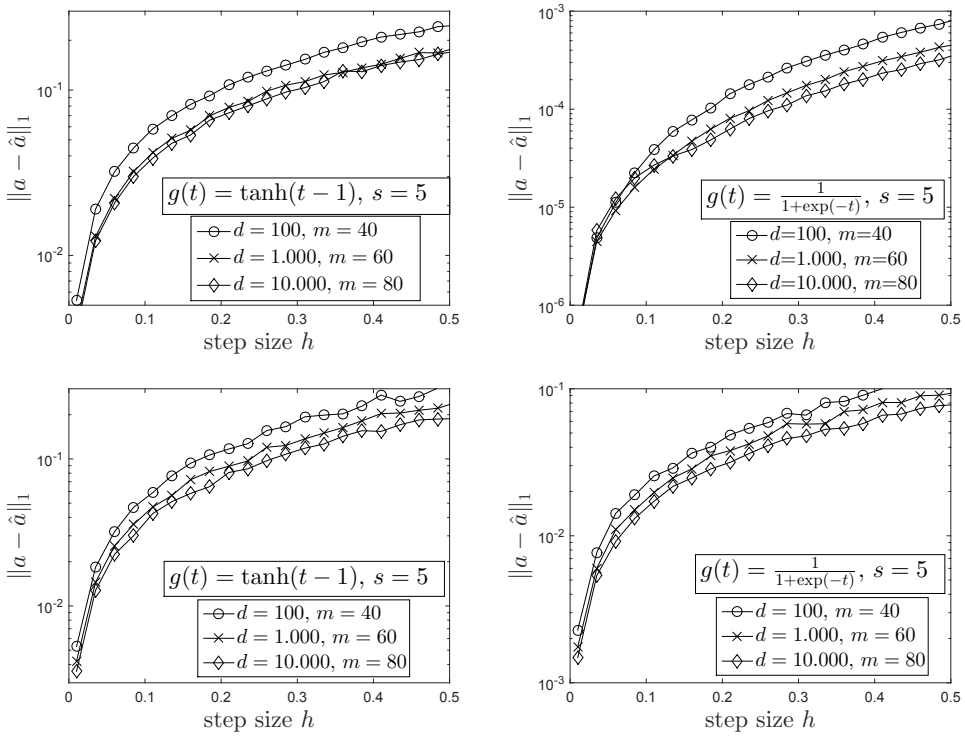

Figure 6.2: Dependency of Algorithm B (top) and Algorithm B' (bottom) on the step size h with respect to different ridge profiles $g(t) \in \{\tanh(t), (1+\exp(-t))^{-1}\}$. For a fixed sparsity level $s = 5$, pairs $(d, m) \in \{(100, 40), (1000, 60), (10000, 80)\}$ and step sizes h varying from 0 to 0.5, we chose an s-sparse vector $\tilde{a} \in \mathbb{R}^d$ at random using the MATLAB command **sprandn**. Afterwards, we set $a = \tilde{a}/\|\tilde{a}\|_1$ and ran Algorithm B and Algorithm B' with the additional parameter $\varepsilon = h/\sqrt{m}$ to obtain the approximations $\hat{a}$. Then we plotted the error of reconstruction $\|a - \hat{a}\|_1$.

2) Dependency on the Amount of Measurements m.

Figure 6.3 shows the dependency of the Algorithms B and B' on the amount of measurements m and the underlying dimension d, where the sparsity $s = 10$, the step size $h = 0.1$ and the ridge profile $g(t) \in \{\tanh(t-1), (1+\exp(-t))^{-1}\}$ are fixed.

We observe that the number of measurements only has to grow logarithmically in the dimension d, if we want to guarantee good approximation results with high probability. Further, we observe a phase transition between configurations where the approximation works and where the approximations are quite bad. This is in accordance with the theory of compressed sensing, where we also have already observed and discussed the phase transition phenomenon in Figure 3.2.

