Compressed Sensing, sparsity and related topics

Dissertation in Mathematics

submitted by

Michael Sandbichler

to the Faculty of Mathematics, Computer Science
and Physics of the University of Innsbruck

universität
innsbruck

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Advisor: Univ.-Prof. Dr. Markus Haltmeier
Co-Advisors: Univ.-Prof. Dr. Felix Krahmer and
Univ.-Prof. Dr. Alexander Ostermann

Innsbruck, 26th February 2018

Abstract

One of the major insights that enabled the development of many modern tools in signal processing was that most data that is encountered in practice admits some kind of sparse representation. This means, that given a suitable basis, a given signal can be approximated by a linear combination of only very few of the basis vectors. Sparsity will be a guiding theme of this thesis.

In the first part of this thesis, we will explore the application of modern signal processing techniques, namely compressed sensing and sparse recovery, to photoacoustic tomography. We will consider possible randomized measurement setups and derive recovery guarantees. Futher, we will take a look at transformations that sparsify the data while maintaining the derived recovery guarantees. In these works, we will use total variation minimization to recover the signals and in the subsequent chapter, we will provide an overview over total variation minimization to recover undersampled signals and in addition give recovery guarantees for subgaussian measurements. Finally, the last chapter is devoted to the study of how such sparse representations can be obtained. We will explore the possibility of learning sparsifying transformations directly from the data without having any additional knowledge about the class of signals. The only prior information we have is that there exists some operator that sparsifys the given data. We will derive algorithms to learn such operators and establish theoretical results.

Acknowledgements

I would like to express my deep gratitude to my advisors Univ.-Prof. Dr. Markus Haltmeier and Univ.-Prof. Dr. Felix Krahmer for their outstanding supervision and encouragement during the process of writing this thesis.

For providing a fresh batch of ideas when mine failed, for many confusing (but of course correct) explanations, fun coffee breaks and of course for providing the necessary funding during the second half of my PhD-studies, I am wholeheartedly thankful to my step-advisor Dr. Karin Schnass.

A big thanks also goes out to Dr. Christian Bargetz, who maybe too often had to be the first one to hear my overly excited explanations of something I thought I had understood, but of course hadn't.

Furthermore, I would like to express my thanks to all other colleagues at the Department of Mathematics at the University of Innsbruck as well as at the Research Unit M15 at the TU Munich for the great working atmosphere.

Apart from my professional environment, I would like to thank my family as well as my girlfriend for providing me with continuous love and support and my friends for often reminding me that life does not entirely consist of equations.

Contents

1 Introduction and Motivation

”The question is - what is the question?”

- HP Baxxter

Let us start our journey by taking a look at the title of this dissertation and define its ingredients. The first part of the title is ’Compressed Sensing’ [Don06a, CRT06a, FR13], which is an area of signal processing built on the insight that ’sparsity’, the second part of the title, is an invaluable prior for a large class of signals.

Perhaps one of the prime examples, in which sparse representations have led to significant improvements is the field of image compression. The JPEG and JPEG2000 standards both rely on transformations into bases that are well known to sparsify large classes of images - the discrete cosine transform (DCT) and waveletbases. By storing only the largest coefficients, often significant reduction of storage space can be achieved. With .jpeg-compression, file sizes of $1/10$ of the original can quite easily be achieved without significant compression artifacts. In Figure 1.1, the resulting images when choosing different levels of .jpeg-compression are depicted. One can clearly see that using 40% of the available coefficients still results in a high quality image. Using 10% of the coefficients already exhibits clear compression artifacts and when using only 5% of the image’s DCT coefficients the undersampling artifacts are really severe. Note however that one can still quite clearly make out the content of the image, although only very few nonzero coefficients are used. The square undersampling artifacts are produced, because .jpeg-compression performs the DCT on 8×8 blocks of the image and afterwards pieces them back together.

Let us define some of the terminology. When we use the term ’signal’, we will always assume that it is given as a vector $x \in \mathbb{R}^d$.

A signal is called sparse, if only few of its entries are nonzero. This means that, although typically the ambient dimension is large, the intrinsic dimension and also the information content of signals is small. Most of the time, sparsity will only be achieved after a suitable change of basis. Finding such a basis for a given class of data is a nontrivial task and subject to the field of dictionary learning [AEB06, Sch15b, Sch16, SQW15], see also Chapter 5.

If a signal is known to be sparse, its information content is much smaller than the Nyquist sampling rate would indicate. The Nyquist rate specifies the required rate at which equidistant

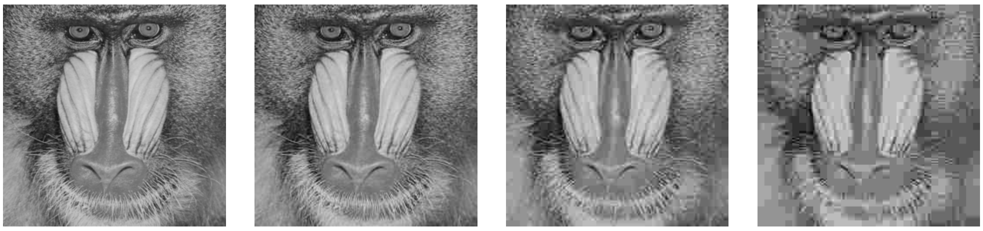

Figure 1.1: The Mandrill test image (1st) compressed using 40% (2nd), 10% (3rd) and 5% (4th) of the largest DCT coefficients.

samples of a bandlimited signal have to be taken, such that these samples determine the signal uniquely. 'Bandlimited' means that the Fourier transform of the signal is compactly supported.

It seems like a waste of resources to actually sample a sparse signal at the Nyquist rate - however there is a caveat. Because the locations of the nonzero entries are not known, simple equispaced sub-Nyquist sampling might fail to recover the signal. For example if the signal is very sparse, sub-Nyquist sampling will very likely return just the zero vector, because the sampling points will hardly ever coincide with the locations of the nonzero entries.

This indicates that regular subsampling is not the way to go for sparse signals and that it is necessary to use special sampling strategies in order to be able to do as few measurements as possible according to the signal's information content. When we use the term 'measurements', we mean a linear map $A \in \mathbb{R}^{m \times d}$, which takes a signal in d dimensions and produces a vector in $\mathbb{R}^m$. The goal is that $m \ll d$, while still being able to perfectly recover the signal from the measurements.

In Compressed Sensing, it has been shown that for s-sparse signals, $m = \mathcal{O}(s \log \frac{d}{s})$ measurements are enough to achieve perfect recovery in the noiseless setting and random measurement strategies have been used in order to achieve this optimal number of measurements [FRon].

From this point of view, namely we are given an underdetermined linear equation $Ax = y$ and a sparsity prior for x, we are solvinge an ill-posed problem. In such problems either the solution is not unique, does not exist or is not stable with respect to the data. In our case, the main problem is that the solution is not unique - but we know that the solution is sparse. If we now resort to finding the sparsest solution of the given underdetermined system of equations, we can hope to restore uniqueness of the solution. Of course this alone is not enough, we also need additional assumptions on the measurement matrix, but a bigger problem is that finding such a sparse solution is, in general, NP hard. The realization that such sparse recovery can also be done via simple convex regularization with the ℓ_1-norm (again, if certain conditions on the measurement matrix A are fulfilled) has initiated a large interest in sparsity-based signal processing during the last decade [DE03, Don06b]. In particular in the field of inverse problems this has sparked a lot of developments and improvements.

Inverse problems are, as their name suggests, the inverse question to a direct or forward problem. So when in a forward problem, a typical question would be to compute the solution of a differential equation given initial conditions and boundary data, the corresponding inverse problem would for example be: compute the initial conditions given the solution of the differential equation on the boundary. Naturally, inverse problems tend to be much more sensitive to noise and perturbations of the data. This is why regularization techniques are often necessary [EHN96, GHS08, GHS11]. Prominent examples of inverse problems are Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). In this thesis we will mostly focus on Photoacoustic Imaging, a hybrid imaging method combining the advantages of optical and ultrasonic imaging.

In Photacoustic Tomography, soft tissue is heated using pulsed electromagnetic radiation. This heating triggers expansion of the tissue, which subsequently emits a sound wave. These sound waves are then measured outside the sample. In Figure 2.1 this principle is illustrated. This means that the governing equation for this problem is the wave equation describing the propagation of pressure waves through the tissue. Mathematically, the setting is as follows,

$$\partial_t^2 p(t,x) = c(x) \cdot \Delta p(t,x), \text{ such that } p(0,x) = f(x) \text{ and } \partial_t p(0,x) = 0, \tag{1.1}$$

where f is compactly supported (typically within a ball of radius R). The speed of sound c may vary in space, but we exclude the case where it can vary in time. In Chapter 2, we will just consider the case of constant speed of sound, where in Chapter 3 also spatially varying speed of sound is admissible. The goal of photoacoustic tomography is to reconstruct the initial function f, given measurements at some hypersurface.

The standard way of performing such a reconstruction is via filtered backprojection [HSS05, HSB$^+$07, Kun07, FPR04, Hal14], but also methods based on variational principles have been examined and shown to work very well in practice.

However, as soon as the number of measurements is decreased, filtered backprojection produces severe artifacts. This is due to the fact that it is not able to take additional structure of the data, as for instance sparsity, into account. In order to be able to produce better results in such undersampling scenarios, we will use iterative schemes to solve an optimization problem suited for sparse recovery. One example of such minimization targets, that has been used frequently in imaging is total variation (TV) minimization. In TV minimization, the regularization functional is $\|\nabla \cdot\|_1$, which is enforcing sparsity of the gradient of a signal in a similar fashion as the regularizer used for basis pursuit, $\|\cdot\|_1$, enforces sparsity of the signal. We will see in Chapter 4 that, at least for one dimensional signals, however, the number of measurements cannot be reduced as much as in the standard Compressed Sensing scenario mentioned above.

Such gradient sparse structure is ubiquitous for example in images, but in many classes of data, we do not know a 'good' representation for our data a priori. This is where dictionary learning techniques come into play.

Gradient sparsity is a special case of cosparsity in a so-called analysis operator Ω, which means that for a given signal y, one has that Ωy is sparse. Such analysis operators can be used for the solution of underdetermined systems in a similar fashion as in TV minimization by using the regularization functional $\|\Omega \cdot\|_1$. Finding a suitable operator can be done for example with a minimization approach. Given a sample of Ω-cosparse signals $y_1, y_2, \dots, y_N \in \mathbb{R}^d$ (possibly corrupted with some noise), we can arrange them in a big matrix $Y = (y_1, y_2, \dots, y_N) \in \mathbb{R}^{d\times N}$. Then the operator Ω is a global minimizer of

$$\hat{\Omega} = \underset{\Gamma \in \mathcal{A}, X \in \mathcal{X}_\ell}{\arg\min} \|\Gamma Y - X\|_F^2, \tag{1.2}$$

where $\mathcal{A}$ consists of all matrices in $\mathbb{R}^{K\times d}$ with normalized rows and $\mathcal{X}_\ell$ is the set of all matrices in $\mathbb{R}^{K\times N}$, with ℓ zeros in each column. However, this optimization problem is highly nonconvex and a lot of local minima are typically possible. The target function in (1.2) has been used for example for Analysis K-SVD [RPE13], Analysis SimCo [DWD$^+$16] and more recently in [SS17]. Modern developments in machine learning suggest to tackle such optimization problems using stochastic gradient methods, however, without any additional postprocessing we cannot expect to end up at a global minimum. This approach to learn analysis operators is somewhat 'dual' to learning dictionaries for sparse data, which can also be seen from the used optimization principle, which is very similar to the one used for K-SVD [AEB06] or ITKM [Sch15a]. We will explore this topic in more depth in Chapter 5.

1.1 Outline

We will start with two chapters on the use of Compressed Sensing in Photoacoustic Tomography. There the measurement setup puts clear limitations on the kinds of random measurements we are able to consider and also the sparse structure is not apparent in the data a priori. We will nevertheless present positive results concerning subsampled reconstruction and also show transformations that sparsify the data.

In Chapter 2, we will use a sparsity prior to use ℓ_1-norm regularization for photoacoustic tomography with integrating line detectors. In order to decrease the necessary number of measurements, an approach based on lossless expander matrices is used to give theoretical results. In the experimental section, we observe that by using total variation minimization, we can improve the quality of recovery, while maintaining the low number of measurements. We will revisit this in Chapter 3. One key feature to obtain the recovery results is the concept of a *sparsifying temporal transform*, which allowed us to sparsify the data without compromising the restricted isometry property of the compressed sensing measurements.

Chapter 3 improves over the results presented in Chapter 2 by considering a different optimization target and by using the second derivative as a temporal transform. Via the wave equation that is used to describe the propagation of the acoustic wave, the second time derivative of the pressure data corresponds (up to a constant) to the Laplacian of the initial data. Under the assumption that the Laplacian of the initial data is sparse, this can be used to get recovery guarantees when sparse recovery techniques are employed.

Chapter 4 is a bookchapter treating total variation minimization for Compressed Sensing. TV Minimization uses the sparsity of the gradient of a signal, so instead of minimizing $\|x\|_1$, as in basis pursuit, one minimizes $\|\nabla x\|_1$. This approach has been pioneered by Rudin, Osher and Fatemi for image denoising, but has recently also been used in compressed sensing. In the chapter, we give a review over existing results, introduce a geometric perspective and provide some new results concerning reconstruction from subgaussian measurements.

While in Chapter 4, we were interested in recovery guarantees for reconstruction when we know that a signal x is sparse when we apply the gradient operator, in Chapter 5 we assume that we are given some signals, from which we know they are sparse when we apply *some unknown* operator Ω. Our aim is to find said operator as fast and accurately as possible. We derive four algorithms each suited for different application scenarios.

Finally, in Chapter 6 we conclude with an outlook on possible future research directions and a discussion of open problems.

2 A Novel Compressed Sensing Scheme for Photoacoustic Tomography

This chapter is based on joint work with

F. Krahmer[2]
T. Berer[3]
P. Burgholzer[3]
M. Haltmeier[1].

It has been published in [SKB+15].

[1] Department of Mathematics, University of Innsbruck, Technikestraße 13, A-6020 Innsbruck, Austria. E-mail: markus.haltmeier@uibk.ac.at

[2] Faculty of Mathematics, Office No.: 02.10.039, Boltzmannstraße 3, 85748 Garching (Munich),Germany. E-mail: felix.krahmer@tum.de

[3] Christian Doppler Laboratory for Photoacoustic Imaging and Laser Ultrasonics, and Research Center for Non-Destructive Testing (RECENDT), Altenberger Straße 69, 4040 Linz, Austria. E-mail: {thomas.berer,peter.burholzer}@recendt.at

Interlude

In this chapter, we take a first look at the possibility of speeding up data acquisition in photoacoustic tomography using compressed sensing techniques. Decreasing measurement time serves two primary purposes. On the one hand, it reduces motion artifacts due to undesired movements, and on the other hand it decreases examination time for the patient. In this chapter, we propose a new scheme for speeding up the data collection process in photoacoustic tomography. As measurement data we use random combinations of pressure values that we use to recover a complete set of pressure data prior to the actual image reconstruction. Sparsity of the data is obtained via a temporal transformation, which commutes with the random measurement matrix and does not perturb its restricted isometry property. We obtain theoretical recovery guarantees based on existing theory on expander graphs and support the theory by reconstruction results on simulated data as well as on experimental data.

2.1 Introduction

Photoacoustic tomography (PAT) is a recently developed non-invasive medical imaging technology whose benefits combine the high contrast of pure optical imaging with the high spatial resolution of pure ultrasound imaging [Bea11, Wan09, XW06a]. In order to speed up the measurement process, in this paper we propose a novel compressed sensing approach for PAT that uses random combinations of the induced pressure as measurement data. The proposed strategy yields recovery guarantees and furthermore comes with an efficient numerical implementation allowing high resolution real time imaging. We thereby focus on a variant of PAT using integrating line detectors proposed in [BHP$^+$05, PNHB07b, BVG$^+$12]. Our strategy, however, can easily be adapted to more classical PAT setups using arrays of point-like detectors.

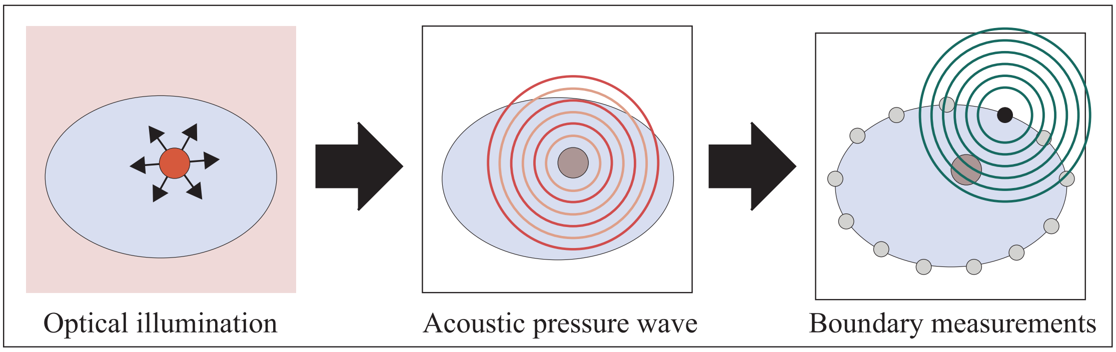

Figure 2.1: BASIC PRINCIPLE OF PAT. A semi-transparent sample is illuminated with a short optical pulse that induces an acoustic pressure wave. The induced pressure is measured outside of the sample and used to recover an image of the interior.

Our proposal is based on the main components of compressed sensing, namely randomness and sparsity. Compressed sensing is one of the most influential discoveries in applied mathematics and signal processing of the past decade [CRT06a, Don06a]. By combining the benefits of data compression and data acquisition it allows to recover a signal from far fewer linear measurements than suggested by Shannon's sampling theorem. It has led to several new proposed sampling strategies in medical imaging, for example for speeding up MRI data acquisition (see [LDP07, LDSP08]) or completing under-sampled CT images [CTL08]. Another prominent application of compressed sensing is the single pixel camera (see [DDT$^+$08]) that circumvents the use of several expensive high resolution sensors in digital photography.

2.1.1 Photoacoustic tomography (PAT)

PAT is based on the generation of acoustic waves by illuminating a semi-transparent sample with short optical pulses (see Figure 2.1). When the sample is illuminated with a short laser pulse, parts of the optical energy become absorbed. Due to thermal expansion a subsequent pressure wave is generated depending on the structure of the sample. The induced pressure waves are recorded outside of the sample and used to recover an image of the interior, see [BBMG$^+$07, KK08, XW06a, Wan09].

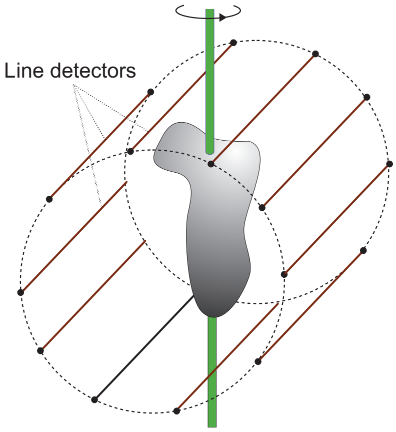

Figure 2.2: PAT WITH INTEGRATING LINE DETECTORS. An array of integrating line detectors measures integrals of the induced acoustic pressure over a certain number of parallel lines. These data are used to recover a linear projection of the object in the first step. By rotating the array of line detectors around a single axis several projection images are obtained and used to recover the actual three dimensional object in a second step.

In this paper, we consider a special variant of photoacoustic tomography that uses integrating line detectors for recording the pressure waves, as proposed in [BHP$^+$05]. As illustrated in Figure 2.2, an array of line detectors is arranged around the investigated sample and measures integrals of the pressure wave over a certain number of parallel lines. Assuming constant speed of sound, the pressure integrated along the direction of the line detectors satisfies the two dimensional wave equation

$$\begin{cases} \partial_t^2 p(x,t) - \Delta p(x,t) = 0\,, & \text{for } (x,t) \in \mathbb{R}^2 \times (0,\infty) \\ p(x,0) = f(x)\,, & \text{for } x \in \mathbb{R}^2 \\ \partial_t p(x,0) = 0\,, & \text{for } x \in \mathbb{R}^2\,, \end{cases} \tag{2.1}$$

where the time scaling is chosen in such a way that the speed of sound is normalized to one. The initial datum f in (2.1) is the two dimensional projection image of the actual, three dimensional initial pressure distribution.

Image reconstruction in PAT with integrating line detectors can be performed via a two-stage approach [BBMG$^+$07, PNHB07a]. In the first step the measured pressure values corresponding to values of the solution of (2.1) outside the support of f are used to reconstruct a linear projection (the initial data in (2.1)) of the three dimensional initial pressure distribution. This procedure is repeated by rotating the array of line detectors around a single axis which yields projection images for several angles. In a second step these projection images are used to recover the three dimensional initial pressure distribution by inverting the classical Radon transform. In this work we focus on the first problem of reconstructing the (two dimensional) initial pressure distribution f in (2.1).

Suppose that the integrating line detectors are arranged on the surface of a circular cylinder of radius $\rho > 0$, and that the object is located inside that cylinder (see Figure 2.2). The data measured by the array of line detectors are then modeled by

$$p_j := p(z_j, \cdot)\colon [0, 2\rho] \to \mathbb{R}\,, \tag{2.2}$$

$$z_j := \begin{pmatrix} \rho\cos(2\pi(j-1)/N) \\ \rho\sin(2\pi(j-1)/N) \end{pmatrix}\,, \qquad \text{for } j = 1, \dots, N\,, \tag{2.3}$$

where p_j is the pressure signal corresponding to the j-th line detector. Since the two dimensional initial pressure distribution f is supported in a disc of radius ρ and the speed of sound is constant and normalized to one, no additional information is contained in data $p(z_j, t)$ for $t > 2\rho$. This can be seen, for example by exploiting the explicit relations between the two dimensional pressure signals $p(z_j, \cdot)$ and the spherical means (compare with Subsection 2.3.3) of the initial pressure distribution; see [PNHB09].

Of course, in practice also the functions $p_j\colon [0, 2\rho] \to \mathbb{R}$ have to be represented by discrete samples. However, temporal samples can easily be collected at a high sampling rate compared to the spatial sampling, where each sample requires a separate sensor. It is therefore natural to consider the semi-discrete data model (2.2). Our compressed PAT scheme could easily be adapted to a fully discretized data model.

2.1.2 Compressed sensing PAT

When using data of the form (2.2), high resolution imaging requires the number N of detector locations to be sufficiently large. As the fabrication of an array of parallel line detectors is demanding, most experiments using integrating line detectors have been carried out using a single line detector, scanned on circular paths using scanning stages [NHK$^+$10, GBB$^+$10]. Recently, systems using arrays of 64 parallel line detectors have been demonstrated [GNW$^+$14, BMFB$^+$15]. The most costly building blocks in such devices are the analog to digital converters (ADC). For completely parallel readout, a separate ADC is required for every detector channel.

In order to reduce costs, in these practical implementations two to four line detector channels are multiplexed to one ADC. For collecting the complete pressure data, the measurements have to be performed two (respectively four) times, because only 32 (respectively 16) of the 64 line measurements can be read out in parallel. This, again, leads to an increased overall measurement time. For example, using an excitation laser with a repetition rate of $10\,\mathrm{Hz}$ and two times multiplexing, a measurement time of $0.2\,\mathrm{s}$ for a projection image has been reported in [GNW$^+$14]. Without multiplexing, this measurement time would reduce to the half.

In order to speed up the scanning process and to reduce system costs, in this paper we propose a novel compressed sensing approach that allows to perform a smaller number of random measurements with a reduced number of ADCs, while retaining high spatial resolution. For that purpose, instead of collecting individually sampled data $p_j(t)$ as in (2.2), we use random combinations

$$y_i(t) = \sum_{j \in J_i} p_j(t) \quad \text{for } i \in \{1, \dots, m\} \text{ and } t \in [0, 2\rho]\,, \tag{2.4}$$

where $m \ll N$ is the number of compressed sensing measurements and $J_i \subset \{1, \dots, N\}$ corresponds to the random set of detector locations contributing to the i-th measurement. In the reconstruction process the random linear combinations $y_i(t)$ are used to recover the full set of pressure data $p_1(t), \dots, p_N(t)$ using compressed sensing techniques. The initial pressure distribution f in (2.1) is subsequently recovered from the completed pressure data by applying standard PAT reconstruction algorithms such as time reversal [BMHP07, HKN08, TC10] or filtered backprojection [BBMG$^+$07, FHR07, FPR04, Hal14, Kun07, XW05].

A naive approach for recovering the pressure data from the random measurements $y_i(t)$ would be to solve (2.4) for $p_j(t)$ separately for each $t \in [0, 2\rho]$. Since $m \ll N$, this is a severely underdetermined system of linear equations and its solution requires appropriate prior knowledge of the unknown parameter vector. Compressed sensing suggests to use the sparsity of the parameter vector in a suitable basis for that purpose. However, recovery guarantees for zero/one measurements of the type (2.4) are basis-dependent and require the parameter to be sparse in the standard basis rather than sparsity in a different basis such as orthonormal wavelets (see Subsection 2.2.2). However, for pressure signals (2.2) of practical relevance such sparsity assumption in the original basis does not hold.

In this work we therefore propose a different approach for solving (2.4) by exploiting special properties of the data in PAT. For that purpose we apply a transformation that acts in the temporal variable only, and makes the transformed pressure values sufficiently sparse in the spatial (angular) component. In Subsection 2.3.3 we present an example of such a transform. The application of a sparsifying transform to (2.4) yields linear equations with unknowns being sparse in the angular variable. It therefore allows to apply sparse recovery results for the zero/one measurements under consideration.

2.1.3 Relations to previous work

A different compressed sensing approach for PAT has been considered in [PL09, GLSW10]. In these articles, standard point samples (such as (2.2)) have been used as measurement data and no recovery guarantees have been derived. Further, in [PL09, GLSW10] the phantom is directly reconstructed from the incomplete data, whereas we first complete the data using sparse recovery techniques. Our approach is more related to a compressed sensing approach for PAT using a planar detector array that has been proposed in [HZB$^+$14] and also uses random zero/one combinations of pressure values and recovers the complete pressure prior to the actual image reconstruction. However, in [HZB$^+$14] the sparsifying transform is applied in spatial domain where recovery guarantees are not available as noted above. We finally notice that our proposal of using a sparsifying temporal transform can easily be extended to planar detector arrays in two or three spatial dimensions; compare Section 2.5.

2.1.4 Outline

The rest of this paper is organized as follows. In Section 2.2 we review basic results from compressed sensing that we require for our proposal. We therefore focus on recovery guarantees for zero/one matrices modeled by lossless expanders; see Subsection 2.2.2. In Section 2.3 we present the mathematical framework of the proposed PAT compressed sensing scheme. The sparsity in the spatial variable, required for ℓ^1-minimization, is obtained by applying a transformation acting in the temporal variable. An example of such a transformation is given in Subsection 2.3.3. In Section 2.4 we present numerical results supporting our theoretical investigations. The paper concludes with a short discussion in Section 2.5.

2.2 Background from compressed sensing

In this section we shortly review basic concepts and results of compressed sensing (sometimes also termed compressive sampling). Our main focus will be on recovery results for lossless expanders, which are the basis of our PAT compressed sensing scheme.

2.2.1 Compressed sensing

Suppose one wants to sense a high dimensional data vector $\mathbf{x} = (x_1, \ldots, x_N) \in \mathbb{R}^N$, such as a digital image. The classical sampling approach is to measure each component of x_i individually. Hence, in order to collect the whole data vector one has to perform N separate measurements, which may be too costly. On the other hand it is well known and the basis of data compression algorithms, that many datasets are compressible in a suitable basis. That is, a limited amount of information is sufficient to capture the high dimensional vector $\mathbf{x}$.

Compressed sensing incorporates this compressibility observation into the sensing mechanism [CRT06a, CT06, Don06a]. Instead of measuring each coefficient of the data vector individually, one collects linear measurements

$$\mathbf{A}\mathbf{x} = \mathbf{y}\,, \tag{2.5}$$

where $\mathbf{A} \in \mathbb{R}^{m \times N}$ is the measurement matrix with $m \ll N$, and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$ is the measurement vector. Any component of the data vector can be interpreted as a scalar linear measurement performed on the unknown $\mathbf{x}$, and the assumption $m \ll N$ means that far fewer measurements than parameters are available. As $m \ll N$, the system (2.5) is highly underdetermined and cannot be uniquely solved (at least without additional information) by standard linear algebra.

Compressed sensing overcomes this obstacle by utilizing randomness and sparsity. Recall that the vector $\mathbf{x} = (x_1, \ldots, x_N)$ is called s-sparse if the support

$$\operatorname{supp}(\mathbf{x}) := \{j \in \{1, \ldots, N\} \colon x_j \neq 0\} \tag{2.6}$$

contains at most s elements. Results from compressed sensing state that for suitable $\mathbf{A}$, any s-sparse $\mathbf{x} \in \mathbb{R}^n$ can be found via the optimization problem

$$\begin{aligned} &\underset{\mathbf{z} \in \mathbb{R}^N}{\text{minimize}} \quad \|\mathbf{z}\|_1 = \sum_{j=1}^{N} |z_j| \\ &\text{such that} \quad \mathbf{A}\mathbf{z} = \mathbf{y}\,. \end{aligned} \tag{2.7}$$

By relaxing the equality constraint $\mathbf{A}\mathbf{z} = \mathbf{y}$, the optimization problem (2.7) can be adapted to data which are only approximately sparse and noisy [CRT06b].

A sufficient condition to guarantee recovery is the so called *restricted isometry property* (RIP), requiring that for any s-sparse vector x, we have

$$(1 - \delta)\|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2 \leq (1 + \delta)\|\mathbf{x}\|_2 \quad \text{for some small } \delta \in (0, 1)\,.$$

The smallest constant δ satisfying this inequality is called the s-restricted isometry constant of A and denoted by δ_s. Under certain conditions on δ_s, recovery guarantees for sparse and approximately sparse data can be obtained, see for example [Can08, CZ14].

While the restricted isometry itself is deterministic, to date all constructions that yield near-optimal embedding dimensions m are based on random matrices. Sub-gaussian random matrices satisfy the RIP with high probability for an order-optimal embedding dimension $m = \mathcal{O}(s \log(N/s))$, see e.g. [BDDW08]. Partial random Fourier matrices (motivated by MRI measurements) and subsampled random convolutions (motivated by remote sensing) have been shown to allow for order-optimal embedding dimensions up to logarithmic factors, see [RV08, Rau07] and [RRT12a, KMR14a], respectively.

The sparsity is often not present in the standard basis of $\mathbb{R}^N$, but in a special *sparsifying* basis, such as wavelets. For matrices with subgaussian rows this does not cause a problem, as the rotation invariance of subgaussian vectors ensures that after incorporating an orthogonal transform, the resulting random matrix construction still yields RIP matrices with high probability. As a consequence, the sparsifying basis need not be known for designing the measurement matrix $\mathbf{A}$, which is often referred to as universality of such measurements [BDDW08].

Many structured random measurement systems including the partial random Fourier and the subsampled random convolution scenarios mentioned above, however, do not exhibit universality. For example, one can easily see that subsampled Fourier measurements cannot suffice if the signal is sparse in the Fourier basis. While it has been shown that this problem can be overcome by randomizing the column signs [KW11], such an alteration often cannot be implemented in the sensing setup. Another way to address this issue is by requiring incoherence between the measurement basis and the sparsity basis [CR07]. That is, one needs that inner products between vectors of the two bases are uniformly small. If not all, but most of these inner products are small, one can still recover, provided that one adjusts the sampling distribution accordingly; this scenario includes the case of Fourier measurements and Haar wavelet sparse signals [KW14a].

Incoherence is also the key to recovery guarantees for gradient sparse signals. Namely, many natural images are observed to have an approximately sparse discrete gradient. As a consequence, it has been argued using a commutation argument that one can recover the signal from uniformly subsampled Fourier measurements via minimizing the ℓ^1 norm of the discrete gradient, the so-called total variation (TV) [CRT06a]. TV minimization had already proven to be a favorable method in image processing, see, for example [ROF92, CL97]. A problem with this approach is that the compressed sensing recovery guarantees then imply good recovery of the gradient, not of the signal itself. Small errors in the gradient, however, can correspond to substantial errors in the signal, which is why this approach can only work if no noise is present. A refined analysis that allows for noisy measurements requires the incoherence of the measurement basis to the Haar wavelet basis [NW13b]. Again, TV minimization is considered for recovery. By adjusting the sampling distribution, these results have also been shown to extend to Fourier measurements and other systems with only most measurement vectors incoherent to the Haar basis [KW14a].

For the measurement matrices considered in this work, namely zero/one matrices based on expander graphs, recovery guarantees build on an ℓ^1-version of the restricted isometry property, namely one requires that

$$(1-\delta)\|\mathbf{x}\|_1 \leq \|\mathbf{A}\mathbf{x}\|_1 \leq (1+\delta)\|\mathbf{x}\|_1$$

for all sufficiently sparse $\mathbf{x}$ and some constant $\delta > 0$; see [BGI+08] and Subsection 2.2.2 below. As the ℓ^1-norm is not rotation invariant, basis transformations typically destroy this property. That is, not even incoherence based recovery guarantees are available; recovery results only hold in the standard basis. Thus an important aspect of our work will be to ensure (approximate) sparsity in the standard basis. This will be achieved by applying a particular transformation in the time variable.

2.2.2 Recovery results for lossless expanders

Recall that we seek recovery guarantees for a measurement setup, where each detector is switched on exactly d out of m times. That is, one obtains a binary measurement matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ with exactly d ones in each column. It therefore can be interpreted as the adjacency matrix of a left d-regular bipartite graph. Under certain additional conditions, such a bipartite graph is a lossless expander (see Definition 2.2.1) which, as we will see, guarantees stable recovery of sparse vectors. Expander graphs have been used since the 1970s in theoretical computer science, originating in switching theory from modeling networks connecting many users, cf. [Kla84] for further applications. They have also been useful in measure theory, where it was possible to solve the Banach-Ruziewicz problem using tools from the construction of expander graphs, see [Lub94] for a detailed examination of this connection. For a survey on expander graphs and their applications, see for example [HLW06, Lub12].

Compressed sensing with expander graphs has been considered in [BGI+08, BIR08, IR08, JXHC09, XH07], where also several efficient algorithms for the solution of compressed sensing problems using expander graphs have been proposed. A short review of sparse recovery algorithms using expander like matrices is given in [GI10]. In this subsection we recall main compressed sensing results using lossless expanders as presented in the recent monograph [FR13], where the proofs of all mentioned theorems can be found in Section 13.

Recall that a bipartite graph consists of a triple (L, R, E), where L is the set of left vertices, R the set of right vertices, and $E \subset L \times R$ is the set of edges. Any element $(j, i) \in E$ represents an edge with left vertex $j \in L$ and right vertex $i \in R$.

A bipartite graph (L, R, E) is called d-regular for some $d \geq 0$, if for every given left vertex $j \in L$, the number of edges $(j, i) \in E$ emerging from j is exactly equal to d. Finally, for any

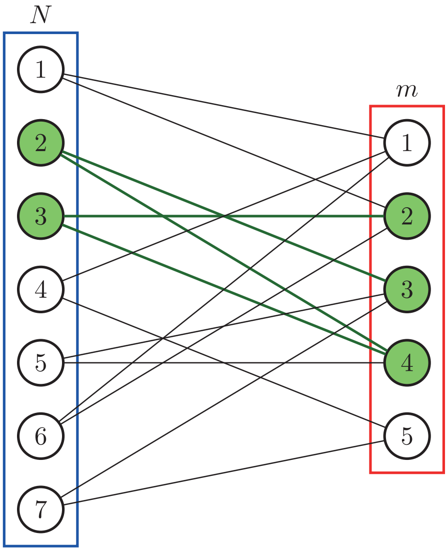

Figure 2.3: Example of a left d-regular bipartite graph with $d = 2$, $N = 7$ left vertices, and $m = 5$ right vertices. Here $d = 2$ because exactly 2 edges emerge at each left vertex. For $J = \{2, 3\}$ the set of right vertices connected to J is given by $R(J) = \{2, 3, 4\}$.

subset $J \subset L$ of left vertices, let

$$R(J) := \{i \in R\colon \text{ there exists some } j \in J \text{ with } (j,i) \in E\}$$

denote the set of right vertices connected to J. Obviously, for any left d-regular bipartite graph and any $J \subset L$, we have $|R(J)| \leq d|J|$.

Definition 2.2.1 (Lossless expander)**.** *Let (L, R, E) be a left d-regular bipartite graph, $s \in \mathbb{N}$, and $\theta \in [0,1]$. Then (L, R, E) is called an (s, d, θ)-lossless expander, if*

$$|R(J)| \geq (1-\theta)d|J| \quad \textit{for all } J \subset L \textit{ with } |J| \leq s\,. \tag{2.8}$$

For any left d-regular bipartite graph, the smallest number $\theta \geq 0$ satisfying (2.8) is called the s-th restricted expansion constant and denoted by θ_s.

The following theorem states that a randomly chosen d-regular bipartite graph will be a lossless expander with high probability.

Theorem 2.2.1 (Regular bipartite graphs are expanders with high probability)**.** *For every $0 < \varepsilon < \frac{1}{2}$, every $\theta \in (0,1)$ and every $s \in \mathbb{N}$, the proportion of (s, d, θ)-lossless expanders among the set of all left d-regular bipartite graphs having N left vertices and m right vertices exceeds*

$1-\varepsilon$*, provided that*

$$d = \left\lceil \frac{1}{\theta} \ln\left(\frac{\mathrm{e}N}{\varepsilon s}\right) \right\rceil \quad \textit{and} \quad m \geq c_\theta s \ln\left(\frac{\mathrm{e}N}{\varepsilon s}\right) . \tag{2.9}$$

Here c_θ *is a constant only depending on* θ*,* e *is Euler's constant,* $\ln(\cdot)$ *denotes the natural logarithm and* $\lceil x \rceil$ *denotes the smallest integer larger or equal to* x*.*

According to Theorem 2.2.1, any randomly chosen left d-regular bipartite graph is a lossless expander with high probability, provided that (2.9) is satisfied. The following theorem states that the adjacency matrix $\mathbf{A} \in \{0,1\}^{m \times N}$,

$$\mathbf{A}_{ij} = 1 :\Longleftrightarrow (j,i) \in E \,, \tag{2.10}$$

of any lossless expander with left vertices $L = \{1, \ldots, N\}$ and right vertices $R = \{1, \ldots, m\}$ yields stable recovery of any sufficiently sparse parameter vector. The result was first established in [BGI$^+$08], we present the version found in [FR13].

Theorem 2.2.2 (Recovery guarantee for lossless expanders)**.** *Let* $\mathbf{A} \in \{0,1\}^{m \times N}$ *be the adjacency matrix of a left* d*-regular bipartite graph having* $\theta_{2s} < 1/6$*. Further, let* $\mathbf{x} \in \mathbb{C}^N$*,* $\eta > 0$ *and* $\mathbf{e} \in \mathbb{C}^m$ *satisfy* $\|\mathbf{e}\|_1 \leq \eta$*, set* $\mathbf{b} := \mathbf{A}\mathbf{x} + \mathbf{e}$*, and denote by* $\mathbf{x}_\star$ *a solution of*

$$\begin{aligned} &\underset{\mathbf{z} \in \mathbb{C}^N}{\text{minimize}} && \|\mathbf{z}\|_1 \\ &\text{such that} && \|\mathbf{A}\mathbf{z} - \mathbf{b}\|_1 \leq \eta \,. \end{aligned} \tag{2.11}$$

Then

$$\|\mathbf{x} - \mathbf{x}_\star\|_1 \leq \frac{2(1-2\theta_{2s})}{(1-6\theta_{2s})} \sigma_s(\mathbf{x})_1 + \frac{4}{(1-6\theta_{2s})d} \eta \,.$$

Here the quantity $\sigma_s(\mathbf{x})_1 := \inf\{\|\mathbf{x} - \mathbf{z}\|_1 \colon \mathbf{z}$ *is* s*-sparse*$\}$ *measures by how much the vector* $\mathbf{x} \in \mathbb{C}^N$ *fails to be* s*-sparse.*

Combining Theorems 2.2.1 and 2.2.2, we can conclude that the adjacency matrix $\mathbf{A}$ of a randomly chosen left d-regular bipartite graph will, with high probability, recover any sufficiently sparse vector $\mathbf{x} \in \mathbb{C}^N$ by basis pursuit reconstruction (2.11).

2.3 Mathematical framework of the proposed PAT compressed sensing scheme

In this section we describe our proposed compressed sensing strategy. As mentioned in the introduction, we focus on PAT with integrating line detectors, which is governed by the two dimensional wave equation (2.1). In the following we first describe the compressed sensing measurement setup in Subsection 2.3.1 and describe the sparse recovery strategy in Subsection 2.3.2. As the used pressure data are not sparse in the original domain we introduce a temporal transform that makes the data sparse in the spatial domain. In Subsection 2.3.3 we present an example of such a sparsifying temporal transform. In Subsection 2.3.4 we finally summarize the whole PAT compressed sensing scheme.

2.3.1 Compressed sensing PAT

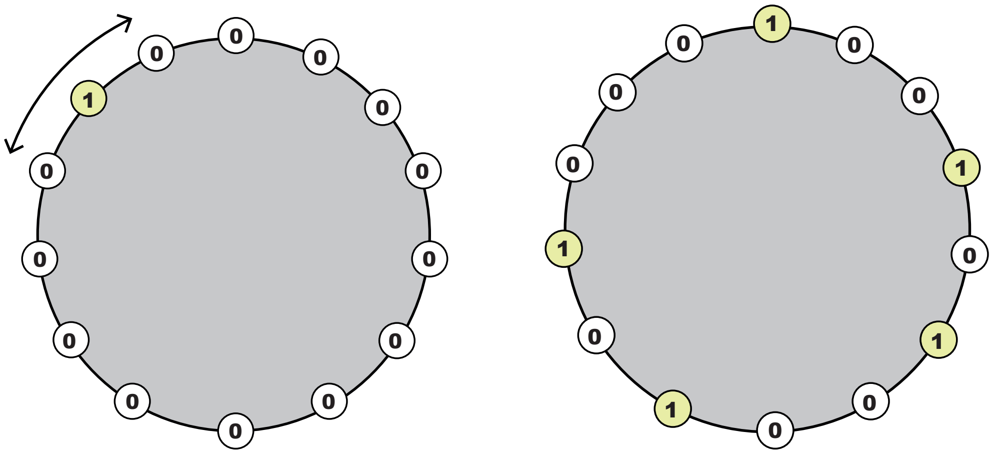

Figure 2.4: LEFT: Classical PAT sampling, where a single detector is moved around the object to collect individual pressure signal p_j. RIGHT: Compressed sensing approach where each measurement consists of a random zero/one combination of individual pressure values.

We define the unknown full sample pressure $p_j \colon [0, 2\rho] \to \mathbb{R}$, for $j \in \{1, \ldots, N\}$ by (2.2), (2.3), where p is the solution of the two dimensional wave equation (2.1). We suppose that the (two dimensional) initial pressure distribution f is smooth and compactly supported in $B_R(0)$ which implies that any p_j is smooth and vanishes in a neighborhood of zero. Furthermore we assume that the data (2.2) are sampled finely enough to allow for f to be reconstructed from $(p_1, \ldots, p_N)$ by standard PAT reconstruction algorithms such as time reversal [BMHP07, HKN08, TC10] or filtered backprojection [BBMG$^+$07, FHR07, FPR04, Hal14, Kun07, XW05].

Instead of measuring each pressure signal p_j separately, we take m compressed sensing measurements. For each measurement, we select sensor locations z_j with $j \in J_i$ at random and take the sum of the corresponding pressure values p_j. Thus the i-th measurement is given by

$$y_i(t) := \sum_{j \in J_i} p_j(t) \quad \text{for } i \in \{1, \ldots, m\}\,, \tag{2.12}$$

where $J_i \subset \{1, \dots, N\}$ corresponds to the set of all detector locations selected for the i-th measurement, and $m \ll N$ is the number of compressed sensing measurements.

In practice, the compressed sensing measurements could be realized by summation of several detector channels, using a configurable matrix switch and a summing amplifier. Even more simply, a summing amplifier summing over all N channels could be used, while individual detector channels z_j are turned on or off, using solid state relays. Thereby, only one ADC is required for one compressed sensing measurement. Performing m compressed sensing measurements in parallel is facilitated by using m ADCs in parallel, and the according number of matrix switches, relays, and summing amplifiers.