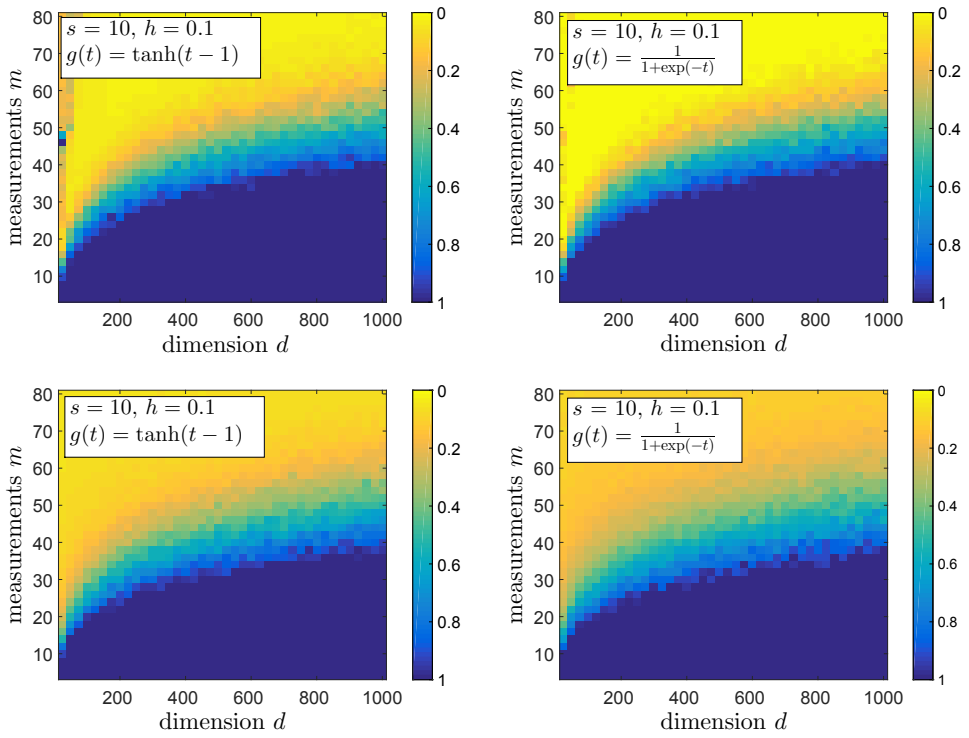

Figure 6.3: Dependency of Algorithm B (top) and of Algorithm B' (bottom) on m and d with respect to different ridge profiles $g(t) \in \{\tanh(t-1), (1+\exp(-t))^{-1}\}$. For different values of m ranging from 0 to 80 and values of d ranging from 0 to 1000 we drew an s-sparse vector $a \in \mathbb{R}^d$ at random and ran the Algorithms B and B' to obtain the reconstruction $\hat{a}$. Afterwards, we calculated and plotted the difference $\|a - \hat{a}\|_1$.

6.4.2 Approximation of Ridge Functions from Noisy Measurements

1) Dependency on the Step Size h.

Figure 6.4 shows the dependency of Algorithm C on the step size h, where we fixed the ridge profile $g(t) \in \{\tanh(t-1), (1+\exp(-t))^{-1}\}$, the noise level $\sigma \in \{0.01, 0.05, 0.001\}$ and the dimension $d = 500$, the sparsity $s = 5$ and the number of measurements $m = 80$.

When taking the first order differences, the noise level σ gets amplified by the factor $1/h$. Hence, if we choose h very small, the noise increases and the approximation completely fails. On the other hand, if we choose h very large, the approximation of the first derivative by first order differences deteriorates as well, as we have already discussed in Remark 6.12. Therefore, there might be an optimal value for h, which we can also guess from the upper left image of Figure 6.12.

The bottom image of Figure 6.12 shows a modification of Algorithm C, where we just used the basis pursuit for the reconstruction, instead of the Dantzig selector. This image can be understood as a demonstration of the success of the Dantzig selector, since the approximation completely fails even if the noise level $\sigma = 0.001$ is quite small.