In the following we write

$$\mathbf{p}\colon (0,\infty) \to \mathbb{R}^N \colon t \mapsto \mathbf{p}(t) = (p_1(t), \dots, p_N(t))^\mathsf{T}\,, \tag{2.13}$$

$$\mathbf{y}\colon (0,\infty) \to \mathbb{R}^m \colon t \mapsto \mathbf{y}(t) = (y_1(t), \dots, y_m(t))^\mathsf{T}\,, \tag{2.14}$$

for the vector of unknown complete pressure signals and the vector of the corresponding measurement data, respectively. Further, we denote by

$$\mathbf{A} \in \{0,1\}^{m \times N} \text{ with entries } \mathbf{A}_{ij} := \begin{cases} 1, & \text{for } j \in J_i \\ 0, & \text{for } j \notin J_i\,, \end{cases}$$

the matrix whose entries in the i-th row correspond the sensor locations selected for the i-th measurement. In order to apply the exact recovery guarantees from Subsection 2.2.2, we require that each column of $\mathbf{A}$ contains exactly d ones, where $d \in \mathbb{N}$ is some fixed number. Practically, this means that each detector location contributes to exactly d of the m measurements, which also guarantees that the measurements are well calibrated and there is no bias towards some of the detector locations.

Recovering the complete pressure data (2.2) from the compressed sensing measurements (2.4) can be written as an uncoupled system of under-determined linear equations,

$$\mathbf{A}\mathbf{p}(t) = \mathbf{y}(t) \quad \text{for any } t \in [0, 2\rho]\,, \tag{2.15}$$

where $\mathbf{A}\mathbf{p}(t) := \mathbf{A}(\mathbf{p}(t))$ for any t. From (2.15), we would like to recover the complete set of pressure values $\mathbf{p}(t)$ for all times $t \in [0, 2\rho]$. Compressed sensing results predict that under certain assumptions on the matrix $\mathbf{A}$ any s-sparse vector $\mathbf{p}(t)$ can be recovered from $\mathbf{A}\mathbf{p}(t)$ by means of sparse recovery algorithms like ℓ^1-minimization.

Similar to many other applications (cf. Section 2.2 above), however, we cannot expect sparsity in the original domain. Instead, one has $\mathbf{p}(t) = \Psi\,\mathbf{x}(t)$, where Ψ is an appropriate orthogonal transform and $\mathbf{x}(t) \in \mathbb{R}^N$ is a sparse coefficient vector. This yields a sparse recovery problem for $\mathbf{x}(t)$ involving the matrix $\mathbf{A}\,\Psi$, which does not inherit the the recovery guarantees of Subsection 2.2.2. Hence we have to find a means to establish sparsity without considering different bases. Our approach will consist of applying a transformation in the time domain that sparsifies the pressure in the spatial domain. A further advantage of working in the original domain is that the structure of $\mathbf{A}$ allows the use of specific efficient algorithms like sparse matching pursuit [BIR08] or certain sublinear-time algorithms like [JXHC09, XH07].

2.3.2 Reconstruction strategy

We denote by $\mathcal{G}([0, 2\rho])$ the set of all infinitely differentiable functions $g\colon [0, 2\rho] \to \mathbb{R}$ that vanish in a neighborhood of zero. To obtain the required sparsity, we will work with a sparsifying transformation

$$\mathbf{T}\colon \mathcal{G}([0, 2\rho]) \to \mathcal{G}([0, 2\rho])\,, \tag{2.16}$$

that is, $\mathbf{Tp}(t) \in \mathbb{R}^N$ can be sufficiently well approximated by a sparse vector for any $t \in [0, 2\rho]$ and certain classes of practically relevant data $\mathbf{p}(t)$. Here we use the convention that $\mathbf{T}$ applied to a vector valued function $\mathbf{g} = (g_1, \dots, g_k)$ is understood to be applied in each component separately, that is,

$$\mathbf{Tg}(t) := ((\mathbf{T}g_1)(t), \dots, (\mathbf{T}g_k)(t))\,, \quad \text{for } t \in [0, 2\rho]\,. \tag{2.17}$$

We further require that $\mathbf{T}$ is an injective mapping, such that any $g \in \mathcal{G}([0, 2\rho])$ can be uniquely recovered from the transformed data $\mathbf{T}g$. See Subsection 2.3.3 for the design of such a sparsifying transformation.

Since any temporal transformation interchanges with $\mathbf{A}$, application of the sparsifying temporal transformation $\mathbf{T}$ to the original system (2.15) yields

$$\mathbf{A}(\mathbf{Tp}(t)) = \mathbf{Ty}(t) \quad \text{for any } t \in [0, 2\rho]\,. \tag{2.18}$$

So the clue is using a sparsifying transformation in the temporal direction, which preserves the structure of the matrix $\mathbf{A}$. As can be observed from Figure 2.5, due to the 'wavy' nature of the pressure data, sparsity in the temporal direction also yields sparsity in the angular direction. Therefore, application of a sparsifying transform $\mathbf{T}$ yields an approximately sparse unknown $\mathbf{Tp}(t)$ for equation (2.18), which therefore can be approached by standard methods for sparse recovery.

According to the choice of the temporal transform, the transformed pressure $\mathbf{Tp}(t)$ can be well approximated by a vector that is sparse in the spatial component. We therefore solve, for any $t \in [0, T]$, the following ℓ^1-minimization problem

$$\begin{aligned} &\underset{\mathbf{q}\in\mathbb{R}^N}{\text{minimize}} && \|\mathbf{q}\|_1 \\ &\text{such that} && \|\mathbf{A}\mathbf{q} - \mathbf{T}\mathbf{y}(t)\|_1 \leq \eta\,, \end{aligned} \tag{2.19}$$

for some error threshold $\eta > 0$. As follows from Theorem 2.2.2, the solution $\mathbf{q}_\star(t)$ of the ℓ^1-minimization problem (2.19) provides an approximation to $\mathbf{T}\mathbf{p}(t)$ and consequently we have $\mathbf{T}^{-1}\mathbf{q}_\star(t) \simeq \mathbf{p}(t)$ for all $t \in [0, 2\rho]$. Note that the use of ℓ^1-norm $\|\cdot\|_1$ for the constraint in (2.19) is required for the stable recovery guarantees for our particular choice of the matrix $\mathbf{A}$ containing zero/one entries using data that is noisy and only approximately sparse.

As a further benefit, compressed sensing measurements may show an increased signal to noise ratio. For that purpose consider Gaussian noise in the pressure data, where instead of the exact pressure data $\mathbf{p}(t)$, we have noisy data $\tilde{\mathbf{p}}(t) = \mathbf{p}(t) + \boldsymbol{\eta}$. For the sake of simplicity assume that the entries $\eta_j \sim \mathcal{N}(0, \sigma^2)$ of $\boldsymbol{\eta}$ are independent and identically distributed. The corresponding noisy (and rescaled) measurements are then given by

$$\frac{1}{|J_i|}\sum_{j\in J_i}(p_j(t) + \eta_j) = \frac{1}{|J_i|}\sum_{j\in J_i} p_j(t) + \frac{1}{|J_j|}\sum_{j\in J_i} \eta_j\,.$$

The variance in the compressed sensing measurements is therefore $\sigma^2/|J_i|$ compared to σ^2 in the individual data $\tilde{p}_j(t)$. Assuming some coherent averaging in the signal part this yields an increased signal to noise ratio reflecting the inherent averaging of compressed sensing measurements.

2.3.3 Example of a sparsifying temporal transform

As we have seen above, in order to obtain recovery guarantees for our proposed compressed scheme, we require a temporal transformation that sparsifies the pressure signals in the angular component. In this section, we construct an example of such a sparsifying transform.

Since the solution of the two dimensional wave equation can be reduced to the spherical means, we will construct such a sparsifying transform for the spherical means

$$\mathbf{M}\, f(z, r) := \frac{1}{2\pi}\int_{\mathbb{S}^1} f(z + r\omega)\mathrm{d}\sigma(\omega) \quad \text{for } (z, r) \in \partial B_R(0) \times (0, \infty)\,.$$

In fact, the solution of the two dimensional wave equation (2.1) can be expressed in terms of the spherical means via $p(z, t) = \partial_t \int_0^t r\,\mathbf{M}\, f(z, r)/\sqrt{t^2 - r^2}\mathrm{d}r$, see [Joh82]. By using standard tools for solving Abel type equations, the last expression can be explicitly inverted, resulting in $(\mathbf{M}\, f)(z,r) = 2/\pi \int_0^r p(z,t)/\sqrt{r^2 - t^2}\mathrm{d}t$ (see [BBMG$^+$07, GV91]). Hence any sparsifying transformation for the spherical means $\mathbf{M}\, f$ also yields a sparsifying transformation for the solution of the wave equation (2.1), and vice versa.

We found empirically, that $\partial_r r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ is sparse in the spatial direction for any function f that is the superposition of few indicator functions of regular domains. Here $\partial_r g$ denotes the derivative in the radial variable,

$$\mathbf{H}_r\, g(z,r) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{g(z,s)}{r-s} \mathrm{d}s\,, \qquad \text{for } (z,r) \in \partial B_R(0) \times (0,\infty)\,,$$

the Hilbert transform of the function $g \colon \partial B_R(0) \times \mathbb{R} \to \mathbb{R}$ in the second component, and r the multiplication operator that maps the function $(z,r) \mapsto g(z,r)$ to the function $(z,r) \mapsto rg(z,r)$. Further, the spherical means $\mathbf{M}\, f \colon \partial B_R(0) \times \mathbb{R} \to \mathbb{R}$ are extended to an odd function in the second variable. For a simple radially symmetric phantom the sparsity of $\partial_r r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ is illustrated in Figure 2.5. Thus we can choose $\mathbf{T} = \partial_r r\, \mathbf{H}_r\, \partial_r$ as a sparsifying temporal transform for the spherical means. For that purpose we used a signal model consisting of superpositions of indicator functions of regular domains. This can be seen as a substitute for the sparsity assumption required for other compressed sensing techniques. Although in practice such a modelling assumption is not always strictly justified, the initial pressure distribution can often be well approximated by such superpositions. Therefore, in future work we will address the robustness of our compressed sensing technique with respect to such a signal model.

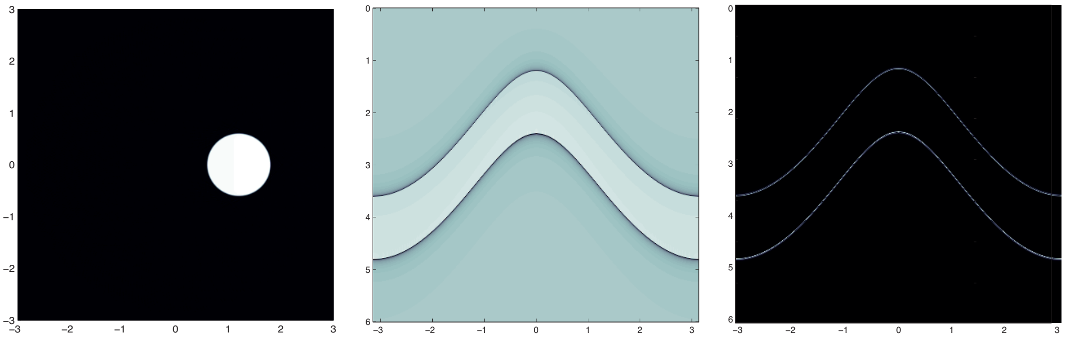

Figure 2.5: SPARSITY INDUCED BY $\partial_r(r\, \mathbf{H}_r\, \partial_r)$. The left image shows the simple disc-like phantom f (characteristic function of a disc), the middle image shows the filtered spherical means $r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ and the right image shows the sparse data $\partial_r(r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f)$.

The function $r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ also appears in the following formula for recovering a function from its spherical means derived in [FHR07].

Theorem 2.3.1 (Exact reconstruction formula for spherical means)**.** *Suppose* $f \in C^\infty(\mathbb{R}^2)$ *is* *supported in the closure of $B_R(0)$. Then, for any $x \in B_R(0)$, we have*

$$f(x) = \frac{1}{2\pi R} \int_{\partial B_R(0)} (r\,\mathbf{H}_r\,\partial_r\,\mathbf{M}\,f)(z, |x - z|)\,\mathrm{d}s(z)\,. \tag{2.20}$$

Proof. See [FHR07, Corollary 1.2]. □

In the practical implementation the spherical means are given only for a discrete number of centers $z_j \in \partial B_R(0)$ yielding semi-discrete data similar to (2.2), (2.3). Formula (2.20) can easily be adapted to discrete or semi-discrete data yielding a filtered backprojection type reconstruction algorithm; compare [FHR07, Section 4]. So if we can find the filtered spherical means

$$(r\,\mathbf{H}_r\,\partial_r\,\mathbf{M}\,f)(z_j, \cdot) \quad \text{for all detector locations } z_j \in \partial B_R(0) \tag{2.21}$$

we can obtain the desired reconstruction of f by applying the backprojection operator (the outer integration in (2.20)) to $r\,\mathbf{H}_r\,\partial_r\,\mathbf{M}\,f$.

2.3.4 Summary of the reconstruction procedure

In this section, we combine and summarize the compressed sensing scheme for photoacoustic tomography as described in the previous sections. Our proposed compressed sensing and sparse recovery strategy takes the following form.

(CS1) Create a matrix $\mathbf{A} \in \{0,1\}^{m \times N}$ as the adjacency matrix of a randomly selected left d-regular bipartite graph. That is, $\mathbf{A}$ is a random matrix consisting of zeros and ones only, with exactly d ones in each column.

(CS2) Perform m measurements, whereby in the i-th measurement pressure signals corresponding to the nonzero entries in i-th row of $\mathbf{A}$ are summed up, see Equations (2.4) and (2.15). This results in measurement data $\mathbf{A}\mathbf{p}(t) = \mathbf{y}(t) \in \mathbb{R}^m$ for any $t \in [0, 2\rho]$.

(CS3) Choose a transform $\mathbf{T}$ acting in the temporal direction, which sparsifies the pressure data $\mathbf{p}$ along the spatial direction; compare Equation (2.17).

(CS4) For any $t \in [0, 2\rho]$ and some given threshold η, perform ℓ^1-minimization (2.19) resulting in a sparse vector $\mathbf{q}_\star(t)$ satisfying $\|\mathbf{A}\mathbf{q}_\star(t) - \mathbf{T}\mathbf{y}(t)\|_1 \leq \eta$.

(CS5) Use $\mathbf{p}_\star(t) = \mathbf{T}^{-1}\mathbf{q}_\star(t)$ as the input for a standard PAT inversion algorithm for complete data, such as time reversal or filtered backprojection.

As we have seen in Subsection 2.2.2 the procedure (CS1)–(CS5) yields a close approximation to the original function f if the transformed data $\mathbf{T}\mathbf{p}(t)$ are sufficiently sparse in the spatial direction. The required sparsity level is hereby given by the expander-properties of the matrix $\mathbf{A}$. Note that for exact data and exactly sparse data, we can use the error threshold $\eta = 0$. In the more realistic scenario of noisy data and $\mathbf{Tp}(t)$ being only approximately sparse, we solve the optimization problem (2.19) to yield a near optimal solution with error level bounded by the noise level.

2.4 Numerical results

To support the theoretical examinations in the previous sections, in this section we present some simulations using the proposed compressed sensing method. We first present reconstruction results using simulated data and then show reconstruction results using experimental data.

2.4.1 Results for simulated data

As in Subsection 2.3.3, we work with the equivalent notion of the spherical means instead of directly working with the solution of the wave equation (2.1). In this case the compressed sensing measurements provide data

$$y_i(r) = \sum_{j \in J_i} m_j(r) \quad \text{for } i \in \{1, \dots, m\} \text{ and } t \in [0, 2\rho] \,, \tag{2.22}$$

where $m_j = (\mathbf{M} f)(z_j, \cdot)$ denote the spherical means collected at the j-the detector location z_j. We further denote by $\mathbf{m}(t) = (m_1(t), \dots, m_N(t))^\mathsf{T}$ the vectors of unknown complete spherical means and by $\mathbf{y}(t) = (y_1(t), \dots, y_m(t))^\mathsf{T}$ the vector of compressed sensing measurement data. Finally, we denote by $\mathbf{A} \in \{0, 1\}^{m \times N}$ the compressed sensing matrix such that (2.22) can be rewritten in the form $\mathbf{Am} = \mathbf{y}$.

As proposed in Subsection 2.3.3 we use $\mathbf{T} = \partial_r (r \, \mathbf{H}_r \, \partial_r)$ as a sparsifying transform for the spherical means. An approximation to $\partial_r r \, \mathbf{H}_r \, \partial_r \, \mathbf{M} \, f$ can be obtained from compressed sensing measurements in combination via ℓ^1-minimization. For the recovery of the original function from the completed measurements, we use one of the inversion formulas of [FHR07] presented given in Theorem 2.3.1. Recall that this inversion formulas can be implemented by applying the circular back-projection to the filtered spherical means $r \, \mathbf{H}_r \, \partial_r \, \mathbf{M} \, f$.

In order to obtain an approximation to the data (2.21) from the sparse intermediate reconstruction $\partial_r (r \, \mathbf{H}_r \, \partial_r \, \mathbf{M} \, f)(\cdot, r)$, one has to perform one numerical integration along the second dimension after the ℓ^1-minimization process. We found that this numerical integration introduces artifacts in the reconstruction of $r \, \mathbf{H}_r \, \partial_r \, \mathbf{M} \, f$, required for application of the inversion formula (2.20), see the middle image in Figure 2.6. These artifacts also yield some undesired blurring in the final reconstruction of f; see the middle image in Figure 2.7.

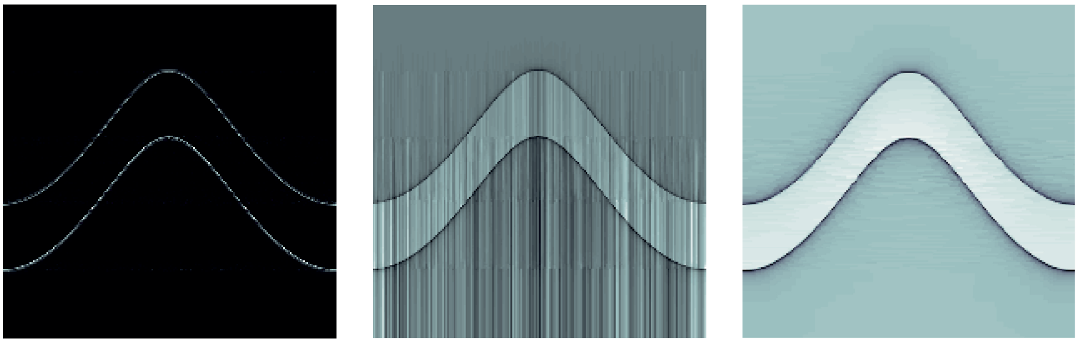

Figure 2.6: RECONSTRUCTION OF FILTERED SPHERICAL MEANS FOR $N = 200$ AND $m = 100$. Left: Reconstruction of $\partial_r r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ using ℓ^1-minimization. Center: Result of integrating the ℓ^1-reconstruction in the radial direction. Right: Result of directly recovering $r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ by TV-minimization.

In order to overcome the artifact introduced by the numerical integration, instead of applying an additional radial derivative to $r\, \mathbf{H}_r\, \partial_r$ to obtain sparsity in the spatial direction, we will apply one-dimensional total variation minimization (TV) for directly recovering $(r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f)(\cdot, r)$. Thereby we avoid performing numerical integration on the reconstructed sparse data. Furthermore, this yields much better results in terms of image quality of the final image f. Clearly, by performing TV-minimization, we solve a problem different to what is covered by the theory. For measurement matrices possessing the 2-RIP recovery guarantees using TV minimization have been established in [NW13b]. However, to the best of our knowledge, no such theory is currently available for 1-RIP measurement matrices, as it is required for our application. The better numerical performance of TV-minimization suggests that such an analysis is possible which will be addressed in future work.

In our interpretation, this performance discrepancy is comparable to the difference between uniform and variable density samples for subsampled Fourier measurements. While [CRT06a] proves recovery of the discrete gradient, this does not carry over to the signal in a stable way – a refined analysis was required [NW13b, KW14a]. Similarly, we expect that a refined analysis to be provided in subsequent work can help explain the quality gap between ℓ^1 and TV minimization that we observe in our scenario.

In order to approximately recover $r\, \mathbf{H}_r\, \partial_r\, \mathbf{M}\, f$ from the compressed sensing measurements, we perform, for any $r \in [0, 2\rho]$, one-dimensional discrete TV-minimization

$$\|\mathbf{A}\mathbf{q} - (r\, \mathbf{H}_r\, \partial_r \mathbf{y})(r)\|_2^2 + \lambda \|\mathbf{q}\|_{\mathrm{TV}} \to \min_{\mathbf{q} \in \mathbb{R}^N} . \tag{2.23}$$

Here $\|\mathbf{q}\|_{\mathrm{TV}} = 2\pi/N \sum_{j=1}^{N} |q_{j+1} - q_j|$ denotes the discrete total variation using the periodic extension $q_{N+1} := q_1$. The one-dimensional total variation minimization problem (2.23) can be efficiently solved using the fast iterative shrinkage thresholding algorithm (FISTA) of Beck and Teboulle [BT09]. The required proximal mapping for the total variation can be computed in $\mathcal{O}(N)$ operation counts by the tautstring algorithm (see [MvdG$^+$97, DK01, GO08]). The approximate solution of (2.23) therefore only requires $\mathcal{O}(NmN_{iter})$ floating point operations, with N_{iter} denoting the number of iterations in the FISTA. Assuming the radial variable to be discretized using $\mathcal{O}(N)$ samples, the whole data completion procedure by (2.23) only requires $\mathcal{O}(N^2 m N_{\text{iter}})$ operation counts. Since we found that fewer than 100 iterations in the FISTA are often sufficient for accurate results, the numerical effort of data completion is only a few times higher than that of standard reconstruction algorithms in PAT.

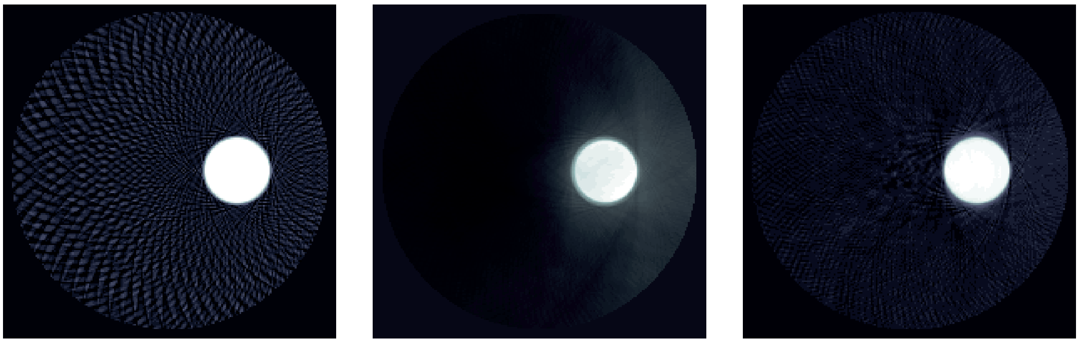

Figure 2.7: Reconstruction results for disc-like phantom using $N = 200$ and $m = 100$. Left: Reconstruction from 100 standard measurements. Center: Compressed sensing reconstruction using ℓ^1-minimization. Right: Compressed sensing reconstruction using TV-minimization. The reconstruction from standard measurements contains under-sampling artifacts which are not present in the compressed sensing reconstructions. Further, the use of TV-minimization yields much less blurred results than the use of ℓ^1-minimization.

Figures 2.6 and 2.7 show results of simulation studies for a simple phantom, where the initial pressure distribution f is the characteristic function of a disc. For the compressed sensing reconstruction, we used $m = 100$ random measurements instead of $N = 200$ standard point measurements. As one can see from the right image in Figure 2.7, the completed data $r\,\mathbf{H}_r\,\partial_r\,\mathbf{M}\,f$, for this simple phantom, is recovered almost perfectly from the compressed sensing measurements by means of TV-minimization. The reconstruction results in Figure 2.7 show that the combination of our compressed sensing approach with TV-minimization yields much better results than the use of ℓ^1-minimization. Fur comparison purpose, the left image in Figure 2.7 shows the reconstruction from 100 standard measurements. Observe that the use of $m = 100$ random measurements yields better result than the use of the 100 standard measurements, where artifacts due to spatial under-sampling are clearly visible.

2.4.2 Results for real measurement data

Experimental data have been acquired by scanning a single integrating line detector on a circular path around a phantom. A bristle with a diameter of $120\,\mu\mathrm{m}$ was formed to a knot and illuminated from two sides with pulses from a frequency doubled Nd:YAG laser with a wavelength of $532\,\mathrm{nm}$. The radiant exposure for each side was below $7.5\,{}^{\mathrm{mJ}}/_{\mathrm{cm}^2}$. Generated photoacoustic signals have been detected by a graded index polymer optical fiber being part of a Mach-Zehnder interferometer, as described in [GBB$^+$10]. Ultrasonic signals have been demodulated using a self-made balanced photodetector, the high-frequency output of which was sampled with a 12-bit data acquisition device. A detailed explanation of the used photo-detector and electronics can be found in [BMFH$^+$13]. The polymer fiber detector has been scanned around the phantom on a circular path with a radius of $6\,\mathrm{mm}$ and photoacoustic signals have been acquired on 121 positions. The scanning curve was not closed, but had an opening angle of $\pi/2\,\mathrm{rad}$. Hence photoacoustic signals have been acquired between $\pi/8\,\mathrm{rad}$ and $15\pi/8\,\mathrm{rad}$.

Using these experimental data, compressed sensing data have been generated, where each detector location was used $d = 10$ times and $m = 60$ measurements are made in total. The reconstruction of the complete measurement data has been obtained by one-dimensional discrete TV-minimization (2.23) as suggested in the Section 2.4.1. The measured and the recovered complete pressure data are shown in the top row in Figure 2.8. The bottom row in Figure 2.8 shows the reconstruction results from 121 standard measurements (bottom left) and the reconstruction from 60 compressed sensing measurements (bottom center). Observe that there is only a small difference between the reconstruction results. This clearly demonstrates the potential of our compressed sensing scheme (CS1)–(CS5) for decreasing the number of measurements while keeping image quality. For comparison purpose we also display the reconstruction using 60 standard measurement (bottom right). Compared to the compressed sensing reconstruction using the same number of measurements the use of standard measurements shows significantly more artifacts which are due to spatial under-sampling.

2.5 Discussion and outlook

In this paper, we proposed a novel approach to compressive sampling for photoacoustic tomography using integrating line detectors providing recovery guarantees for suitable datasets. Instead of measuring pressure data p_j at any of the N individual line detectors, our approach uses m random combinations of p_j with $m \ll N$ as measurement data. The reconstruction strategy

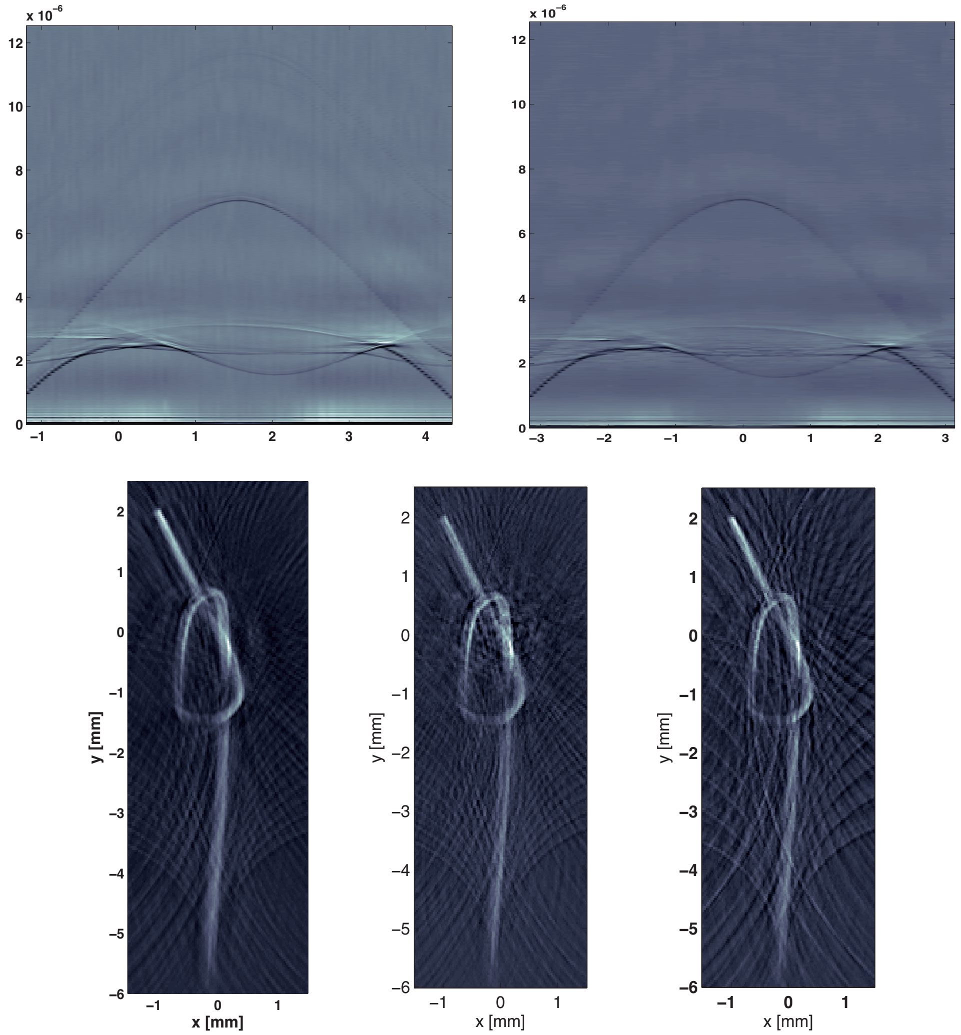

Figure 2.8: RECONSTRUCTION RESULTS FOR PAT MEASUREMENTS. Top left: Measured pressure data for $N = 121$ detector positions. Top right: Compressed sensing reconstruction of the full pressure data from $m = 60$ compressed measurements. Bottom left: Reconstruction from the full measurement data using $N = 121$ detector positions. Bottom center: Reconstruction from $m = 60$ compressed sensing measurements. Bottom right: Reconstruction from half of the measurement data using 60 detector positions.

consists of first recovering the pressure values p_j for $j \in \{1, \ldots, N\}$ and then applying a standard PAT reconstruction for obtaining the final photoacoustic image. For recovering the individual pressure data, we propose to apply a sparsifying transformation that acts in the temporal variable and makes the data sparse in the angular component. After applying such a transform the complete pressure data is recovered by solving a set of one dimensional ℓ^1-minimization problems.

This decomposition also makes our reconstruction algorithm numerically efficient.

Although we focused on PAT using integrating line detectors, we emphasize that a similar framework can be developed for standard PAT based on the three dimensional wave equation and a two dimensional array of point-like detectors. In such a situation, finding a sparsifying transform is even simpler. Recalling that N-shaped profile of the thermoacoustic pressure signal induced by the characteristic function of a ball suggests to use $\mathbf{p} \mapsto \partial_t^2 \mathbf{p}$ as a sparsifying transform.

Note that the recovery guarantees in this paper crucially depend on the choice of an appropriate sparsifying transform. In Subsection 2.3.3 we proposed a candidate for such a transformation that works well in our numerical examples. A more theoretical study of such sparsifying transforms and the resulting recovery guarantees for simple (piecewise constant) phantoms is postponed to further research. In this context, we will also investigate the use of different sparsifying temporal transforms, such as the 1D wavelet transform in the temporal direction. Further research includes using a fixed number of detectors in each measurement process. This requires novel results for right k-regular expander graphs and compressive sampling.

Acknowledgement

The authors thank Johannes Bauer-Marschallinger and Karoline Felbermayer for performing measurements and for providing the experimental data. The work of M. Sandbichler has been supported by a doctoral fellowship from the University of Innsbruck and the DK+ Project 'Computational Interdisciplinary Modeling' W1227. The work of F. Krahmer has been supported in parts by the German Science Foundation (DFG) in the context of the Emmy Noether Junior Research Group KR 4512/1-1 (RaSenQuaSI) and in parts by the German Federal Ministry of Education and Reseach (BMBF) through the cooperative research project ZeMat. T. Berer and P. Burgholzer have been supported by the Austrian Science Fund (FWF), project number S10503-N20, by the Christian Doppler Research Association, the Federal Ministry of Economy, Family and Youth, the European Regional Development Fund (EFRE) in the framework of the EU-program Regio 13, and the federal state of Upper Austria.

3 A New Sparsification and Reconstruction Strategy for Compressed Sensing Photoacoustic Tomography

This chapter is based on joint work with

M. HALTMEIER[1]
T. BERER[2]
J. BAUER-MARSCHALLINGER[2]
P. BURGHOLZER[2]
L. NGUYEN[3].

It has been submitted to the Journal of the Acoustical Society of America. A preprint can be found on arXiv [HSB$^+$17].

[1] Department of Mathematics, University of Innsbruck, Technikestraße 13, 6020 Innsbruck, Austria. E-mail: markus.haltmeier@uibk.ac.at

[2] Research Center for Non-Destructive Testing (RECENDT), Altenberger Straße 69, 4040 Linz, Austria. Also affiliated with Christian Doppler Laboratory for Photoacoustic Imaging and Laser Ultrasonics, Linz.

[3] Department of Mathematics, University of Idaho 875 Perimeter Dr, Moscow, ID 83844, US.

Interlude

In the previous chapter, we took a first look at the possibility of using compressed sensing to reduce the required number of measurements in photoacoustic imaging. However, where the previous chapter introduced a two-step scheme, in this chapter we perform the reconstruction in a single step. The two-step scheme had the drawback that first the measurement data was reconstructed and then, still, a filtered backprojection step had to be employed. Combining these two steps into one reduces the possibilty of an accumulation of errors of the individual methods and yields higher quality results. We again collect random combinations of pressure data at multiple detector locations, but this time we use the second derivative of the pressure data as a sparsifying transformation. Based on the insight that this second derivative corresponds to the Laplacian of the data in the spatial domain and that typical sources consist of flat regions and singularities along boundaries, which means that the Laplacian is typically sparse, we derive a single step scheme, which jointly recovers the initial data as well as its Laplacian. We present reconstruction results with simulated as well as experimental data.

3.1 Introduction

Photoacoustic tomography (PAT) is a non-invasive hybrid imaging technology, that beneficially combines the high contrast of pure optical imaging and the high spatial resolution of pure ultrasound imaging (see [Bea11, Wan09, XW06b]). The basic principle of PAT is as follows (see Fig. 3.1). A semitransparent sample (such as a part of a human body) is illuminated with short pulses of optical radiation. A fraction of the optical energy is absorbed inside the sample which causes thermal heating, expansion, and a subsequent acoustic pressure wave depending on the interior absorbing structure of the sample. The acoustic pressure is measured outside of the sample and used to reconstruct an image of the interior.

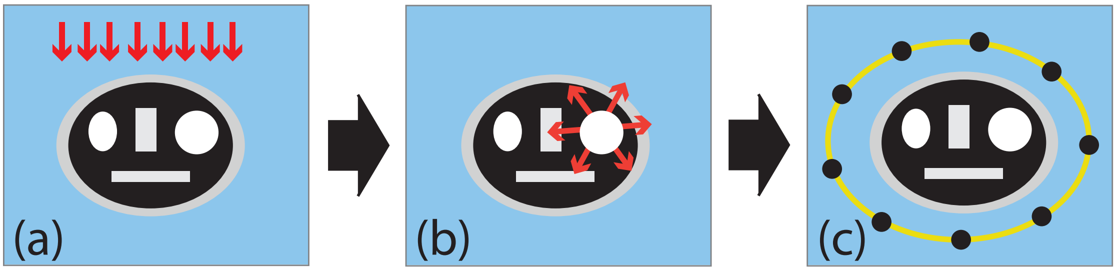

Figure 3.1: (a) An object is illuminated with a short optical pulse; (b) the absorbed light distribution causes an acoustic pressure; (c) the acoustic pressure is measured outside the object and used to reconstruct an image of the interior.

In this paper we consider PAT in heterogeneous acoustic media, where the acoustic pressure satisfies the wave equation [WA11, XW06b]

$$\partial_t^2 p(\mathbf{r},t) - c^2(\mathbf{r})\Delta_{\mathbf{r}} p(\mathbf{r},t) = \delta'(t)\, f(\mathbf{r}) \quad \text{for } (\mathbf{r},t) \in \mathbb{R}^d \times \mathbb{R}\,. \tag{3.1}$$

Here $\mathbf{r} \in \mathbb{R}^d$ is the spatial location, $t \in \mathbb{R}$ the time variable, $\Delta_{\mathbf{r}}$ the spatial Laplacian, $c(\mathbf{r})$ the speed of sound, and $f(\mathbf{r})$ the photoacoustic (PA) source that has to be recovered. The factor $\delta'(t)$ denotes the distributional derivative of the Dirac δ-distribution in the time variable and the product $\delta'(t)\, f(\mathbf{r})$ on right hand side of the equation represents the PA source due to the pulsed optical illumination. The wave equation (3.1) is augmented with the initial condition $p(\mathbf{r},t) = 0$ on $\{t < 0\}$. The acoustic pressure is then uniquely defined and referred to as the causal solution of (3.1). Both cases $d = 2, 3$ for the spatial dimension are relevant in PAT: The case $d = 3$ arises in PAT using classical point-wise measurements; the case $d = 2$ is relevant for PAT with integrating line detectors [BMFB17, BBMG$^+$07, PHKN17].

To recover the PA source, the pressure is measured with sensors distributed on a surface or curve outside of the sample; see Fig. 3.1. Using standard sensing, the spatial sampling step size limits the spatial resolution of the measured data and therefore the spatial resolution of the final reconstruction. Consequently, high spatial resolution requires a large number of detector locations. Ideally, for high frame rate, the pressure data are measured in parallel with a large array made of small detector elements. However, producing a detector array with a large number of parallel readouts is costly and technically demanding. In this work we use techniques of compressed sensing (CS) to reduce the number of required measurements and thereby accelerating PAT while keeping high spatial resolution [SKB$^+$15, ABB$^+$16, BCH$^+$17, HBMB16].

CS is a novel sensing paradigm introduced in [CRT06a, CT06, Don06a] that allows to capture high resolution signals using much less measurements than advised by Shannon's sampling theory. The basic idea is to replace point samples by linear measurements, where each measurement consists of a linear combination of sensor values. It offers the ability to reduce the number of measurements while keeping high spatial resolution. One crucial ingredient enabling CS PAT is sparsity, which refers to the requirement that the unknown signal is sparse, in the sense that it has only a small number of entries that are significantly different from zero (possibly after a change of basis).

3.1.1 Main contributions

In this work we develop a new framework for CS PAT that allows to bring sparsity into play. Our approach is rooted in the concept of sparsifying temporal transforms developed for PAT in [SKB$^+$15, HBMB16] for two and three spatial dimensions. However, the approach in this present paper extends and simplifies this transform approach considerably. First, it equally applies to any detection surface and arbitrary spatial dimension. Second, the new method can even be applied to heterogenous media. In order to achieve this, we use the second time derivative applied to the pressure data as a sparsifying transform. Opposed to [SKB$^+$15, HBMB16], where the transform was used to sparsify the measured signals, in the present work we exploit this for obtaining sparsity in the original imaging domain.

Our new approach is based on the following. Consider the second time derivative $\partial_t^2 p(\mathbf{r}, t)$ of the PA pressure. We will show that this transformed pressure again satisfies the wave equation, however with the modified PA source $c^2(\mathbf{r})\Delta_{\mathbf{r}} f(\mathbf{r})$ in place of the original PA source $f(\mathbf{r})$. If the original PA source consists of smooth parts and jumps, the modified source consists of smooth parts and sparse structures; see Fig. 3.2 for an example. This enables the use of efficient CS reconstruction algorithms based on sparse recovery. One possible approach is based on the following two-step procedure. First, recover an approximation $h \simeq c^2 \Delta_{\mathbf{r}} f$ via ℓ^1-minimization. In a second step, recover an approximation to f by solving the Poisson equation $\Delta_{\mathbf{r}} f = h/c^2$.

While the above two-stage approach very well recovers the singularities of the PA source in our performed numerical tests, at the same time it showed disturbing low-frequency arti-

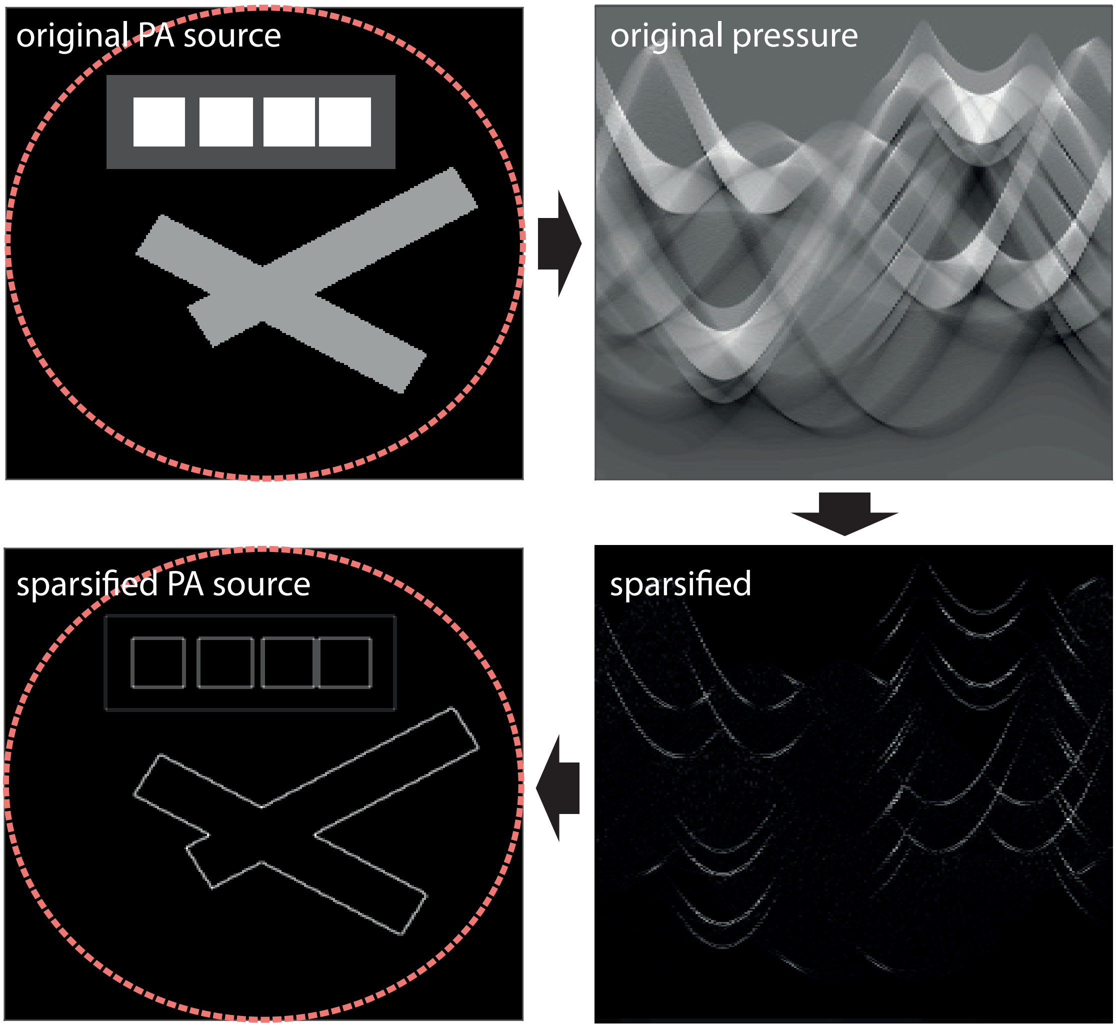

Figure 3.2: Top left: PA source f (circle indicates the detector locations). Top right: Corresponding pressure p (detector locations are in horizontal direction and time varies in vertical direction). Bottom right: Second derivative $\partial_t^2 p$ Bottom left: Sparsified source $\Delta_{\mathbf{r}} f$.

facts. Therefore, in this paper we develop a completely novel strategy to jointly recovering f as well as its Laplacian. Similarly to the two stage-procedure, the joint sparse reconstruction approach aims at finding the sparsified PA source h, corresponding to the second derivative of the measured data, via ℓ^1-minimization. Additionally, it requires the original PA source f to be consistent with the original data. Finally, the two source terms are connected by adding the constraint $\Delta_{\mathbf{r}} f = h/c^2$. The resulting optimization problem will be formally introduced in Section 3.3.3. In the case of noisy data, we consider a penalized version that can be efficiently solved via various modern optimization techniques, such as forward backward splitting.

3.1.2 Outline

In Section 3.2 we describe PAT and existing CS approaches, wherein we focus on the role of sparsity in PAT. The proposed framework for CS PAT will be presented in Section 3.3. This includes the sparsification of the PA source and the joint sparse reconstruction algorithm. Numerical and experimental results are presented in Section 3.4. The paper concludes with a summary and an outlook on future research presented in Section 3.5.