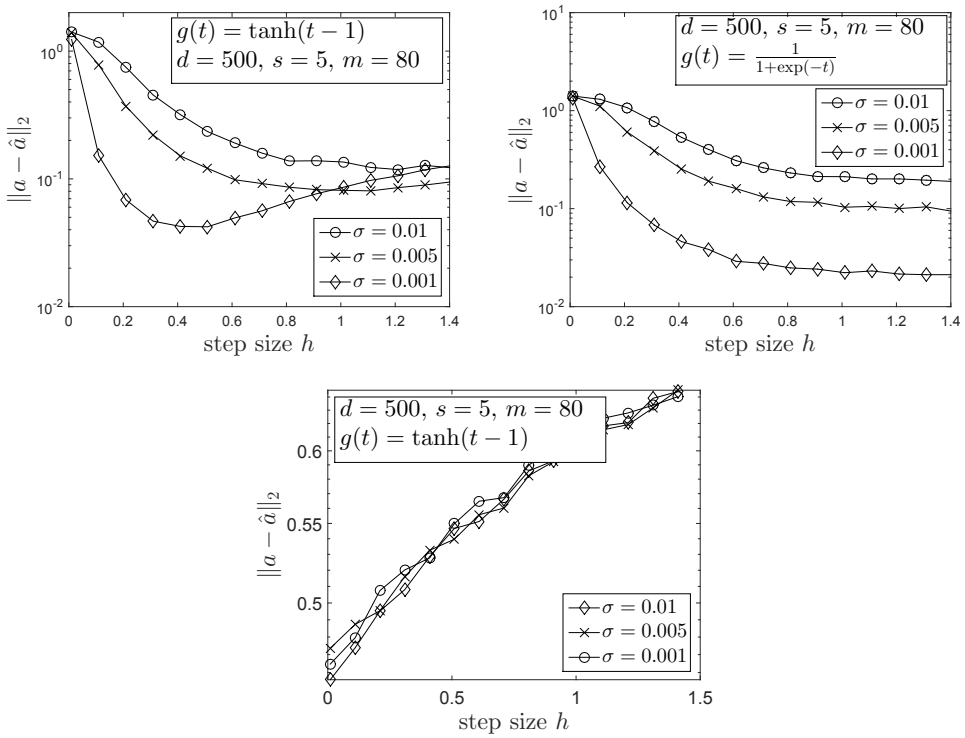

Figure 6.4: Dependency of Algorithm C on the step size h for different ridge profiles $g(t) \in \{\tanh(t-1), (1+\exp(-t))^{-1}\}$ (top) and the performance of the basis pursuit in the presence of Gaussian noise (bottom) for the ridge profile $g(t) = \tanh(t-1)$. Beside the ridge profile, we fixed the sparisty level $s = 5$, the dimension $d = 500$ and the number of measurements $m = 80$. For the noise levels $\sigma \in \{0.001, 0.005, 0.01\}$ and values of h ranging from 0 to 1.4 we drew an s-sparse signal $\tilde{a} \in \mathbb{R}^d$ at random using the MATLAB command **sprandn** and afterwards set $a = \tilde{a}/\|\tilde{a}\|_2$. Then we ran Algorithm C with the Dantzig selector and the choice $\lambda_d = \sqrt{2\log(d)}$ (top) and the basis pursuit (bottom) to obtain the approximation $\hat{a}$ of a.

2) Dependency on the Amount of Measurements m.

In Figure 6.5 we tested the dependency of Algorithm C on the number of measurements m and the dimension d, where we fixed the noise level $\sigma = 0.005$, the ridge profile $g(t) = \tanh(t-1)$, the sparsity $s = 10$ and the step size $h = 1$. We observed a similar phase transition as for Algorithm B.

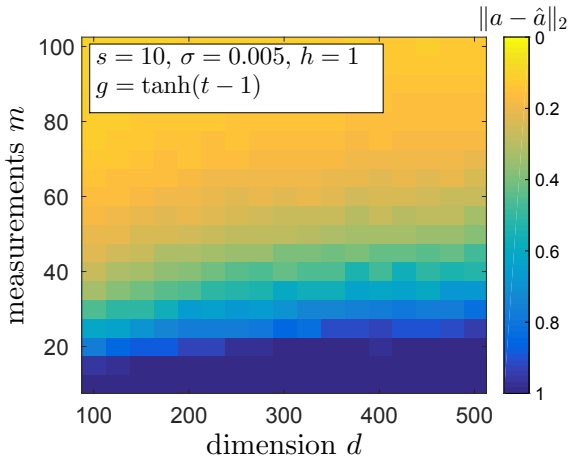

Figure 6.5: Dependency of Algorithm C on the dimension d and the number of measurements m. We fixed the noise level $\sigma = 0.005$, the step size $h = 1$, the sparsity $s = 10$ and the ridge profile $g(t) = \tanh(t - 1)$. Then, for different values of m ranging from 0 to 100 and values for d ranging from 0 to 500 we generated an s-sparse signal $\tilde{a} \in \mathbb{R}^d$ using the MATLAB command **sprandn**. Afterwards, we set $a = \tilde{a}/\|\tilde{a}\|_2$ and used Algorithm C to obtain the approximation $\hat{a}$ of a. Then we calulated their difference $\|a - \hat{a}\|_2$.

6.4.3 Approximation of Translated Radial Functions

Figure 6.6 illustrates the performance of Algorithm D (top) and Algorithm E (bottom left) in dependency on the step size h, and the performance of the approximation of translated radial functions from noisy measurements using the Dantzig selector (bottom right), as discussed in Remark 6.16. In particular, here we have to highlight that we used the profile $g(t) = -1/t$. In this case the radial function $f(x) = g(\|a - x\|_2^2)$ has a singularity in a, hence, a uniform approximation of f is out of reach. However, as we have already discussed in Remark 6.14, in this case we are still able to recover a, that is, the position of the singularity.