3.2 Compressed photoacoustic tomography

3.2.1 Photoacoustic tomography

Suppose that $V \subseteq \mathbb{R}^d$ is a bounded region and let $\mathbf{r}_i$ for $i = 1, \dots, n$ denote admissible detector locations distributed on the boundary ∂V. We define the forward wave operator $\mathbf{W}$ that maps a PA source f vanishing outside V to the corresponding pressure data by

$$(\mathbf{W}f)(\mathbf{r}, t) \triangleq p(\mathbf{r}, t) \qquad \text{for } \mathbf{r} \in \partial V \text{ and } t \in [0, T] \,, \tag{3.2}$$

where $p(\mathbf{r}, t)$ is the solution of (3.1). The inverse source problem of PAT is to recover the PA source f from data $\mathbf{W}f(\mathbf{r}_i, t)$, known for a finite number of admissible detector locations $\mathbf{r}_i$. The measured data $\mathbf{W}f$ is considered to be fully/completely sampled if the transducers are densely located on the whole boundary ∂V, such that the function f can be stably reconstructed from the data. Finding necessary and sufficient sampling conditions for PAT is still on-going research [Hal16].

Let us mention that most of the theoretical works on PAT consider the continuous setting where the transducer locations are all points of a surface or curve $\Gamma \subseteq \partial V$; see [FPR04, XW05, Kun07, Ngu09, Hal13]. On the other hand, most works on discrete settings consider both discrete spatial and time variables [HWNW13, Hal16]. The above setting (3.2) has been considered in a few works [SKB$^+$15, HBMB16, CN17]. It well reflects the high sampling rate of the time variable in many practical PAT systems.

3.2.2 Compressive measurements in PAT

The number n of detector positions in (3.2) is directly related to the resolution of the final reconstruction. Namely, consider the case V being the disc of radius R and f being essentially wavenumber limited with maximal wavenumber give by λ_0. Then, $N_\varphi \geq 2R\lambda_0$ equally spaced transducers are sufficient to recover f with small error; see [Hal16]. This condition, however, requires a very high sampling rate, especially when the PA source contains narrow features, such as blood vessels, or sharp interface. Consequently, full sampling in PAT is costly and time consuming.

To reduce the number of measurements while preserving resolution, we use CS measurements in PAT. Instead of collecting n individually sampled signals as in (3.2), we take CS measurements

$$y(j,t) \triangleq (\mathbf{M}f)(j,t) \triangleq (\mathbf{AW}f)(j,t) = \sum_{i=1}^{n} \mathbf{A}[j,i]p(\mathbf{r}_i,t) \quad \text{for } j \in \{1,\dots,m\}\,, \tag{3.3}$$

with $m \ll n$. [SKB$^+$15, HBMB16] we proposed to take the measurement matrix $\mathbf{A}$ in (3.3) as the adjacency matrix of a lossless expander graph. Hadamard matrices have been proposed in [BCH$^+$17, HZB$^+$16]. In this work, we take $\mathbf{A}$ as a random Bernoulli matrix with entries $\pm 1/\sqrt{m}$ with equal probability or a Gaussian random matrix consisting of i.i.d. $\mathcal{N}(0,1/m)$-Gaussian random variables in each entry. These choices are validated by the fact that Gaussian and Bernoulli random matrices satisfy the restricted isometry property (RIP) with high probability (see Section 3.2.3 below). Subsampling matrices, which do not satisfy the RIP, have have be used in [PL09].

3.2.3 The role of sparsity

A central aspect in the theory of CS is sparsity of the given data in some basis or frame [CRT06a, Don06a, FR13]. Recall that a vector $x \in \mathbb{R}^N$ is called s-sparse if $|\{i \mid x_i \neq 0\}| \leq s$ for some number $s \ll N$, where $|\cdot|$ is used to denote the number of elements of a set. If the data is known to have sparse structure, then reconstruction procedures using ℓ_1-minimization or greedy-type methods can often be guaranteed to yield high quality results even if the problem is severely ill-posed [CRT06a, GHS11]. If we are given measurements $\mathbf{M}x = y$, where $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^m$ with $m \ll N$, the success of the aforementioned reconstruction procedures can, for example, be guaranteed if the matrix $\mathbf{M}$ satisfies the restricted isometry property of order s (s-RIP), i.e. for all s-sparse vectors z we have

$$(1-\delta_s)\|z\|^2 \leq \|\mathbf{M}z\|^2 \leq (1+\delta_s)\|z\|^2\,, \tag{3.4}$$

for an RIP constant $\delta_s < 1/\sqrt{2}$; see [FR13]. Gaussian and Bernoulli random matrices satisfy the s-RIP with high probability, provided $m \geq Cs\log(\mathrm{e}n/s)$ for some reasonable constant C and with e denoting Euler's number [BDDW08].

In PAT, the possibility to sparsify the data has recently been examined [SKB$^+$15, HBMB16].

In these works it was observed that the measured pressure data could be sparsified and the sparse reconstruction methods were applied directly to the pressure data. As a second step, still a classical reconstruction via filtered backprojection had to be performed. The sparsification of the data was achieved with a transformation in the time direction of the pressure data. In two dimensions, the transform is a first order pseudo-differential operator [SKB$^+$15], while in three dimensions the transform is of second order [HBMB16].

3.3 Proposed Framework

3.3.1 Sparsifying transform

The following theorem is the foundation of our CS approach. It shows that the required sparsity in space can be obtained by applying the second time derivative to the measured data.

Theorem 3.3.1. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a given smooth source term vanishing outside V, and let $p(\mathbf{r},t)$ denote the causal solution of* (3.1). *Then $\partial_t^2 p(\mathbf{r},t)$ is the causal solution of*

$$\partial_t^2 p'' - c^2(\mathbf{r})\Delta_{\mathbf{r}} p''(\mathbf{r},t) = \delta'(t)\, c^2(\mathbf{r})\Delta_{\mathbf{r}} f(\mathbf{r})\,. \tag{3.5}$$

In particular, we have $\partial_t^2 \mathbf{M}[f] = \mathbf{M}[c^2\Delta_{\mathbf{r}} f]$, where $\mathbf{M}$ denotes the CS PAT forward operator defined by (3.3).

Proof. The proof is split in two parts. First, we show that the causal solution of (3.1) is, for positive times, equal to the solution of the initial value problem

$$(\partial_t^2 - c^2\Delta_{\mathbf{r}})p(\mathbf{r},t) = 0 \qquad \text{for } (\mathbf{r},t) \in \mathbb{R}^d \times (0,\infty)\,, \tag{3.6}$$

$$p(\mathbf{r},0) = f(\mathbf{r}) \qquad \text{for } \mathbf{r} \in \mathbb{R}^d\,, \tag{3.7}$$

$$\partial_t p(\mathbf{r},0) = 0 \qquad \text{for } \mathbf{r} \in \mathbb{R}^d\,. \tag{3.8}$$

Second, we take the second time derivative of the even extension and derive (3.5).

To see the equivalence of (3.6)-(3.8) and (3.1), denote by q the even extension of the solution of (3.6)-(3.8) to $\mathbb{R}^d \times \mathbb{R}$, which satisfied the homogeneous wave equation for all $t \in \mathbb{R}$. By using the characteristic function of the positive semiaxis, $\chi(\mathbf{r},t) := 1$ if $t > 0$ and zero otherwise, we obtain that $\chi \cdot q$ is a causal function. By the product rule, we have

$$(\partial_t^2 - c^2\Delta_{\mathbf{r}})(\chi q) = \chi \underbrace{(\partial_t^2 - c^2\Delta_{\mathbf{r}})q}_{=0} + \underbrace{\delta' q}_{=\delta' f} + \underbrace{\delta q'}_{=0} = \delta' f\,. \tag{3.9}$$

This shows that χq is the (unique) causal solution of (3.1) and establishes the equivalence of the two formalisms. Next notice that $\partial_t^2 q$ satisfies the wave equation as well, $(\partial_t^2 - c^2 \Delta_{\mathbf{r}})(\partial_t^2 q)(\mathbf{r}, t) = 0$. Therefore, a computation similar to (3.9) gives $(\partial_t^2 - c^2 \Delta_{\mathbf{r}})(\chi \partial_t^2 q) = \delta' c^2 \Delta_{\mathbf{r}} f + \delta \partial_t^3 q(\cdot, 0)$. Because q is smooth and even in t, its third time derivative is continuous and odd in t, and thus $\partial_t^3 q(\mathbf{r}, 0) = 0$. This shows that $\chi \partial_t^2 q$ is the causal solution of (3.5) and concludes the proof. □

3.3.2 Spatial discretization and sparse recovery

Throughout the following, we assume a semi-discrete setting, where the spatial variable $\mathbf{r} \in \{0, \dots, N_{\mathbf{r}}\}^d$ is already discretized. As we show in Theorem 3.6.1 in the appendix, an analog of Theorem 3.3.1 for discrete sources $f \colon \{0, \dots, N_{\mathbf{r}}\}^d \to \mathbb{R}$ holds as well, where the discrete Laplacian $\Delta_{\mathbf{r}}$ may be defined in the spatial domain using finite differences or in the spectral domain using the Fourier transform. Working with discrete sources significantly simplifies treating the sparsity of $\Delta_{\mathbf{r}} f$. In the continuous setting applying the Laplacian to piecewise smooth functions would require working with generalized functions (or distributions), which is completely avoided by taking the discretized spatial variable (which is anyway required for the numerical implementation).

Typical phantoms consist of smoothly varying parts and rapid changes at interfaces. For such PA sources, the modified source $c^2 \Delta_{\mathbf{r}} f$ is sparse or at least compressible. The theory of CS therefore predicts that the modified source can be recovered by solving

$$\min_h \|h\|_1 \quad \text{such that } \mathbf{M} h = \partial_t^2 y \,, \tag{3.10}$$

where $\mathbf{M}$ is the forward operator (including wave propagation and compressed measurements) and $\|\cdot\|_1$ denotes the ℓ^1-norm guaranteeing sparsity of h.

In the case the unknown is only approximately sparse or the data are noisy, one instead minimizes the penalized functional problem $\frac{1}{2}\|\mathbf{M}h - \partial_t^2 y\|_2^2 + \beta \|h\|_1$, where $\beta > 0$ is a regularization parameter which gives trade-off between the data-fitting term $\|\mathbf{M}h - \partial_t^2 y\|_2^2$ and the regularization term $\|h\|_1$. Having obtained an approximation of h by either solving (3.10) or the relaxed version, one can recover the original PA source f by subsequently solving the Poisson equation $\Delta_{\mathbf{r}} f = h/c^2$ with zero boundary conditions.

While the above two-stage procedure recovers boundaries well, we observed disturbing low frequency artifacts in the reconstruction of f. Therefore, below we introduce a new joint sparse reconstruction approach based on Theorem 3.3.1 that jointly recovers f and h.

3.3.3 Joint sparse reconstruction approach

As argued above, the second derivative p'' is well suited (via $c^2(\mathbf{r})\Delta_{\mathbf{r}} f$) to recover the singularities of f, but hardly contained low-frequency components of f. On the other hand, the low frequency information is contained in the original data, which is still available to us. Therefore we propose the following joint sparsity constrained optimization problem

$$\begin{aligned}&\min_{(f,h)} \|h\|_1 + I_C(f)\\ &\text{such that } \Big[\mathbf{M}f, \mathbf{M}h, \Delta_{\mathbf{r}} f - h/c^2\Big] = \Big[y, y'', 0\Big] \,.\end{aligned} \tag{3.11}$$

Here I_C is the indicator function of the positive cone $C \triangleq \{f \mid f(\mathbf{r}) \geq 0\}$, defined by $I_C(f) = 0$ if $f \in C$ and $I_C(f) = \infty$ and guaranteeing non-negativity.

We have the following result.

Theorem 3.3.2. *Assume that $f\colon \{0, \dots, N_{\mathbf{r}}\}^d \to \mathbb{R}$ is non-negative and that $\Delta_{\mathbf{r}} f$ is s-sparse. Moreover, suppose that the measurement matrix $\mathbf{M}$ satisfies the RIP of order $2s$ with $\delta_{2s} < 0.6246$ (see* (3.4)*) and denote $y = \mathbf{M}f$. Then, the pair $[f, \Delta_{\mathbf{r}} f]$ can be recovered as the unique solution of* (3.11).

Proof. According to Theorem 3.6.1, the second derivative $p''(\mathbf{r}, t)$ is the unique causal solution of the wave equation (3.5) with modified source term $h = c^2(\mathbf{r})\Delta_{\mathbf{r}} f$. As a consequence $c^2(\mathbf{r})\Delta_{\mathbf{r}} h$ satisfies $\mathbf{M}h = y''$, which implies that the pair $[f, h]$ is a feasible solution for (3.11). It remains to verify that $[f, h]$ is the only solution of (3.11). To show this, note that for any solution $[f^*, h^*]$ of (3.11) its second component h^* is a solution of (3.10). Because $\mathbf{M}$ satisfies the $2s$-RIP, and $h = c^2(\mathbf{r})\Delta_{\mathbf{r}} f$ is s-sparse, CS theory implies that (3.10) is uniquely solvable (see Theorem 6.12 in [FR13]) and therefore $h = h^*$. The constraint $\Delta_{\mathbf{r}} f = h/c^2$ then implies that $f^* = f$. □

In the case the data only approximately sparse or noisy, we propose, instead of (3.11), to solve the ℓ^2-relaxed version

$$\frac{1}{2}\|\mathbf{M}f - y\|_2^2 + \frac{1}{2}\|\mathbf{M}h - y''\|_2^2 + \frac{\alpha}{2}\|\Delta_{\mathbf{r}} f - h/c^2\|_2^2 + \beta\|h\|_1 + I_C(f) \to \min_{(f,h)} \,. \tag{3.12}$$

Here $\alpha > 0$ is a tuning and $\beta > 0$ a regularization parameter. There are several modern methods to efficiently solve (3.12), for example, proximal or stochastic gradient algorithms[Ber11, CP11]. In this work we use a proximal forward-backward splitting method with the quadratic terms used in the explicit (forward) step and $\beta\|h\|_1 + I_C(f)$ for the implicit (backward) step. For an overview over proximal forward-backward splitting methods in signal processing, see for example [CP11].

3.3.4 Numerical minimization

We will solve (3.12) using a proximal gradient algorithm [CP11], which is an algorithm well suited for minimizing the sum of a smooth and a non-smooth but convex part. In the case of (3.12) we take the smooth part as

$$\Phi(f,h) \triangleq \frac{1}{2}\|\mathbf{M}f - y\|_2^2 + \frac{1}{2}\|\mathbf{M}h - y''\|_2^2 + \frac{\alpha}{2}\|\Delta_{\mathbf{r}} f - h/c^2\|_2^2 \tag{3.13}$$

and the non-smooth part as $\Psi(f,h) \triangleq \beta\|h\|_1 + I_C(f)$.

The proximal gradient algorithm then alternately performs an explicit gradient step for Φ and an implicit proximal step for Ψ. For the proximal step, the proximity operator of a function must be computed. The proximity operator of a given convex function $F\colon \mathbb{R}^d \to \mathbb{R}$ is defined by [CP11]

$$\operatorname{prox}_F(f) \triangleq \operatorname{argmin}\{F(z) + \tfrac{1}{2}\|f - z\|_2^2 \mid z \in \mathbb{R}^d\}\,.$$

The regularizers we are considering here have the advantage, that their proximity operators can be computed explicitly and do not cause a significant computational overhead. The gradient $[\nabla_f \Phi, \nabla_h \Phi]$ of the smooth part can easily be computed to be

$$\begin{aligned}\nabla_f \Phi(f,h) &= \mathbf{M}^*(\mathbf{M}f - y) - \alpha\Delta_{\mathbf{r}}(\Delta_{\mathbf{r}} f - h/c^2)\\ \nabla_h \Phi(f,h) &= \mathbf{M}^*(\mathbf{M}h - y'') - \frac{\alpha}{c^2}(\Delta_{\mathbf{r}} f - h/c^2)\,.\end{aligned}$$

The proximal operator of the non-smooth part is given by

$$\begin{aligned}\operatorname{prox}(f,h) &:= [\operatorname{prox}_{I_C}(f), \operatorname{prox}_{\beta\|\cdot\|_1(h)}]\,,\\ \operatorname{prox}_{I_C}(f)_i &= (\max(f_i, 0))_i\,,\\ \operatorname{prox}_{\beta\|\cdot\|_1}(h)_i &= (\max(|h_i| - \beta, 0)\operatorname{sign}(h_i))_i\end{aligned}$$

With this, the proximal gradient algorithm is given by

$$f^{k+1} = \operatorname{prox}_{I_C}\left(f^k - \mu_k \nabla_f \Phi(f^k, h^k)\right) \tag{3.14}$$

$$h^{k+1} = \operatorname{prox}_{\beta\|\cdot\|_1}\left(h^k - \mu_k \nabla_k \Phi(f^k, h^k)\right), \tag{3.15}$$

where (f^k, h^k) is the k-th iterate and μ_k the step size in the k-th iteration. We initialize the proximal gradient algorithm with $f^0 = h^0 = 0$.

Remark 3.3.1. Note that the optimization problem (3.11) is further equivalent to the analysis-ℓ_1

problem

$$\begin{aligned} &\min_f \|c\Delta_{\mathbf{r}} f\|_1 + I_C(f) \\ &\text{such that } \mathbf{M}f = y\,. \end{aligned} \tag{3.16}$$

Implementation of (3.16) avoids taking the second time derivative of the data y. Because the proximal map of $f \mapsto \|c\Delta_{\mathbf{r}} f\|_1$ in not available explicitly, (3.16) and its relaxed versions cannot be straightforwardly addressed with the proximal gradient algorithm. Therefore, in the present paper we only use the model (3.12) and the algorithm (3.14), (3.15) for its minimization. Different models and algorithms will be investigated in future research.

3.4 Experimental and numerical results

3.4.1 Numerical results

For the presented numerical results, the two dimensional PA source term $f\colon \{0,\dots,N_{\mathbf{r}}\}^2 \to \mathbb{R}$ depicted in Figure 3.2 is used which is assumed to be supported in a disc of radius R. For simplicity, we assume the speed of sound c to be constant. Additional results are presented using an MRI image. The synthetic data is recorded on the boundary circle of radius R, where the the time was discretized with 301 equidistant sampling points in the interval $[0, 2R/c]$.

The phantom images were then used to compute the pressure data at $n = 200$ equispaced detector locations in a circle around the source. For the simulated experiments, we assume ideal detectors with $\delta(t)$ impulse response. The reconstruction of both phantoms via the filtered backprojection algorithm of [FPR04] from the full measurements is shown in Figure 3.3.

CS measurements $\mathbf{M}f$ (see (3.3)) have been generated in two random ways and one deterministic way. The random matrices $\mathbf{A}$ have be taken either as random Bernoulli matrix with entries $\pm 1/\sqrt{m}$ with equal probability or a Gaussian random matrix consisting of i.i.d. $\mathcal{N}(0, 1/m)$-Gaussian random variables in each entry. The deterministic subsampling was performed by choosing m equispaced detectors. In the joint sparse reconstruction with (3.14), (3.15), the step-size and regularization parameters are chosen to be $\mu_k = 0.1$, $\alpha = 0.1$ and $\beta = 0.005$; 5000 iterations are performed. For the random subsampling matrices the recovery guarantees from the theory of CS can be employed, but they do not provably hold for the deterministic subsampling - although the results are equally convincing even for subsampling on the order of $10\,\%$ of the full data, cf. Figure 3.4. All results are compared to the standard filtered backprojection (FBP) reconstruction applied to $\mathbf{A}^T(\mathbf{M}f)$.

If one increases the number of measurements to about $25\,\%$ of the data, the reconstruction results become almost indistinguishable from the results obtained from FBP on the full data, cf. Figure 3.5.

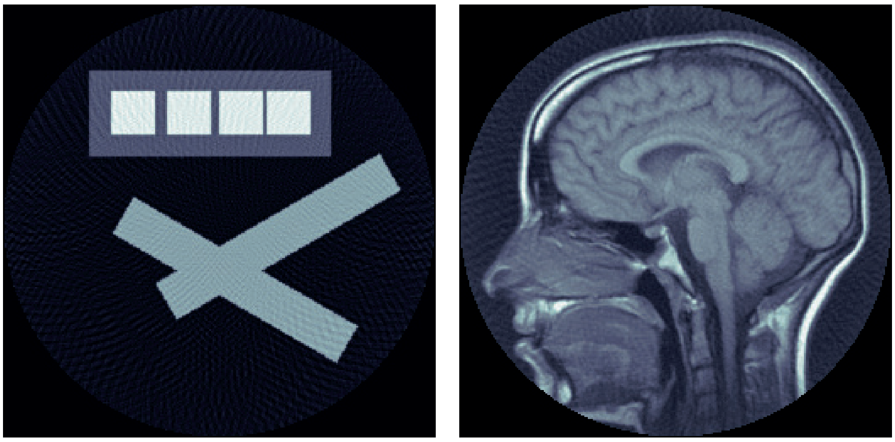

Figure 3.3: RECONSTRUCTIONS FROM FULL DATA (CROSS PHANTOM AND MRI PHANTOM): Reconstruction from 200 equispaced (fully sampled) measurements of the cross and the MRI image used in the numerical experiments.

In Figure 3.6 the application of the method developed in this article to an MRI image is presented. As the sparsity of the Laplacian is not as pronounced as in the synthetic example, the MRI image requires more measurements to achieve high qualitative outcome.

For noisy data, the algorithm still produces good results, although more samples need to be taken to achieve good results. For the synthetic phantom, Gaussian noise amounting to an SNR of approximately $15\,\mathrm{dB}$ was added, as shown in Figures 3.7 and 3.8. The reconstruction results using $m = 20$ and $m = 50$ measurements are depicted in Figure 3.9 and Figure 3.10, respectively.