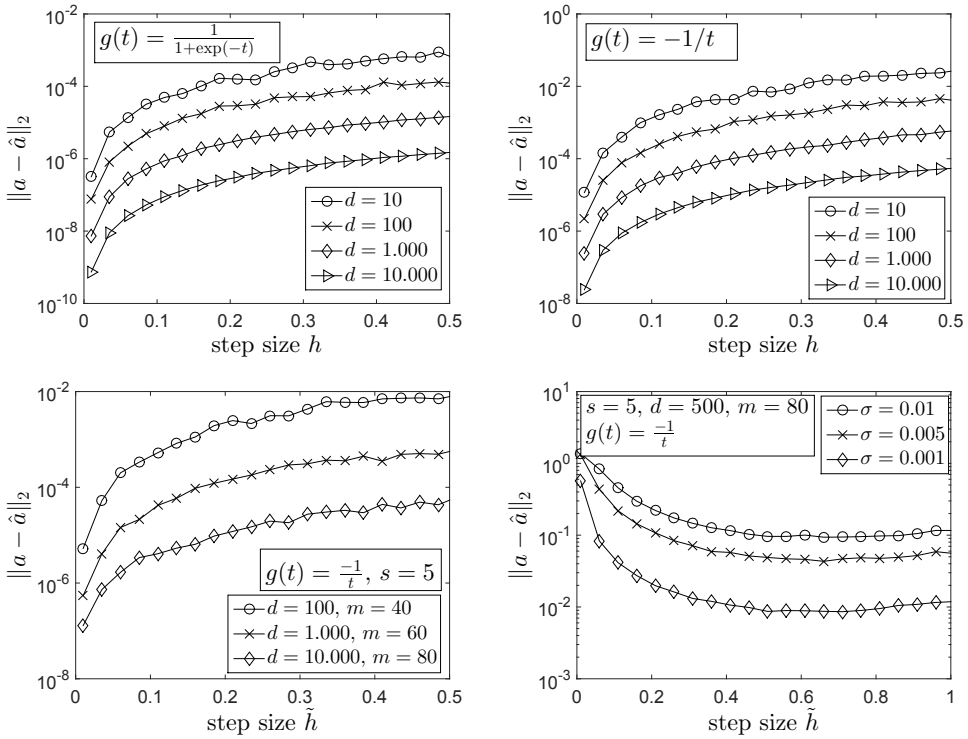

Figure 6.6: Dependency of Algorithm D on the step size h for different profiles $g(t) \in \{(1 + \exp(-t))^{-1}, -1/t\}$ (top), the dependency of Algorithm E on the step size $\tilde{h}$ for the ridge profile $g(t) = -1/t$ (bottom left) and the performance of approximation of radial functions from noisy measurements (bottom right). Note that the approximation errors for Algorithm E (bottom) are plotted against the modified step size $\tilde{h} = h/2 \cdot \sqrt{d/m}$.

Chapter 7

Conclusion and Outlook

In this thesis we have studied different aspects of compressed sensing. The first main result of this thesis is given in Chapter 4. Here we have proven the last missing part of Carl's inequality namely for Gelfand numbers on quasi-Banach spaces. This in particular can be used to fill a gap in Donoho's argument estimating the compressive n-width.

In the further parts of this thesis we demonstrated that the ideas of compressed sensing for the recovery of sparse signals in high dimensions from linear measurements can also be used if the measurement process is disturbed by a certain nonlinearity.

In Chapter 5 we have analyzed the performance of the ℓ_1-support vector machine for the problem of 1-bit compressed sensing, that is, for the recovery of effectively sparse signals from binary measurements. In order to obtain good recovery results, we have shown that the number of measurements m has to grow logarithmically in the dimension d and almost linear in the sparsity s, which is actually the same rate we would need for the recovery from linear measurements using basic recovery algorithms from the field of compressed sensing.

However, we have not discussed the optimality on the other parameters r and ε, which we left out for further research. Furthermore, we have only shortly discussed the recovery from noisy measurements. Here, we have given an idea to show an appropriate recovery result, once the model of the noise is fixed.

As a third open problem let us mention the particular choice of the measurement vectors. We have always chosen the measurement vectors to be i.i.d. Gaussian, but recovery results for other distributions are of interest. Furthermore, it would also be beneficial to have recovery results if the measurement vectors are not independent of each other.

Another open question for further research we would like to highlight is the choice of the particular loss function and also the set of constraints. To improve the performance of the ℓ_1-SVM, we have suggested the $\ell_{1,2}$-SVM by adding an additional penalty term. However, it might be possible to further improve the performance, by suggesting other constraints, but also by considering other loss functions than the hinge loss.

In Chapter 6 we have answered some open questions in the theory of ridge functions. First, we have discussed the approximation of ridge functions defined on the unit cube $[-1,1]^d$ instead of the Euclidean unit ball B_2^d. Afterwards, we

have discussed the approximation of ridge functions from noisy measurements by using the Dantzig selector. The last contribution of this chapter is given by the approximation scheme for translated radial functions, where we demonstrated that one can transfer the approximation idea of ridge functions to other, but similar function classes.