3.4.2 Experimental results

Experimental data have been acquired by an all-optical photoacoustic projection imaging (O-PAPI) system as described in [BMFB17]. The system featured 64 integrating line detector (ILD) elements distributed along a circular arc of radius $4\,\mathrm{cm}$, covering an angle of $289\,\mathrm{degree}$. For an imaging depth of $20\,\mathrm{mm}$, the imaging resolution of the O-PAPI system was estimated to be between $100\,\mu\mathrm{m}$ and $260\,\mu\mathrm{m}$; see [BMFB17]. PA signals were excited by illuminating the sample from two sides with pulses from a frequency-doubled Nd:YAG laser (Continuum Surelite, $20\,\mathrm{Hz}$ repetition rate, $6\,\mathrm{ns}$ pulse duration, $532\,\mathrm{nm}$ center wavelength) at a fluence of $21\,\mathrm{mJ/cm}^2$ and recorded by the ILD elements with a sample rate of $60\,\mathrm{MS/s}$. The sample consisted an approximately triangular shaped piece of ink-stained leaf skeleton, embedded in a cylinder consisting of agarose gel with a diameter of $36\,\mathrm{mm}$ and a height of $40\,\mathrm{mm}$. Intralipid was added to the agarose to increase optical scattering. The strongest branches of the leaf

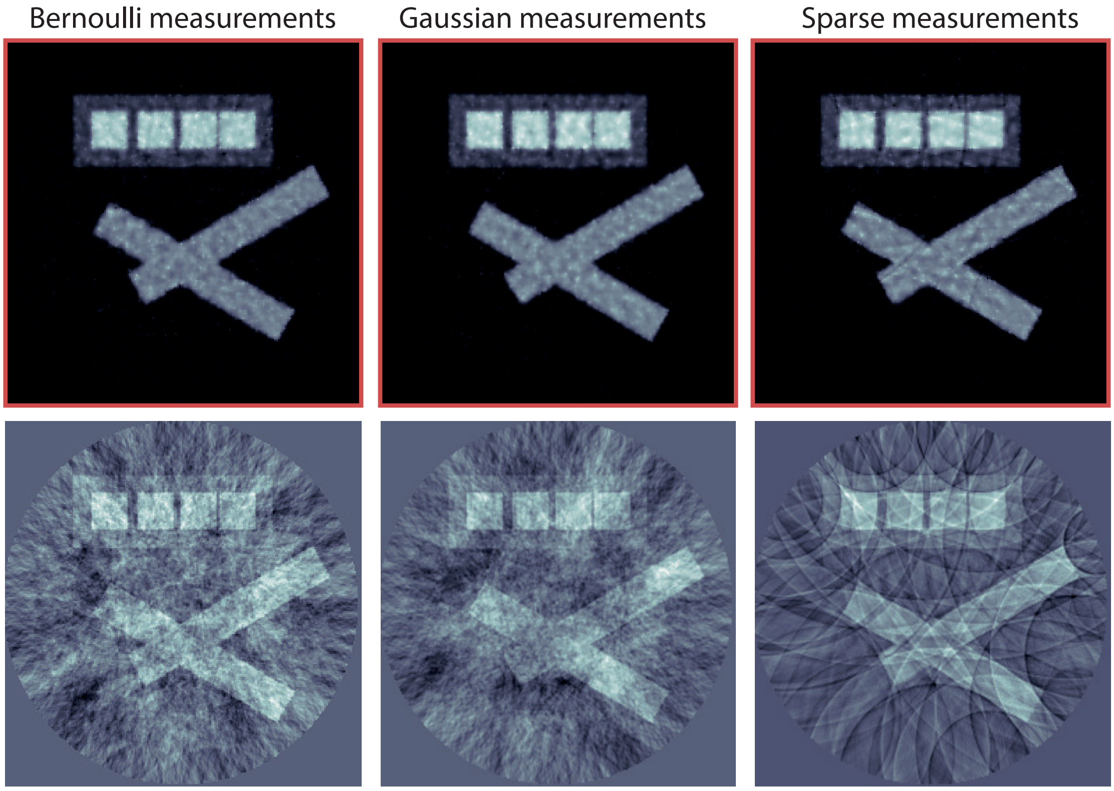

Figure 3.4: Reconstructions from 20 noisefree measurements (cross phantom): Reconstruction from $m = 20$ noisefree CS measurements (i.e. $10\,\%$ of the fully sampled data) using the method presented in this article (top) and FBP (bottom).

had diameters of approximately $160\,\mu\text{m}$ to $190\,\mu\text{m}$ and the smallest branches of about $50\,\mu\text{m}$. Results are only for 2D (projection imaging).

Reconstruction results for the leaf phantom from 60 sparsely sampled sensor locations after 500 iterations with the proposed joint sparse minimization algorithm are shown in Figure 3.11. For this, the regularization and step-size parameters were chosen as in the previous section. From the experimental data we also generated $m = 30$ random Bernoulli measurements. The reconstruction results using this data are shown in Figure 3.12. For all results, the PA source is displayed on a $1.6\,\text{cm} \times 1.33\,\text{cm}$ rectangle with step size $26\,\mu\text{m}$ inside the detection arc.

3.4.3 Discussion

All numerical results presented in Figures (3.4)–(3.12) demonstrate that the proposed joint sparse reconstruction method yields artifact-free reconstructions from a smaller number of spatial measurements. In particular, our approach demonstrated robustness with respect to noise

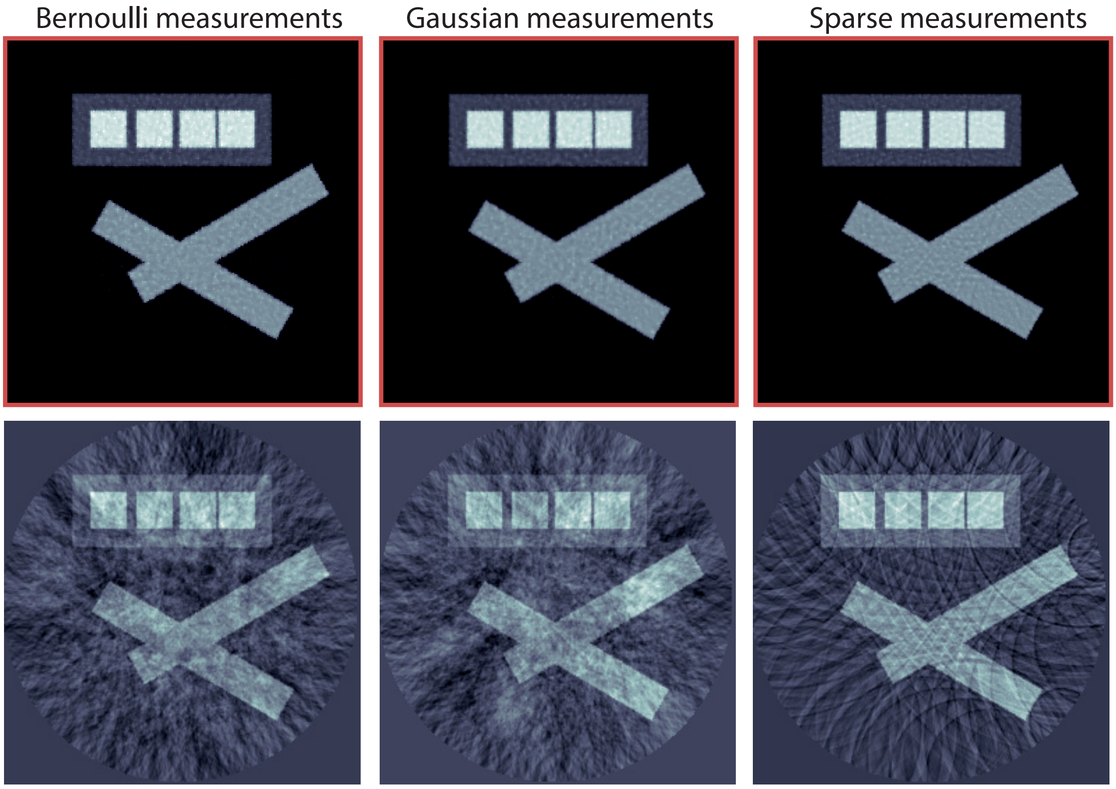

Figure 3.5: RECONSTRUCTIONS FROM 50 NOISEFREE MEASUREMENTS (CROSS PHANTOM): Reconstruction from $m = 50$ (i.e. $25\,\%$ of the fully sampled data) noisefree CS measurements using the method presented in this article (top) and FBP (bottom).

(see Figures 3.9 and 3.10) and also good performance on experimental data (see Figures 3.12 and 3.11). In the case of the cross phantom we even obtained almost artifact free reconstructions from only 20 spatial measurements (3.4).

We point out, that opposed to most existing reconstruction schemes for sparse data, our algorithms come with a uniqueness result given in Theorem 3.3.2. Theorem 3.3.2 is based on the RIP which is satisfied, for example, by Gaussian and Bernoulli measurements matrices. It does not hold for the subsampling matrices and, therefore, it is surprising that even in this case, the joint sparse reconstruction methods shows similar performance as for Gaussian or Bernoulli matrices. This fact will be investigated theoretically and numerically in a future work. We also will compare our approach with other standard methods such as TV minimization.

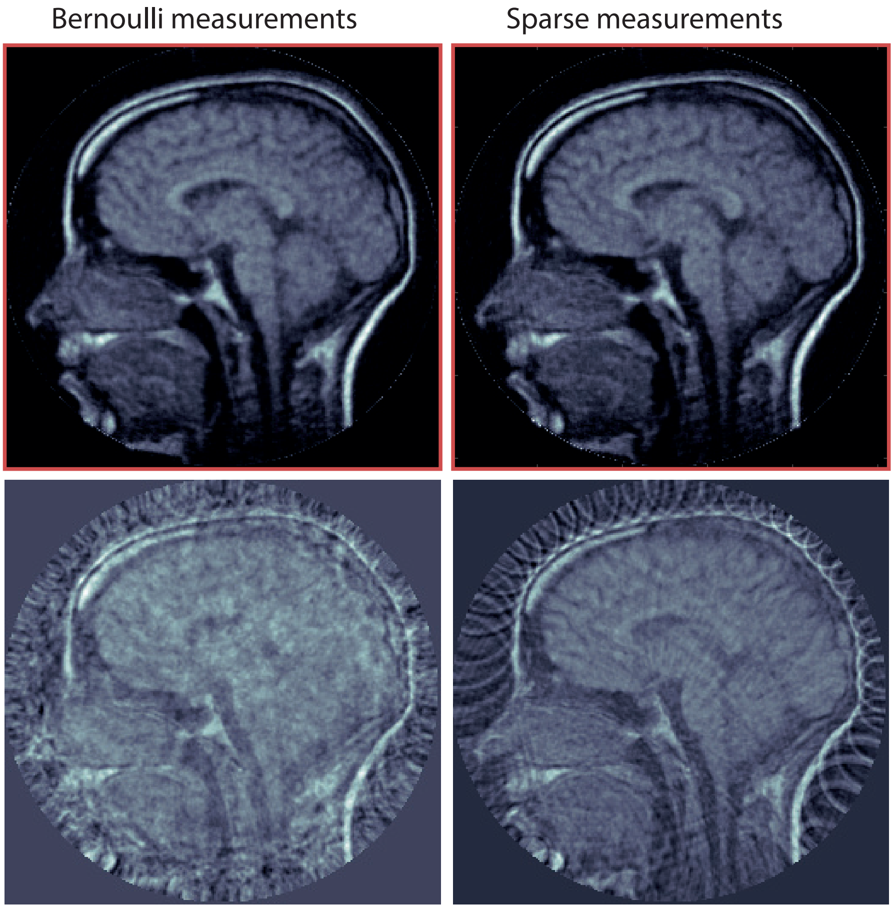

Figure 3.6: RECONSTRUCTIONS FROM 60 NOISEFREE MEASUREMENTS (MRI PHANTOM): Reconstruction of an MRI image from $m = 60$ (i.e. $33\,\%$ of the fully sampled data) synthetically generated noisefree measurements using the method presented in this article (top) and FBP (bottom).

3.5 Conclusion

In order to achieve high spatial resolution in PAT, standard measurement and reconstruction schemes require a large number of spatial measurements with high bandwidth detectors. In order to speed up the measurement process, systems allowing a large number of parallel measurements are desirable. However such systems are technically demanding and costly to fabricate. For example, in PAT with integrating detectors, the required analog to digital converters are among the most costly building blocks. In order to increase measurement speed and to minimize system costs, CS aims to reduce the number of measurements while preserving high resolution of the reconstructed image.

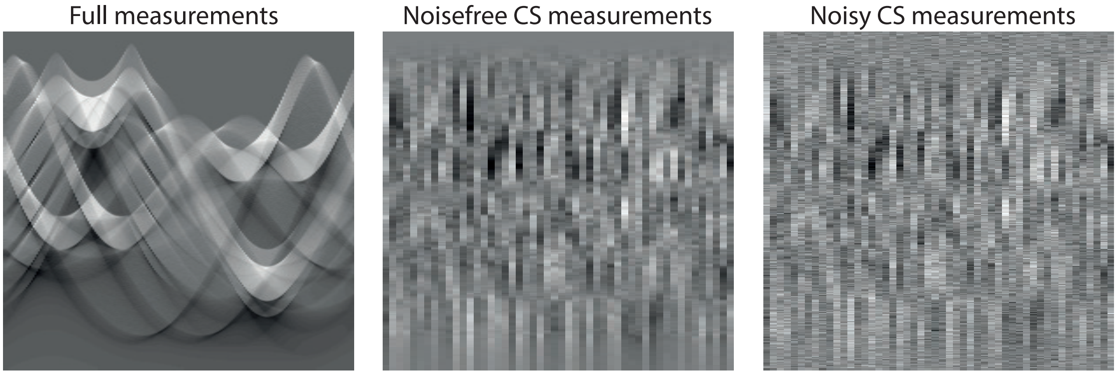

Figure 3.7: Measurements for the setup of Figure 3.10. The leftmost image shows the full measurements for all 200 detector locations. The center image shows the 50 compressed sensing measurements obtained by using a Gaussian random matrix, the right image shows the CS measurements with added noise.

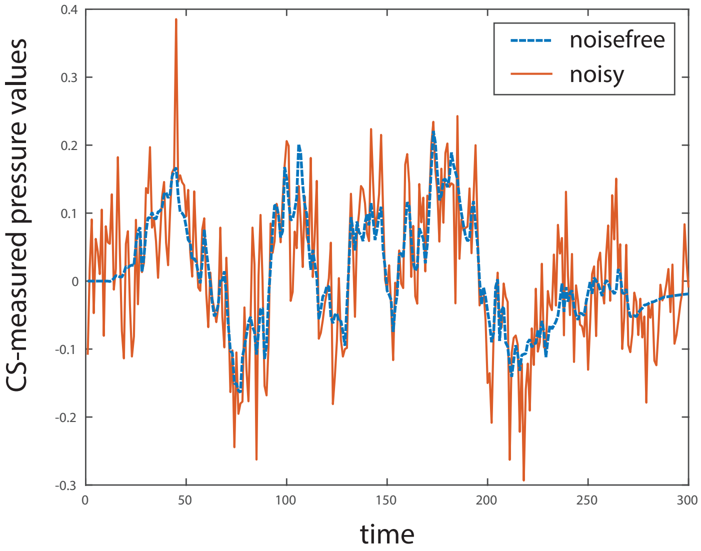

Figure 3.8: Measurements at a single detector for all times in the setup of Figure 3.10. The blue dash-dotted line shows the original data, the red solid line is the noisy data.

One main ingredient enabling CS in PAT is sparsity of the image to be reconstructed. To bring sparsity into play, in this paper we introduced a new approach based on the commutation relation $\partial_t^2 \mathbf{W}[f] = \mathbf{W}[c^2 \Delta_{\mathbf{r}} f]$ between the PAT forward operator $\mathbf{W}$ and the the Laplacian. We developed a new reconstruction strategy for jointly reconstructing the pair $[f, \Delta_{\mathbf{r}} f]$ by minimizing (3.12) and thereby using sparsity of $\Delta_{\mathbf{r}} f$. The commutation relation further allows to rigorously

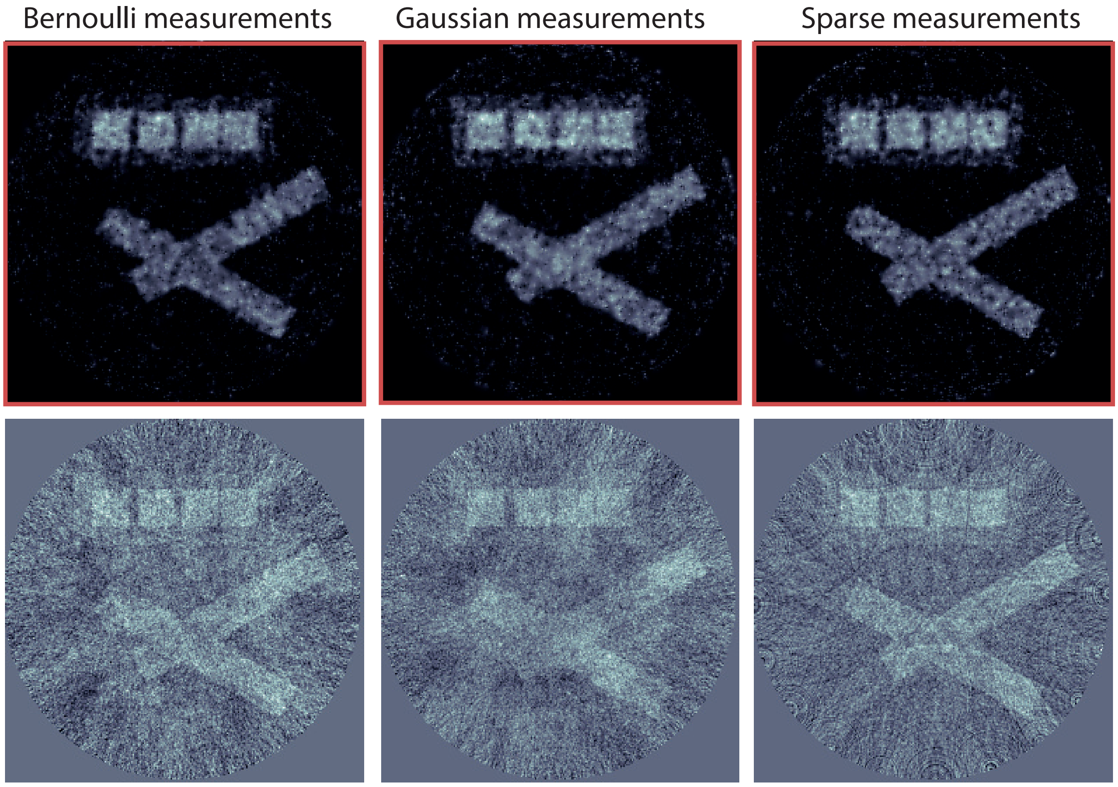

Figure 3.9: RECONSTRUCTIONS FROM 20 NOISY MEASUREMENTS (CROSS PHANTOM): Reconstruction from $m = 20$ (i.e. $10\,\%$ of the fully sampled data) noisy measurements using the method presented in this article (top) and FBP (bottom).

study generalized Tikhonov regularization of the form $\frac{1}{2}\|\mathbf{M}f - y\|_2^2 + \beta\|c\Delta_{\mathbf{r}}f\|_1 + I_C(f)$ for CS PAT. Such an analysis as well as the development of more efficient numerical minimization schemes are subjects of further research.

3.6 Sparsifying concept in semi-discrete setting

Suppose the discrete source is given by $f\colon \{0,\dots,N_{\mathbf{r}}\}^d \to \mathbb{R}$ which we can identify with $f\colon \mathbb{Z}^d \to \mathbb{R}$ by assigning the value $f(\mathbf{r}) = 0$ for $\mathbf{r} \notin \{0,\dots,N_{\mathbf{r}}\}^d$. We consider the spatially discretized wave equation

$$\partial_t^2 p(\mathbf{r},t) - c^2(\mathbf{r})\Delta_{\mathbf{r}} p(\mathbf{r},t) = \delta'(t)\, f(\mathbf{r}) \quad \text{for } (\mathbf{r},t) \in \mathbb{Z}^d \times \mathbb{R}\,, \tag{3.17}$$

together with the causality condition $p(\mathbf{r},t) = 0$ for $t < 0$. To be specific, we take the discrete Laplacian $\Delta_{\mathbf{r}} p(\mathbf{r},t)$ to be defined via symmetric finite differences. However, we note that al-

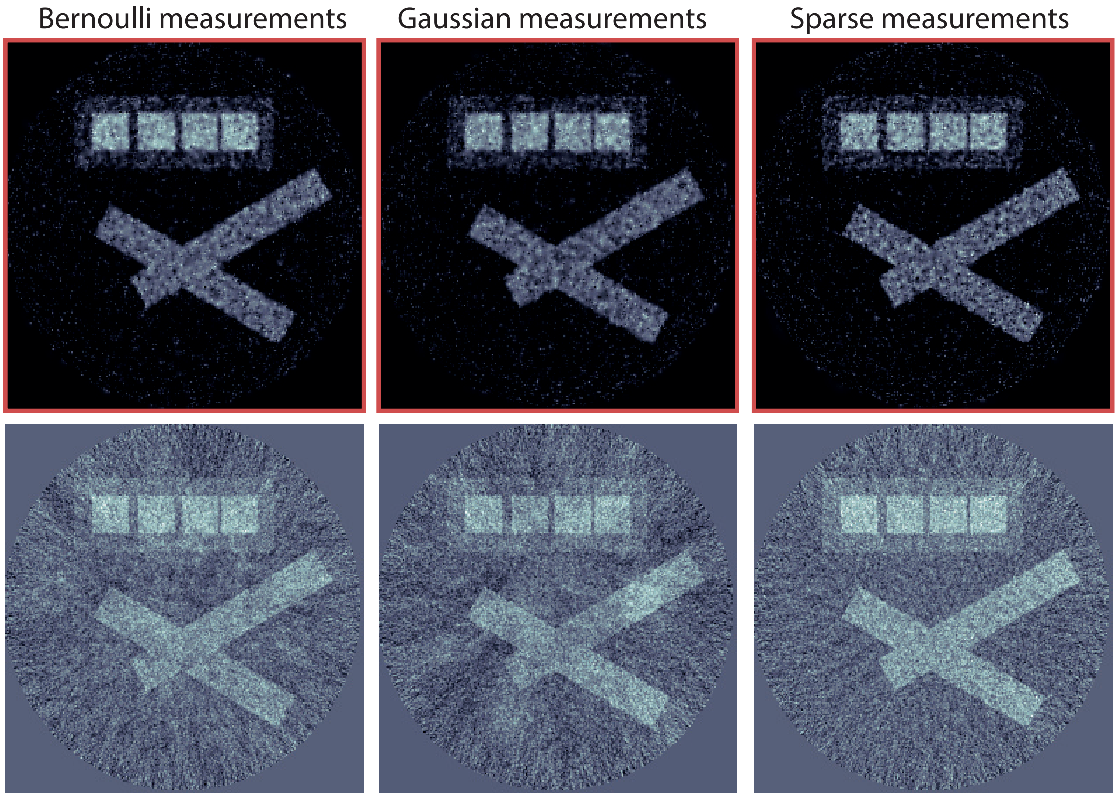

Figure 3.10: RECONSTRUCTIONS FROM 50 NOISY MEASUREMENTS (CROSS PHANTOM): Reconstruction from $m = 50$ (i.e. $25\,\%$ of the fully sampled data) noisy measurements using the method presented in this article (top) and FBP (bottom).

ternatively, it may be also defined in the spectral domain via the time-discrete spatial Fourier transform.