Let us highlight some open questions which we left out for further research. First, one can think of the approximation of ridge functions defined on a general convex domain. By using the so-called *Minkowski functional*, we hope to transfer the approximation scheme for the recovery of ridge functions defined on the unit cube to the more general case.

Second, we have only considered the approximation of ridge functions with univariate ridge profile g. Hence, an interesting open question is the recovery of ridge functions f defined on the unit cube, which are of the form $f(x) = g(Ax)$ for some ridge profile $g\colon \mathbb{R}^k \to \mathbb{R}$ and an unknown (low-rank) matrix $A \in \mathbb{R}^{k,d}$.

Furthermore, also a combination of the 1-bit compressed sensing problem and the approximation of ridge functions may be of interest, i.e., the recovery of the sparse signal x from measurements of the form $y = \operatorname{sign}(g(\langle a, x\rangle))$. Here the ridge profile g can be understood as a generalization of the classification model, which means, the sample points are not separated by a hyperplane anymore, but by the level set $\{g = 0\}$.

Last, one can also discuss the approximation of more general functions of a similar type. Here we have already discussed translated radial functions, whose function values depend on the distance to a certain point. This can be further generalized by considering functions, whose values are given by the distance to a certain manifold, for instance, to a k-dimensional plane.

Bibliography

[1] T. Aoki. Locally bounded linear topological spaces. Proc. Imp. Acad. Tokyo 18 (1942), pp. 588–594.

[2] E. Ardestanizadeh, M. Cheraghchi, A. Shokrollahi. Bit precision analysis for compressed sensing. In IEEE International Symposium on Information Theory (ISIT) (2009)

[3] E. Arias-Castro, Y. Eldar. Noise Folding in Compressed Sensing. IEEE Signal Proc. Letters 18 (2011), no. 8, pp. 478–481.

[4] R. Balan, P. Casazza, D. Edidin. On signal reconstruction without phase. Appl. Comput. Harmon. Anal. 20 (2006), no. 3, pp. 345–356.

[5] A. S. Bandeira, J. Cahill, D. G. Mixon, A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. Appl. Comput. Harmon. Anal. 37 (2014), no. 1, pp. 106–125.

[6] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices. Constr. Approx. 28 (2008), no. 3, pp. 253–263.

[7] J. Bastero, J. Bernués, A. Peña. An extension of Milman's reverse Brunn-Minkowski inequality. Geom. Func. Anal. 5 (1995), no. 3, pp. 572–581.

[8] T. Blumensath, M. Davis. Iterative Hard Thresholding for Compressed Sensing. Appl. Comput. Harmon. Anal. 27 (2009), no. 3, pp. 265–274.

[9] H. Boche, R. Calderbank, G. Kutyniok, J. Vybíral. A survey of compressed sensing. In Compressed Sensing and its Applications, Birkhäuser, Boston (2015).

[10] P. Boufounos, R. Baraniuk. 1-Bit Compressive Sensing. In Proc. 42nd Annu. Conf. Inf. Sci. Syst. (2008), Princeton, NJ, pp. 16–21.

[11] S. Boyd, L. Vandenberghe. Convex Optimization. Cambridge University Press, Cambridge (2004).

[12] P. S. Bradley, O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In Proceedings of the 15th International Conference on Machine Learning (1998), pp. 82–90.

[13] L. M. Brekhovskikh. Waves in Layered Media. Series: Applied Mathematics and Mechanics 16 (1980), Academic Press, New York-London.
[11] P. Bühlmann, S. van de Geer. Statistics for High-Dimensional Data - Methods, Theory and Applications. Springer, Heidelberg (2011).

[14] E. Candès. The restricted isometry property and its implications for compressed sensing. C. R. Math. Acad. Sci. Paris, Ser. 346 (2008), no. 9–10, pp. 589–592.

[15] E. Candès. Harmonic analysis of neural networks. Appl. Comput. Harmon. Anal. 6 (1999), no. 2, pp. 197–218.

[16] E. Candès, D. L. Donoho. Ridgelets: a key to higher-dimensional intermittency? R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. 357 (1999), no. 1760, pp. 2495–2509.

[17] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory 52 (2006), no. 2, pp. 489–509.

[18] E. J. Candès, T. Strohmer, V. Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. Comm. Pure Appl. Math. 66 (2013), no. 8, pp. 1241–1274.

[19] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory 52 (2006), no. 12, pp. 5406–5425.

[20] E. Candès, T. Tao. The Dantzig Selector: statistical estimation when p is much larger than n. Ann. Stat. 35 (2007), no. 6, pp. 2313–2351.

[21] B. Carl. Entropy numbers, s-numbers, and eigenvalue problems. J. Funct. Anal. 41 (1981), no. 3, pp. 290–306.

[22] B. Carl, A. Pajor. Gel'fand numbers of operators with values in a Hilbert space. Invent. Math. 94 (1988), no. 3, pp. 479–504.

[23] B. Carl, I. Stephani. Entropy, compactness and the approximation of operators. Cambridge University Press, Cambridge (1990).

[24] R. Chartrand, R. Saab, Ö. Yilmaz. Stable sparse approximation via nonconvex optimization. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2008), Las Vegas, Nevada.

[25] S. S. Chen, D. L. Donoho, M. A. Saunders. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1998), no. 1, pp. 33–61.

[26] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, D. Picard. Capturing ridge functions in high dimensions from point queries. Constr. Approx. 35 (2012), no. 2, pp. 225–243.

[27] A. Cohen, W. Dahmen, R. DeVore. Compressed Sensing and best k-term approximation. J. Amer. Math. Soc. 22 (2009), no. 1, pp. 211–231.

[28] C. Cortes, V. Vapnik. Support-vector networks. Mach. Learn. 20 (1995), pp. 273–297.

[29] F. Cucker, D. X. Zhou. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007).

[30] S. Dasgupta, A. Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. Random Structures Algorithms 22 (2003), no. 1, pp. 60–65.

[31] I. Daubechies. Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics 61 (SIAM) (1992), Philadelphia.

[32] R. DeVore, G. G. Lorentz. Constructive Approximation. in: Grundlehren der Mathematischen Wissenschaften 303 (1993), Springer, Berlin.

[33] R. DeVore, G. Petrova, P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. Constr. Approx. 33 (2011), no. 1, pp. 125–143.

[34] R. DeVore, G. Petrova, P. Wojtaszczyk. Instance-optimality in Probability with an ℓ_1-Minimization Decoder. Appl. Comput. Harmon. Anal. 27 (2009), no. 3, pp. 275–288.

[35] D. L. Donoho. Compressed Sensing. IEEE Trans. Inform. Theory 52 (2006), no. 4, pp. 1289–1306.

[36] D. L. Donoho. Sparse components of images and optimal atomic decompositions. Constr. Approx. 17 (2001), no. 3, pp. 353–382.

[37] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. Proc. Natl. Acad. Sci. 100 (2003), no. 5, pp. 2197–2202.

[38] D. L. Donoho, M. Lustig, J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. Magnetic resonance in medicine 58 (2007), no. 6, pp. 1182–1195.

[39] D. E. Edmunds, H. Triebel. Function spaces, entropy numbers and differential operators. Cambridge Tracts in Mathematics 120 (1996), Cambridge University Press, Cambridge.

[40] Y. C. Eldar, G. Kutyniok. Compressed sensing: Theory and Applications. Cambridge University Press, Cambridge (2012).

[41] A. Flinth, G. Kutyniok. PROMP: A Sparse Recovery Approach to Lattice-Valued Signals. Submitted (2016).

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

[43] S. Foucart, M.J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. 26 (2009), no. 3, pp. 395–407.

[44] S. Foucart, A. Pajor, H. Rauhut, T. Ullrich. The Gelfand width of ℓ_p-balls for $0 < p \leq 1$. J. Compl. 26 (2010), no. 6, pp. 629–640.

[45] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York (2013).

[46] J. Friedmann, T. Hastie, S. Rosset, R. Tibshirani, J. Zhou. Discussion of boosting papers. Ann. Statist. 32 (2004), no. 1, pp. 102–107.

[47] M. Gerhold. Entropy, approximation and Kolmogorov numbers on quasi-Banach spaces. Bachelor thesis, Friedrich-Schiller University, Jena (2011).

[48] C. Güntürk, M. Lammers, A. Powell, R. Saab, Ö. Yilmaz. Sigma delta quantization for compressed sensing. IEEE 44th Annual Conference on Information Sciences and Systems (CISS) (2010).

[49] K. Guo, G. Kutyniok, D. Labate. Sparse multidimensional representations using anisotropic dilation and shear operators. Wavelets and Splines: Athens (2005), pp. 189–201, Nashboro Press, Brentwood.

[50] K. Guo, D. Labate. Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal, 39 (2007), no. 1, pp. 298–318.

[51] A. Gupta, R. Nowak, B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. IEEE International Symposium on Information Theory (ISIT) (2010).

[52] W. Härdle, L. Simar. Applied multivariate statistical analysis. Springer, Berlin (2003).

[53] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001).

[54] T. Hastie, R. Tibshirani, M. Wainwright. Statistical Learning with Sparsity: The Lasso and Generalizations. Chapmann and Hall/CRC (2015).

[55] T. Hemant, V. Cevher. Active learning of multi-index function models. In Advances in Neural Information Processing Systems 25 (2012), pp. 1475–1483.

[56] M. Hilario, A. Kalousis. Approaches to dimensionalty reduction in proteomic biomarker studies. Brief Bioinform 9 (2008), no.2, pp. 102–118.