Note that we assume (3.17) to be satisfied for all $\mathbf{r} \in \mathbb{Z}^d$ (instead of finitely many sampling points) in order not to suffer from boundary effects wich would be induced by considering a finite spatial domain. The spatially discretized wave equation (3.17) is therefore an infinite dimensional coupled system of second order differential equations that can be analyzed in an abstract Hilbert space.

Remark 3.6.1 (Existence and uniqueness of (3.17))**.** Consider the (two-sided) initial value prob-

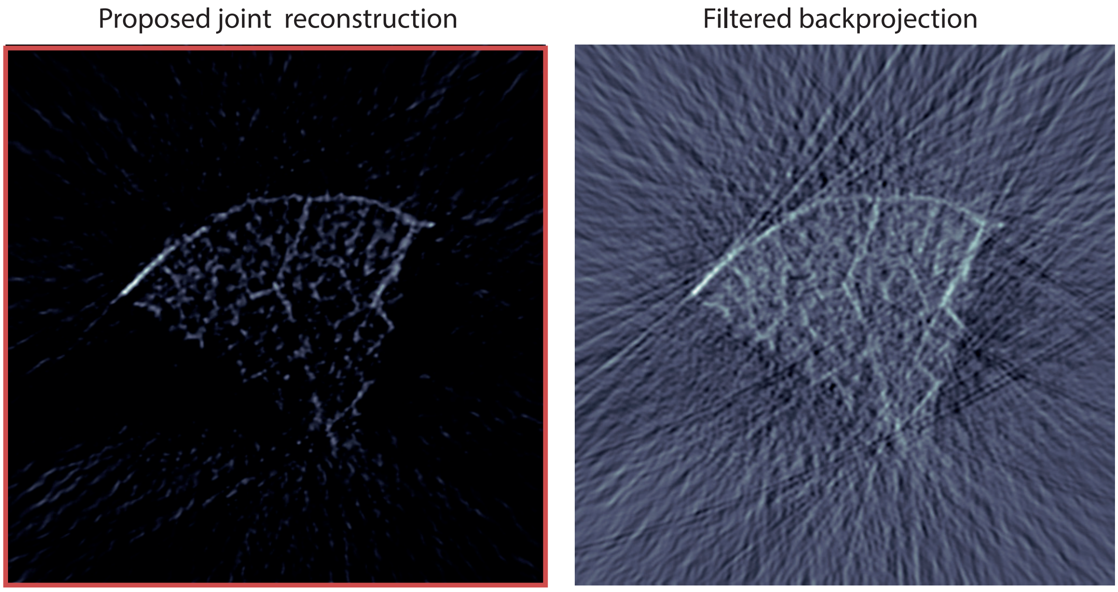

Figure 3.11: Reconstruction from experimental data using 60 sparse samples: Left: PAT image reconstructed with the proposed method. Right: FBP reconstruction.

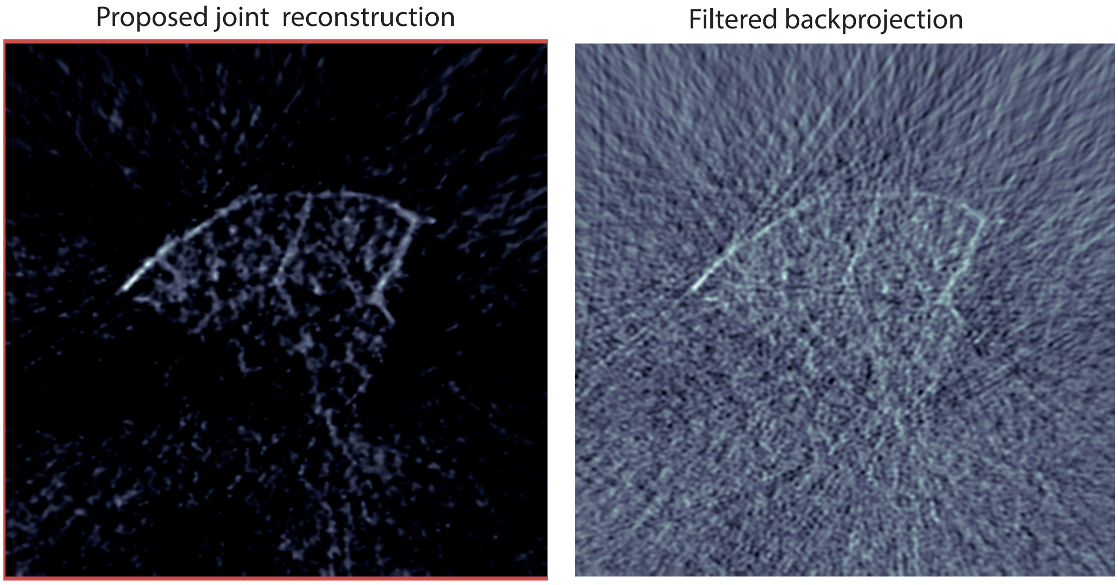

Figure 3.12: Reconstruction using 30 Bernoulli measurements: Left: PAT image reconstructed with the proposed method. Right: FBP reconstruction.

lem

$$
\begin{aligned}
(\partial_t^2 - \mathcal{L}_{\mathbf{r}})q(\mathbf{r}, t) &= 0\,, && \text{for } (\mathbf{r}, t) \in \mathbb{Z}^d \times \mathbb{R}\,, && (3.18)\\
q(\mathbf{r}, 0) &= f(\mathbf{r}) && \text{for } \mathbf{r} \in \mathbb{Z}^d\,, && (3.19)\\
\partial_t q(\mathbf{r}, 0) &= 0 && \text{for } \mathbf{r} \in \mathbb{Z}^d\,, && (3.20)
\end{aligned}
$$

where $\mathcal{L}_{\mathbf{r}} \colon \ell^2(\mathbb{Z}^d) \to \ell^2(\mathbb{Z}^d)$ is linear and bounded and $f \in \ell^2(\mathbb{Z}^d)$. From general theory of evolution equations in Hilbert spaces [Sho10], it follows that (3.18)–(3.20) has a unique solution $p \in C^\infty(\mathbb{R}, \ell^2(\mathbb{Z}^d))$. As in the case of a continuous spatial variable, this shows that χq solves (3.17) for general $\mathcal{L}_{\mathbf{r}}$ in place of $c^2(\mathbf{r})\Delta_{\mathbf{r}}$, and therefore assures the existence of a causal solution of (3.17). It uniqueness is implied by the uniqueness of a solution of (3.18)–(3.20).

The following Theorem establishes a discrete version of the Theorem 3.3.1 concerning the sparsification of PA source.

Theorem 3.6.1. *Suppose the discrete source $f \colon \mathbb{Z}^d \to \mathbb{R}$ vanishes outside the set $\{0, \dots, N_{\mathbf{r}}\}^d$, let $\mathcal{L}_{\mathbf{r}} \colon \ell^2(\mathbb{Z}^d) \to \ell^2(\mathbb{Z}^d)$ be a continuous linear operator and let p denote the unique causal solution of* (3.17). *Then $\partial_t^2 p$ is the unique causal solution of*

$$
\partial_t^2 p'' - \mathcal{L}_{\mathbf{r}} p''(\mathbf{r}, t) = \delta'(t)\, \mathcal{L}_{\mathbf{r}} f(\mathbf{r}) \quad \textit{for } (\mathbf{r}, t) \in \mathbb{Z}^d \times \mathbb{R}\,. \tag{3.21}
$$

Proof. Taking into account Remark 3.6.1, the conclusion can be shown in a similar fashion as for the continuous version (Theorem 3.3.1). □

Acknowledgments

Linh Nguyen's research is supported by the NSF grants DMS 1212125 and DMS 1616904. Markus Haltmeier and Thomas Berer acknowledge support of the Austrian Science Fund (FWF), project P 30747. Michael Sandbichler was supported by the Austrian Science Fund (FWF) under Grant no. Y760. In addition, the computational results presented have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck.

4 Total Variation Minimization in Compressed Sensing

This chapter is based on joint work with

F. KRAHMER[1]
C. KRUSCHEL

It has been published in [KKS17].

[1] Faculty of Mathematics, Office No.: 02.10.039, Boltzmannstraße 3, 85748 Garching (Munich),Germany. E-mail: felix.krahmer@tum.de

Interlude

In Chapter 2, we saw that some instability problems could be avoided when using TV minimization instead of ℓ_1-minimization. This chapter gives an overview over recovery guarantees for total variation minimization in compressed sensing for different measurement scenarios. In addition to summarizing the results in the area, we illustrate why an approach that is common for synthesis sparse signals fails and different techniques are necessary. Lastly, we discuss a generalizations of recent results for Gaussian measurements to the subgaussian case.

4.1 Introduction

The central aim of Compressed Sensing (CS) [CRT06a, Don06a] is the recovery of an unknown vector from very few linear measurements. Put formally, we would like to recover $x \in \mathbb{R}^n$ from $y = Ax + e \in \mathbb{R}^m$ with $m \ll n$, where e denotes additive noise.

For general x, recovery is certainly not possible, hence additional structural assumptions are necessary in order to be able to guarantee recovery. A common assumption used in CS is that the signal is *sparse*. Here for x we assume

$$\|x\|_0 := |\{k \in [n] \colon \ x_k \neq 0\}| \leq s,$$

that is, there are only very few nonzero entries of x. And say that x is s-sparse for some given sparsity level $s \ll n$. We call a vector *compressible*, if it can be approximated well by a sparse vector. To quantify the quality of approximation, we let

$$\sigma_s(x)_q := \inf_{\|z\|_0 \leq s} \|z - x\|_q$$

denote the error of the best s-sparse approximation of x.

In most cases, the vector x is not sparse in the standard basis, but there is a basis Ψ, such that $x = \Psi z$ and z is sparse. This is also known as *synthesis sparsity* of x. To find an (approximately) synthesis sparse vector, we can instead solve the problem of recovering z from $y = A\Psi z$. A common strategy in CS is to solve a basis pursuit program in order to recover the original vector. For a fixed noise level ε, it is given by

$$\text{minimize } \|z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon. \tag{4.1}$$

While this and related approaches of convex regularization have been studied in the inverse problems and statistics literature long before the field of compressed sensing developed, these works typically assumed the measurement setup was given. The new paradigm arising in the context of compressed sensing was to attempt to use the remaining degrees of freedom of the measurement system to reduce the ill-posedness of the system as much as possible. In many measurement systems, the most powerful known strategies will be based on randomization, i.e., the free parameters are chosen at random.

Given an appropriate amount of randomness (i.e., for various classes of random matrices A, including some with structure imposed by underlying applications), one can show that the

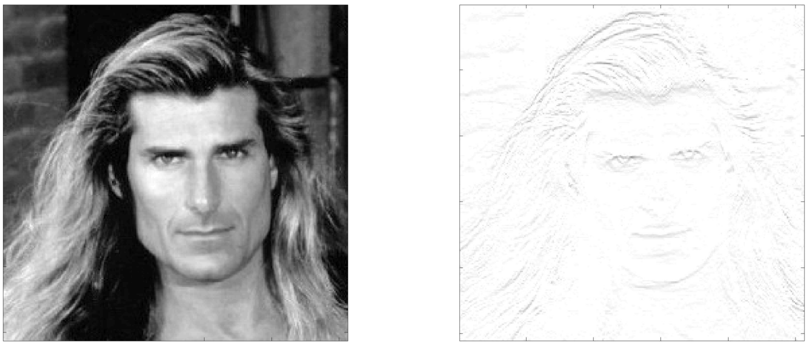

Figure 4.1: The original Fabio image (left) and the absolute values after application of a discrete gradient operator(right).

minimizer $\hat{x}$ of (4.1) recovers the original vector x with error

$$\|x-\hat{x}\|_2 \leq c\left(\frac{\sigma_s(x)_1}{\sqrt{s}}+\varepsilon\right), \tag{4.2}$$

see, e.g., [BDDW08] for an elementary proof in the case of subgaussian matrices without structure, and [KR14] for an overview, including many references, of corresponding results for random measurement systems with additional structure imposed by applications. Note that (4.2) entails that if x is s-sparse and the measurements are noiseless, the recovery is exact.

For many applications, however, the signal model of sparsity in an orthonormal basis has proven somewhat restrictive. Two main lines of generalization have been proposed. The first line of work, initiated by [RSV08] is the study of sparsity in redundant representation systems, at first under incoherence assumptions on the dictionary. More recently, also systems without such assumptions have been analyzed [CENR10, KNW15]. The main idea of these works is that even when one cannot recover the coefficients correctly due to conditioning problems, one may still hope for a good approximation of the signal.

The second line of work focuses on signals that are sparse after the application of some transform, one speaks of *cosparsity* or *analysis sparsity* [NDEG13], see, e.g., [KR15] for an analysis of the Gaussian measurement setup in this framework. A special case of particular importance, especially for imaging applications, is that of sparse gradients. Namely, as it turns out, natural images often admit very sparse approximations in the gradient domain, see, e.g., Figure 4.1. Here the discrete gradient at location $i=(i_1,\ldots,i_n)$ is defined as the vector with its n entries given by $\big((\nabla z)_i\big)_j = z_{i+e_j}-z_i$, $j=1,\ldots,n$, where e_j is the j-th standard basis vector.

A first attempt to recover a gradient sparse signal is to formulate a compressed sensing problem in terms of the sparse gradient. When this is possible (for instance in the example of Fourier measurements [CRT06a]), applying (4.1) will correspond to minimizing $\|\nabla z\|_1 =: \|z\|_{TV}$, the *total variation seminorm*. Then (under some additional assumptions) compressed sensing recovery guarantees of the form (4.2) can apply. This proof strategy, however, only allows for showing that the gradient can be approximately recovered, not the signal. When no noise is present and the gradient is exactly sparse (which is not very realistic), this allows for signal recovery via integrating the gradient, but in case of noisy measurements, this procedure is highly unstable.

Nevertheless, the success motivates to minimize the total variation seminorm if one attempts to recover the signal directly, not the gradient. In analogy with (4.1), this yields the following minimization problem.

$$\text{minimize } \|z\|_{TV} = \|\nabla z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon.$$

For A the identity (i.e., not reducing the dimension), this relates to the famous Rudin-Osher-Fatemi functional, a classical approach for signal and image denoising [ROF92]. Due to its high relevance for image processing, this special case of analysis sparsity has received a lot of attention recently also in the compressed sensing framework where A is dimension reducing. The purpose of this chapter is to give an overview of recovery results for total variation minimization in this context of compressed sensing (Section 4.2) and to provide some geometric intuition by discussing the one-dimensional case under Gaussian or subgaussian measurements (to our knowledge, a generalization to the latter case does not appear yet in the literature) with a focus on the interaction between the high-dimensional geometry and spectral properties of the gradient operator (Section 4.3).

4.2 An overview over TV recovery results

In this section, we will give an overview of the state of the art guarantees for the recovery of gradient sparse signals via total variation minimization. We start by discussing in Section 4.2.1 sufficient conditions for the success of TV minimization.

Subsequently, we focus on recovery results for random measurements. Interestingly, the results in one dimension differ severely from the ones in higher dimensions. Instead of obtaining a required number of measurements roughly on the order of the sparsity level s, we need $\sqrt{sn}$ measurements for recovery. We will see this already in Subsection 4.2.2, where we present the results of Cai and Xu [CX15] for recovery from Gaussian measurements. In Section 4.3, we will use their results to obtain refined results for noisy measurements as well as guarantees for subgaussian measurements, combined with an argument of Tropp [Tro15]. In Subsection 4.2.3 we will present results by Ward and Needell for dimensions larger or equal than two showing that recovery can be achieved from Haar incoherent measurements.

4.2.1 Sufficient Recovery Conditions

Given linear measurements $Ax = y$ for an arbitrary $A \in \mathbb{R}^{m \times n}$ and a signal x with $\|\nabla x\|_0 \leq s$, a natural way to recover x is by solving

$$\text{minimize } \|\nabla z\|_1 \text{ such that } Az = y. \tag{4.3}$$

For $I \subset [n]$ we denote A_I as the columns of A indexed by I, and for a consecutive notation we denote $\mathcal{I}_I^T \nabla$ as the rows of ∇ indexed by I and $\mathcal{I}$ as the identity matrix. The following results can also be easily applied to *analysis ℓ_1-minimization*, where any arbitrary matrix $D \in \mathbb{R}^{p \times n}$ replaces ∇ in (4.3), as well as to any real Hilbert space setting [Kru15].

In many applications it is important to verify whether there is exactly one solution of (4.3). Since ∇ is not injective here, we cannot easily use the well-known recovery results in compressed sensing [FR13] for the matrix $A\nabla^\dagger$. However, a necessary conditon can be given since x can only satisfy $Ax = y$ and $(\nabla x)_{I^c} = 0$ if

$$\ker(\mathcal{I}_{I^c}^T \nabla) \cap \ker(A) = \{0\}.$$

If ∇ is replaced by the identity, this is equivalent to A_I being injective. Since this injectivity condition is unavoidable, we assume for the rest of this section that it is satisfied.

The paper [NDEG13] provides sufficient and necessary conditons for uniform recovery via (4.3). The conditions rely on the null space of the measurements and are hard to verify similar to the classical compressed sensing setup [TP14]. The following result is a corollary of these conditions. It no longer provides a necessary condition, but is more manageable.

Corollary 4.2.1. *[NDEG13] For all $x \in \mathbb{R}^n$ with $s := \|\nabla x\|_0$, the solution of* (4.3) *with $y = Ax$ is unique and equal to x if for all $I \subset [n]$ with $|I| \leq s$ it holds that*

$$\forall w \in \ker(A) \backslash \{0\} \colon \ \|(\nabla w)_I\|_1 < \|(\nabla w)_{I^c}\|_1.$$

To consider measurements for specific applications, where it is difficult to prove whether uniform recovery is guaranteed, one can empirically examine whether specific elements x solve (4.3) uniquely. For computed tomography measurements, a *Monte Carlo Experiment* is considered in [JKL15] to approximate the fraction of all gradient s-sparse vectors to uniquely solve

(4.3). The results prompt that there is a sharp transition between the case that every vector with a certain gradient sparsity is uniquely recoverable and the case that TV-minimization will find a different solution than the desired vector. This behavior empirically agrees with the phase transition in the classical compressed sensing setup with Gaussian measurements [Don04].

To efficiently check whether many specific vectors x can be uniquely recovered via (4.3), one needs to establish characteristics of x which must be easily verifiable. Such a non-uniform recovery condition is given in the following theorem.

Theorem 4.2.2. *[JKL15] It holds that $x \in \mathbb{R}^n$ is a unique solution of* (4.3) *if and only if there exists $w \in \mathbb{R}^m$ and $v \in \mathbb{R}^{n-1}$ such that*

$$\nabla^T v = A^T w, v_I = sign(\nabla x)_I, \|v_{I^c}\|_\infty < 1. \tag{4.4}$$

The basic idea of the proof is to use the optimality condition for convex optimization problems [Roc72]. Equivalent formulations of the latter theorem can be found in [ZMY16, KR15] where the problem is considered from a geometric perspective. However, verifying the conditions in Theorem 4.2.2 still requires solving a linear program where an optimal v for (4.4) needs to be found. In classical compressed sensing, the *Fuchs Condition* [Fuc04] is known as a weaker result as it suggests a particular w in (4.4) and avoids solving the consequential linear program. The following result generalizes this result to general analysis ℓ_1-minimization.

Corollary 4.2.3. *If $x \in \mathbb{R}^n$ satisfies*

$$\|(\mathcal{I}_{I^c}^T \nabla(\nabla^T \mathcal{I}_{I^c} \mathcal{I}_{I^c}^T \nabla + A^T A)^{-1} \nabla sign(\nabla x))_I\|_\infty < 1$$

then x is the unique solution of (4.3).

4.2.2 Recovery from Gaussian measurements

As discussed above, to date no deterministic constructions of compressed sensing matrices are known that get anywhere near an optimal number of measurements. Also for the variation of aiming to recover approximately gradient sparse measurements, the only near-optimal recovery guarantees have been established for random measurement models. Both under (approximate) sparsity and gradient sparsity assumptions, an important benchmark is that of a measurement matrix with independent standard Gaussian entries. Even though such measurements are hard to realize in practice, they can be interpreted as the scenario with maximal randomness, which often has particularly good recovery properties. For this reason, the recovery properties of total variation minimization have been analyzed in detail for such measurements. Interestingly, as shown by the following theorem, recovery properties in the one-dimensional case are significantly worse than for synthesis sparse signals and also for higher dimensional cases. That is why we focus on this case in Section 4.3, providing a geometric viewpoint and generalizing the results to subgaussian measurements.

Theorem 4.2.4. *[CX15] Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. standard Gaussian random variables and let $\hat{x}$ be a solution of* (4.3) *with input data $y = Ax_0$. Then*

1. *There exist constants $c_1, c_2, c_3, c_4 > 0$, such that for $m \geq c_1\sqrt{sn}(\log n + c_2)$*

$$\mathbb{P}(\forall x_0 \colon \|\nabla x_0\|_0 \leq s \colon \hat{x} = x_0) \geq 1 - c_3 \mathrm{e}^{-c_4\sqrt{m}}.$$

2. *For any $\eta \in (0,1)$, there are constants $\tilde{c}_1, \tilde{c}_2 > 0$ and a universal constant $c_2 > 0$, such that for $s \geq \tilde{c}_0$ and $(s+1) < \frac{n}{4}$. If $m \leq \tilde{c}_1\sqrt{sn} - \tilde{c}_2$, there exist infinitely many $x_0 \in \mathbb{R}^n$ with $\|\nabla x_0\|_0 \leq s$, such that $\mathbb{P}(\hat{x} \neq x_0) \geq 1 - \eta$.*

This scaling is notably different from what is typically obtained for synthesis sparsity, where the number of measurements scales linearly with s up to $\log$ factors. Such a scaling is only obtained for higher dimensional signals, e.g., images. Indeed, in [CX15], it is shown that for dimensions at least two the number of Gaussian measurements sufficient for recovery is

$$m \geq \begin{cases} c_2 s \log^3 n, & \text{if } d = 2 \\ c_d s \log n, & \text{if } d \geq 3, \end{cases}$$

where the constant c_d depends on the dimension.

Furthermore, as we can see in Theorem 4.2.7 below, this is also the scaling one obtains for dimensions larger than 1 and Haar incoherent measurements. Thus the scaling of $\sqrt{sn}$ is a unique feature of the 1-dimensional case. Also note that the square-root factor in the upper bound makes the result meaningless for a sparsity level on the order of the dimension. This has been addressed in [KRZ15], showing that a dimension reduction is also possible if the sparsity level is a (small) constant multiple of the dimension.