[57] A. Hinrichs, A. Kolleck, J. Vybíral. Carl's inequality for quasi-Banach spaces. J. Funct. Anal. 271 (2016), no. 8, pp. 2293–2307.

[58] A. Hinrichs, E. Novak, H. Woźniakowski. The Curse of Dimensionalty for Monotone and Convex Functions of Many Variables. J. Approx. Theory 163 (2011), no. 8, pp. 955–965.

[59] J. L. Horowitz. Semiparametric and nonparametric methods in econometrics 692, Springer, New York (2009).

[60] J. Huang, T. Zhang. The Benefit of Group Sparsity. Ann. Statist. 38 (2010), no. 4, pp. 1978–2004.

[61] L. Jacques, J. Laska, P. Boufounos, R. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Trans. Inform. Theory 59 (2013), no. 4, pp 2082–2102.

[62] W. B. Johnson, J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. Contemp. Math 26 (1984), pp. 189–206.

[63] N. J. Kalton. Basic sequences in F-spaces and their applications. Proc. Edinburgh Math. Soc. (2) 19 (1974/75), no. 2, pp. 151–167.

[64] N. J. Kalton. Quasi-Banach Spaces. Handbook of the Geometry of Banach Spaces 2 (2003), pp. 1099–1130, North-Holland, Amsterdam.

[65] B. Kashin, V. Temlyakov. A remark on the problem of compressed sensing. Mat. Zametki 82 (2007), no. 6, pp. 829–837.

[66] S. Keiper, G. Kutyniok, D. G. Lee, G. E. Pfander. Compressed Sensing for Finite-Valued Signals. Submitted (2016), available at arXiv:1609.09450.

[67] A. Kolleck, J. Vybíral. Non-asymptotic Analysis of ℓ_1-norm Support Vector Machines. Submitted (2016), available at arXiv:1509.08083.

[68] A. Kolleck, J. Vybíral. On some aspects of approximation of ridge functions. Jour. of Approx. Theo. 194 (2015), pp. 35–61.

[69] T. Kühn. A lower estimate for entropy numbers. J. Approx. Theory 110 (2001), no. 1, pp. 120–124.

[70] G. Kutyniok, D. Labate. Shearlets: Multiscale analysis for multivariate data. Springer (2012).

[71] J. Li, Y. Yingmin, D. Junping, Y. Fashan. A new support vector machine for microarray classification and adaptive gene selection. In 2009 IEEE American Control Conference (2009), pp. 5410-5415.

[72] V. Y. Lin, A. Pinkus. Fundamentality of ridge functions. J. Approx. Theory 75 (1993), no. 3, pp. 295–311.

[73] A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. Adv. Math. 195 (2005), no. 2, pp. 491–523.

[74] M. Ledoux. The Concentration Of Measure Phenomenon. Math. Surveys Monogr. 89 (2001), Amer. Math. Soc.

[75] M. Ledoux, M. Talagrand. Probability in Banach Spaces: Isoperimetry and Processes. Springer, Berlin (1991).

[76] B. P. Logan, L. A. Shepp. Optimal reconstruction of a function from its projections. Duke Math. J. 42 (1975), no. 4, pp. 645–659.

[77] J. Ma. Stable reconstructions for the analysis formulation of ℓ^p-minimization using redundant systems. Submitted (2016), available at arXiv:1509.05512.

[78] O. L. Mangasarian. Arbitrary-norm separating plane. Oper. Res. Lett. 24 (1999), no. 1–2, pp. 15–23.

[79] P. Mathé. s-Numbers in Information-Based Complexity. J. Compl. 6 (1990), no. 1, pp. 41–66.

[80] S. Mayer, T. Ullrich, J. Vybíral. Entropy and sampling numbers of classes of ridge functions. Constr. Approx. (2015), no. 2, pp. 231–264.

[81] D. Needell, J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. 26 (2009), no. 3, pp. 301–321.

[82] E. Novak. Optimal recovery and n-widths for convex classes of functions. J. Approx. Theory 80 (1995), no. 3, pp. 390–408.

[83] E. Novak, H. Woźniakowski. Approximation of infinitely differentiable multivariate functions is intractable. J. Complexity 25 (1009), no. 4, pp. 398–404.

[84] E. Novak, H. Woźniakowski. Tractability of Multivariate Problems, Volume I: Linear Information. EMS Tracts in Mathematics, Vol. 6, Eur. Math. Soc. Publ. House, Zürich (2008).

[85] E. Novak, H. Woźniakowski. Tractability of Multivariate Problems, Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, Vol. 12, Eur. Math. Soc. Publ. House, Zürich (2010).

[86] E. Novak, H. Woźniakowski. Tractability of Multivariate Problems, Volume III: Standard Information for Operators. EMS Tracts in Mathematics, Vol. 18, Eur. Math. Soc. Publ. House, Zürich (2012).

[87] H. Ohlsson, Y. C. Eldar. On conditions for uniqueness in sparse phase retrieval. In IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) (2014), pp. 1841–1845.

[88] Y. Pati, R. Rezaiifar, P. Krishaprasad. Orthogonal Matching Pursuit: recursive function approximation with application to wavelet decomposition. In Asilomar Conf. on Signals, Systems and Comput. (1993).

[89] M. E. Pfetsch, A. Tillmann. The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing. IEEE Trans. Inform. Theory 60 (2014), no. 2, pp. 1248–1259.

[90] A. Pietsch. Operator ideals. North-Holland Publishing Co., Amsterdam-New York (1980).

[91] A. Pietsch. Eigenvalues and s-Numbers. Cambridge University Press, Cambridge (1987).

[92] A. Pinkus. N-widths in approximation theory. Springer, Berlin (1985).

[93] A. Pinkus. Approximation theory of the MLP model in neural networks. Acta Numer. 8 (1999), pp. 143–195.

[94] A. Pinkus. Approximating by ridge functions. Surface fitting and multiresolution methods (1996), pp. 279–292, Vanderbilt Univ. Press, Nashville.

[95] A. Pinkus. N-widths and optimal recovery. Proc. Symp. Appl. Math. 36. (1986), pp. 51–66.

[96] Y. Plan, R. Vershynin. Robust 1-Bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach. IEEE Trans. Inform. Theory 59 (2013), no. 1, pp 482–494.

[97] Y. Plan, R. Vershynin. The generalized Lasso with non-linear observations. IEEE Trans. Inform. Theory 62 (2016), no. 3, pp. 1528–1537.

[98] Y. Plan, R. Vershynin, E. Yudovina. High-dimensional estimation with geometric constraints. Information and Inference (2016), pp. 1–40.

[99] S. Rolewicz. On a certain class of linear metric spaces. Bull. Acad. Polon. Sci. Cl. III. 5 (1957), pp. 471–473.

[100] M. Rossi, A. M. Haimovich, Y. C. Eldar. Spatial compressive sensing for MIMO radar. IEEE Trans. Signal Process. 62 (2014), no. 2, pp. 419–430.

[101] K. Schnass, J. Vybíral. Compressed learning of high-dimensional sparse functions. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2011), pp. 3924–3927.

[102] C. Schütt. Entropy numbers of diagonal operators between symmetric Banach spaces. J. Approx. Theory 40 (1984), no. 2, pp. 121–128.

[103] M. Stojnic, F. Parvaresh, B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. IEEE Trans. Signal Process. 57 (2009), no. 8, pp. 3075–3085.

[104] I. Steinwart. Support vector machines are universally consistent. J. Complexity 18 (2002), no. 3, pp. 768–791.

[105] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. IEEE Trans. Inform. Theory 51 (2005), no. 1, pp. 128–142.

[106] I. Steinwart, A. Christmann. Support Vector Machines. Springer, Berlin (1995).

[107] R. Tibshirani. Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B 58 (1996), no. 1, pp. 267–288.

[108] V. Vapnik, A. Chernovenkins. A note on one class perceptrons. In Automation and Remote Control 25 (1964), no. 1.

[109] J. Vybíral. Widths of embeddings in function spaces. J. Complexity 24 (2008), no. 4, pp. 545–570.

[110] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. IEEE Trans. Inform. Theory 55 (2009), no. 12, pp. 5728–5741.

[111] Y. Wang, Z. Xu. Phase retrieval for sparse signals. Appl. Comput. Harmon. Anal. 37 (2014), no. 3, pp. 531–544.

[112] L. Wang, J. Zhu, H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics 24 (2008), no. 3, pp. 412–419.

[113] L. Wang, J. Zhu, H. Zou. The doubly regularized support vector machine. Statist. Sinica 16 (2006), no. 2, pp. 589–615.

[114] P. Wojtaszczyk. Complexity of approximation of functions of few variables in high dimensions. J. Complexity 27 (2011), no. 2, pp. 141–150.

[115] H. Zhan, J. Ahn, X. Lin, C. Park. Gene selection using support vector machines with non-convex penalty. Bioinformatics 22 (2006), no. 1, pp. 88–95.

[116] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani. 1-norm Support Vector Machines. In Proc. Advances in Neural Information Processing Systems 16 (2004), pp. 45–56.

[117] H. Zou, T. Hastie. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005), no. 2, pp. 301–320.

[118] H. Zou, T. Hastie, R. Tibshirani. Sparse principal component analysis. J. Comput. Graph. Statist. 15 (2006), no. 2, pp. 265–286.