The proof of Theorem 4.2.4 uses Gordon's escape through the mesh Theorem [Gor88]. We will elaborate on this topic in Section 4.3.

In case we are given noisy measurements $y = Ax_0 + e$ with $\|e\|_2 \leq \varepsilon$, we can instead of solving (4.3) consider

$$\text{minimize } \|\nabla z\|_1 \text{ such that } \|Az - y\|_2 \leq \varepsilon. \tag{4.5}$$

If ∇x_0 is not exactly, but approximately sparse, and our measurements are corrupted with noise, the following result can be established.

Theorem 4.2.5. *[CX15] Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. standard Gaussian random variables and let $\hat{x}$ be a solution of* (4.5) *with input data y satisfying $\|Ax_0 - y\|_2 \le \varepsilon$. Then for any $\alpha \in (0,1)$, there are positive constants $\delta, c_0, c_1, c_2, c_3$, such that for $m = \alpha n$ and $s = \delta n$*

$$\mathbb{P}\left(\|x_0 - \hat{x}\|_2 \le c_2 \frac{\min_{|S|\le s} \|(\nabla x_0)_{S^c}\|_1}{\sqrt{n}} + c_3 \frac{\varepsilon}{\sqrt{n}}\right) \ge 1 - c_0 \mathrm{e}^{-c_1 n}.$$

This looks remarkably similar to the recovery guarantees obtained for compressed sensing, note however that the number of measurements needs to be proportional to n, which is not desirable. We will present a similar result with improved number of measurements in Section 4.3.5.

Theorem 4.2.6. *(Corollary of Theorem 4.3.8) Let $x_0 \in \mathbb{R}^n$ be such that $\|\nabla x_0\| \le s$ for $s > 0$ and $A \in \mathbb{R}^{m\times n}$ with $m \ge C\sqrt{ns}\log(2n)$ be a standard Gaussian matrix. Furthermore, set $y = Ax_0 + e$, where $\|e\| \le \varepsilon$ denotes the (bounded) error of the measurement and for some absolute constants $c, \tilde{c} > 0$ the solution $\hat{x}$ of* (4.12) *satisfies*

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{c\sqrt[4]{ns}(\sqrt{\log(2n)} - 1)}\right) \le \mathrm{e}^{-\tilde{c}\sqrt{ns}}.$$

Note, however that in contrast to theorem 4.2.5, this theorem does not cover the case of gradient compressible vectors, but on the other hand Theorem 4.3.8 also incorporates the case of special subgaussian measurement ensembles. Also, if we set $s = \delta n$, we reach a similar conclusion as in Theorem 4.2.5.

4.2.3 Recovery from Haar-incoherent measurements

For dimensions $d \ge 2$, Needell and Ward [NW13a, NW13b] derived recovery results for measurement matrices having the restricted isometry property (RIP) when composed with the Haar wavelet transform. Here we say that a matrix Φ has the RIP of order k and level δ if for every k-sparse vector x it holds that

$$(1-\delta)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1-\delta)\|x\|_2^2.$$

The results of [NW13a, NW13b] build upon a connection between a signal's wavelet representation and its total variation seminorm first noted by Cohen, Dahmen, Daubechies and DeVore [CDDD03].

Their theorems yield stable recovery via TV minimization for N^d dimensional signals. For $d = 2$, notably these recovery results concern images of size $N \times N$.

Several definitions are necessary in order to be able to state the theorem. The d dimensional discrete gradient is defined via $\nabla \colon \mathbb{R}^{C^d} \to \mathbb{C}^{N^d \times d}$ and maps $x \in \mathbb{C}^{N^d}$ to its discrete derivative which, for each $\alpha \in [N]^d$ is a vector $(\nabla x)_\alpha \in \mathbb{C}^d$ composed of the derivatives in all d directions. Up to now, we have always used the anisotropic version of the TV seminorm, which can be seen as taking the ℓ_1 norm of the discrete gradient. The isotropic TV seminorm is defined via a combination of ℓ_2 and ℓ_1 norms. It is given by $\|z\|_{TV_2} := \sum_{\alpha \in [N]^d} \|(\nabla z)_\alpha\|_2$. The result in [NW13a] is given in terms of the isotropic TV seminorm but can also be formulated for the anisotropic version.

Furthermore, we will need to concatenate several measurement matrices in order to be able to state the theorem. This will be done via the concatenation operator $\oplus \colon \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_1}) \times \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_2}) \to \mathrm{Lin}(\mathbb{C}^n, \mathbb{C}^{k_1+k_2})$, which 'stacks' two linear maps.

Finally, we need the notion of shifted operators. For an operator $\mathcal{B} \colon \mathbb{C}^{N^{l-1} \times (N-1) \times N^{d-l}} \to \mathbb{C}^q$, these are defined as the operators $\mathcal{B}_{0_l} \colon \mathbb{C}^{N^d} \to \mathbb{C}^q$ and $\mathcal{B}^{0_l} \colon \mathbb{C}^{N^d} \to \mathbb{C}^q$ concatenating a column of zeros to the end or beginning of the l-th component, respectively.

Theorem 4.2.7 ([NW13a])**.** *Let $N = 2^n$ and fix integers p and q. Let $\mathcal{A} \colon \mathbb{C}^{N^d} \to \mathbb{C}^p$ be a map that has the restricted isometry property of order $2ds$ and level $\delta < 1$ if it is composed with the orthonormal Haar wavelet transform. Furthermore let $\mathcal{B}_1, \dots, \mathcal{B}_d$ with $\mathcal{B}_j \colon \mathbb{C}^{(N-1)N^{d-1}} \to \mathbb{C}^q$ be such that $\mathcal{B} = \mathcal{B}_1 \oplus \mathcal{B}_2 \oplus \cdots \oplus \mathcal{B}_d$ has the restricted isometry property of order $5ds$ and level $\delta < \frac{1}{3}$. Consider the linear operator $\mathcal{M} = \mathcal{A} \oplus [\mathcal{B}_1]_{0_1} \oplus [\mathcal{B}_1]^{0_1} \oplus \cdots \oplus [\mathcal{B}_d]_{0_d} \oplus [\mathcal{B}_d]^{0_d}$. Then $\mathcal{M} \colon \mathbb{C}^{N^d} \to \mathbb{C}^m$ with $m = 2dq + p$ and for all $x \in \mathbb{C}^{N^d}$ we have the following. Suppose we have noisy measurements $y = \mathcal{M}(x) + e$ with $\|e\|_2 \le \varepsilon$, then the solution to*

$$\hat{x} = \mathop{arg\,min}_{z} \|z\|_{TV_2} \quad \textit{such that } \|\mathcal{M}(z) - y\|_2 \le \varepsilon$$

satisfies

1. $\|\nabla(x - \hat{x})\|_2 \le c_1 \left(\frac{\|\nabla x - (\nabla x)_S\|_{1,2}}{\sqrt{s}} + \sqrt{d}\varepsilon \right)$,
2. $\|x - \hat{x}\|_{TV_2} \le c_2 \left(\|\nabla x - (\nabla x)_S\|_{1,2} + \sqrt{sd}\varepsilon \right)$,
3. $\|x - \hat{x}\|_2 \le c_3 d \log N \left(\frac{\|\nabla x - (\nabla x)_S\|_{1,2}}{\sqrt{s}} + \sqrt{d}\varepsilon \right)$,

for some absolute constants c_1, c_2, c_3.

From the last point of the previous theorem, we see that for noiseless measurements and gradient sparse vectors x, perfect recovery can be achieved provided the RIP assumption holds.

Subgaussian measurement matrices, for example, will have the RIP, also when composed with the Haar wavelet transform H (this is a direct consequence of rotation invariance). Moreover, as shown in [KW11], randomizing the column signs of an RIP matrix will, with high probability, also yield a matrix that has the RIP when composed with H. An important example is a subsampled Fourier matrix with random column signs, which relates to spread spectrum MRI (cf. [PMG$^+$12]).

4.2.4 Recovery from subsampled Fourier measurements

Fourier measurements are widely used in many applications. Especially in medical applications as parallel-beam tomography and magnetic resonance imaging it is desirable to reduce the number of samples to spare patients burden. In Section 4.2.1, this is a motivation for introducing algorithmic checks for unique solutions of (4.3). In this section, we consider a probabilistic approach where an incomplete measurement matrix $A \in \mathbb{C}^{m \times n}$ chosen from the discrete Fourier transform on $\mathbb{C}^N$ is considered. Therefore we consider a subset Ω of the index set $\{-\lfloor n/2 \rfloor + 1, ..., \lceil n/2 \rceil\}$, where Ω consists of m integers chosen uniformly at random and, additionally, $0 \in \Omega$. Hence, we want to recover a signal, sparse in the gradient domain, with a measurement matrix $A = (e^{2\pi ikj/n})_{k \in \Omega, j \in [n]}$. In [CRT06a] the optimal sampling cardinality for s-sparse signals in the gradient domain was given and enables to recover one-dimensional signals signals from $\mathcal{O}(k \log(n))$ Fourier samples. It naturally extends to two dimensions.

Theorem 4.2.8. *[CRT06a] With probability exceeding* $1 - \eta$*, a signal* z*, which is* k*-sparse in the gradient domain is the unique solution of* (4.3) *if*

$$m \gtrsim k(\log(n) + \log(\eta^{-1})).$$

As already discussed in the introduction, the proof of this result proceeds via recovering the gradient and then using that the discrete gradient (with periodic boundary conditions) is injective. Due to the poor conditioning of the gradient, however, this injectivity results do not directly generalize to recovery guarantees for noisy measurements. For two (and more) dimensions, such results can be obtained via the techniques discussed in the previous subsection.

These techniques, however, do not apply directly. Namely, the Fourier (measurement) basis is not incoherent to the Haar wavelet basis; in fact, the constant vector is contained in both, which makes them maximally coherent. As observed in [PVW11], this incoherence phenomenon only occurs for low frequencies, the high frequency Fourier basis vectors exhibit small inner products to the Haar wavelet basis. This can be taken into account using a *variable density* sampling scheme with sampling density that is larger for low frequencies and smaller for high frequencies. For such a sampling density, one can establish the restricted isometry for the corresponding randomly subsampled discrete Fourier matrix combined with the Haar wavelet transform with appropriately rescaled rows [KW14b]. This yields the following recovery guarantee.

Theorem 4.2.9. *[KW14b] Fix integers $N = 2^p, m,$ and s such that $s \gtrsim \log(N)$ and*

$$m \gtrsim s \log^3(s) \log^5(N). \tag{4.6}$$

Select m frequencies $\{(\omega_1^j, \omega_2^j)\}_{j=1}^m \subset \{-N/2+1, \dots, N/2\}^2$ i.i.d. according to

$$\mathbb{P}\big[(\omega_1^j, \omega_2^j) = (k_1, k_2)\big] = C_N \min\left(C, \frac{1}{k_1^2 + k_2^2}\right) =: \eta(k_1, k_2), \quad -N/2+1 \leq k_1, k_2 \leq N/2, \tag{4.7}$$

where C is an absolute constant and C_N is chosen such that η is a probability distribution. Consider the weight vector $\rho = (\rho_j)_{j=1}^m$ with $\rho_j = (1/\eta(\omega_1^j, \omega_2^j))^{1/2}$, and assume that the noise vector $\xi = (\xi_j)_{j=1}^m$ satisfies $\|\rho \circ \xi\|_2 \leq \varepsilon\sqrt{m}$, for some $\epsilon > 0$. Then with probability exceeding $1 - N^{-C\log^3(s)}$, the following holds for all images $f \in \mathbb{C}^{N \times N}$:
Given noisy partial Fourier measurements $y = \mathcal{F}_\Omega f + \xi$, the estimation

$$f^{\#} = \underset{g \in \mathbb{C}^{N\times N}}{\mathit{arg\,min}} \|g\|_{TV} \quad \textit{such that} \quad \|\rho \circ (\mathcal{F}_\Omega g - y)\|_2 \leq \varepsilon\sqrt{m}, \tag{4.8}$$

where $\circ$ denotes the Hadamard product, approximates f up to the noise level and best s-term approximation error of its gradient:

$$\|f - f^{\#}\|_2 \lesssim \frac{\|\nabla f - (\nabla f)_s\|_1}{\sqrt{s}} + \varepsilon. \tag{4.9}$$

A similar optimality result is given in [Poo15], also for noisy data and inexact sparsity. In contrast to the previous result, this result includes the one-dimensional case. The key to obtaining such a result is showing that the stable gradient recover implies the stable signal recovery, i.e.,

$$\|z\|_2 \lesssim \gamma + \|z\|_{TV} \text{ with } \|Az\|_2 \leq \gamma. \tag{4.10}$$

Again the sampling distribution is chosen as a combination of the uniform distribution and a decaying distribution. The main idea is to use this sampling to establish (4.10) via the RIP. We skip technicalities for achieving the optimality in the following theorem and refer to the original article for more details.

Theorem 4.2.10. *[Poo15] Let $z \in \mathbb{C}^n$ be fixed and x be a minimizer of* (4.5) *with $\varepsilon = \sqrt{m}\delta$ for some $\delta > 0$, $m \gtrsim k\log(n)(1 + \log(\eta^{-1}))$, and an appropriate sampling distribution. Then with* *probability exceeding* $1-\eta$*, it holds that*

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C_1\frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \frac{\|z-x\|_2}{\sqrt{n}} \lesssim C_2\left(\frac{\delta}{\sqrt{s}} + C_1\frac{\|P\nabla z\|_1}{k}\right),$$

where P *is the orthogonal projection onto a* k*-dimensional subspace,*

$$C_1 = \log(k)\log^{1/2}(m), \textit{ and } C_2 = \log^2(k)\log(n)\log(m).$$

In the two-dimensional setting the result changes to

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C_3\frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \|z-x\|_2 \lesssim C_2\left(\delta + C_3\frac{\|P\nabla z\|_1}{k}\right),$$

with remaining C_2 and

$$C_3 = \log(k)\log(n^2/k)\log^{1/2}(n)\log^{1/2}(m).$$

These results are optimal since the best error one can achieve [NW13b] is $\|z-x\|_2 \lesssim \frac{\|P\nabla z\|_1}{\sqrt{k}}$.

The optimality in the latter theorems is achieved by considering a combination of uniform random samling and variable density sampling. Uniform sampling on its own can achieve robust and stable recovery. However, the following theorem shows that the signal error is no longer optimal but the bound on the gradient error is still optimal up to log factors. Here (4.10) is obtained by using the Poincaré inequality.

Theorem 4.2.11. *[Poo15] Let* $z \in \mathbb{C}^n$ *be fix and* x *be a minimizer of* (4.5) *with* $\varepsilon = \sqrt{m}\delta$ *for some* $\delta > 0$ *and* $m \gtrsim k\log(n)(1+\log(\eta^{-1}))$ *with random uniform sampling. Then with probability exceeding* $1-\eta$*, it holds that*

$$\|\nabla z - \nabla x\|_2 \lesssim \left(\delta\sqrt{k} + C\frac{\|P\nabla z\|_1}{\sqrt{k}}\right), \frac{\|z-x\|_2}{\sqrt{n}} \lesssim (\delta\sqrt{s} + C\|P\nabla z\|_1),$$

where P *is the orthogonal projection onto a* k*-dimensional subspace and* $C = \log(k)\log^{1/2}(m)$*.*

4.3 TV-recovery from subgaussian measurements in 1D

In this section, we will apply the geometric viewpoint discussed in [Ver15] to the problem, which will eventually allow us to show the TV recovery results for noisy subgaussian measurements mentioned in Section 4.2.2.

As in the original proof of the 1D recovery guarantees for Gaussian measurements [CX15], the *Gaussian mean width* will play an important role in our considerations.

Definition 4.3.1. *The (Gaussian) mean width of a bounded subset K of $\mathbb{R}^n$ is defined as*

$$w(K) := \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle,$$

where $g \in \mathbb{R}^n$ is a vector of i.i.d. $\mathcal{N}(0,1)$ random variables.

In [CX15], the mean width appears in the context of the *Gordon's escape through the mesh* approach [Gor88] (see Section 4.3.4 below), but as we will see, it will also be a crucial ingredient in applying the Mendelson small ball method [KM15, Men14].

The mean width has some nice (and important) properties, it is for example invariant under taking the convex hull, i.e.,

$$w(\mathrm{ch}(K)) = w(K).$$

Furthermore, it is also invariant under translations of K, as $(K - x_0) - (K - x_0) = K - K$. Due to the rotational invariance of Gaussian random variables, that is $Ug \sim g$, we also have that $w(UK) = w(K)$. Also, it satisfies the inequalities

$$w(K) = \mathbb{E} \sup_{x \in K-K} \langle g, x \rangle \leq 2\mathbb{E} \sup_{x \in K} \langle g, x \rangle \leq 2\mathbb{E} \sup_{x \in K} |\langle g, x \rangle|,$$

which are equalities if K is symmetric about 0, because then $K = -K$ and hence $K - K = 2K$.

4.3.1 M^* bounds and recovery

In order to highlight the importance of the Gaussian mean width in signal recovery, we present some arguments from [Ver15]. Thus in this section we present a classical result, the M^* bound, which connects the mean width to recovery problems, cf. [Ver15]. Namely, recall that due to rotational invariance, the kernel of a Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ is a random subspace distributed according to the uniform distribution (the Haar measure) on the Grassmannian

$$G_{n,n-m} := \{V \leq \mathbb{R}^n \colon \dim(V) = n - m\}.$$

Consequently, the set of all vectors that yield the same measurements directly correspond to such a random subspace.

The average size of the intersection of this subspace with a set reflecting the minimization objective now gives us an average bound on the worst case error.

Theorem 4.3.1 (M^* bound, Theorem 3.12 in [Ver15])**.** *Let K be a bounded subset of $\mathbb{R}^n$ and E be a random subspace of $\mathbb{R}^n$ of drawn from the Grassmanian $G_{n,n-m}$ according to the Haar* *measure. Then*

$$\mathbb{E}\,\mathrm{diam}(K \cap E) \leq C\frac{w(K)}{\sqrt{m}}, \tag{4.11}$$

where C is absolute constant.

Given the M^*-bound it is now straightforward to derive bounds on reconstructions from linear observations. We first look at feasibility programs - which in turn can be used to obtain recovery results for optimization problems. For that, let $K \subset \mathbb{R}^n$ be bounded and $x \in K$ be the vector we seek to reconstruct from measurements $Ax = y$ with a Gaussian matrix $A \in \mathbb{R}^{m\times n}$.

Corollary 4.3.2. *[MPTJ07] Choose $\hat{x} \in \mathbb{R}^n$, such that*

$$\hat{x} \in K \ \textit{and}\ A\hat{x} = y,$$

then one has, for an absolute constant C',

$$\mathbb{E} \sup_{x\in K} \|\hat{x} - x\|_2 \leq C' \frac{w(K)}{\sqrt{m}}.$$

This corollary directly follows by choosing $C' = 2C$, observing that $\hat{x} - x \in K - K$, and that the side constraint enforces $A(\hat{x} - x) = 0$.

Via a standard construction in functional analysis, the so called *Minkowski functional*, one can now cast an optimization problem as a feasiblity program so that Corollary 4.3.2 applies.

Definition 4.3.2. *The Minkowski functional of a bounded, symmetric set $K \subset \mathbb{R}^n$ is given by*

$$\|\cdot\|_K \colon \mathbb{R}^n \to \mathbb{R}\colon x \mapsto \inf\{t > 0\colon x \in tK\}.$$

So the Minkowski functional tells us, how much we have to ’inflate’ our given set K in order to capture the vector x. Clearly, from the definition we have that if K is closed

$$K = \{x\colon \|x\|_K \leq 1\}.$$

If a convex set K is closed and symmetric, then $\|\cdot\|_K$ defines a norm on $\mathbb{R}^n$.

Recall that a set K is star shaped, if there exists a point $x_0 \in K$, which satisfies that for all $x \in K$ we have $\{tx_0 + (1 - t)x\colon t \in [0, 1]\} \subset K$. It is easy to see that convex sets are star shaped, but for example unions of subspaces are not convex, but star shaped.

For bounded, star shaped K, the notion of $\|\cdot\|_K$ now allows to establish a direct correspondence between norm minimization problems and feasibility problems. With this observation, Corollary 4.3.2 translates to the following result.

Corollary 4.3.3. *For K bounded, symmetric and star-shaped, let $x \in K$ and $y = Ax$. Choose $\hat{x} \in \mathbb{R}^n$, such that it solves*

$$\min \|z\|_K \text{ with } Az = y,$$

then

$$\mathbb{E} \sup_{x \in K} \|\hat{x} - x\|_2 \leq C' \frac{w(K)}{\sqrt{m}}.$$

Here $\hat{x} \in K$ is due to the fact that the minimum satisfies $\|\hat{x}\|_K \leq \|x\|_K \leq 1$, as $x \in K$ by assumption.

This result directly relates recovery guarantees to the mean width, it thus remains to calculate the mean width for the sets under consideration. In the following subsections, we will discuss two cases. The first one directly corresponds to the desired signal model, namely gradient sparse vectors. These considerations are mainly of theoretical interest, as the associated minimization problem closely relates to support size minimization, which is known to be NP hard in general. The second case considers the TV minimization problem introduced above, which then also yields guarantees for the (larger) set of vectors with bounded total variation.

Note, however, that the M^*-bound only gives a bound for the expected error. We can relate this result to a statement about tail probabilities using Markov's inequality, namely

$$\mathbb{P}(\sup_{x \in K} \|x - \hat{x}\|_2 > t) \leq t^{-1} \mathbb{E} \sup_{x \in K} \|x - \hat{x}\|_2 \leq C' \frac{w(K)}{t\sqrt{m}}.$$

In the next section we compute the mean width for the set of gradient sparse vectors, that is we now specify the set K in Corollary 4.3.2 to be the set of all vectors with energy bounded by one that only have a small number of jumps.

4.3.2 The mean width of gradient sparse vectors in 1d

Here [PV13] served as an inspiration, as the computation is very similar for the set of sparse vectors.

Definition 4.3.3. *The jump support of a vector x is given via*

$$\operatorname{Jsupp}(x) := \{i \in [n-1] \colon x_{i+1} - x_i \neq 0\}.$$

The jump support captures the positions, in which a vector x changes its values. With this, we now define the set

$$K_0^s := \{x \in \mathbb{R}^n \colon \|x\|_2 \leq 1, |\operatorname{Jsupp}(x)| \leq s\}.$$

The set K_0^s consists of all s-gradient sparse vectors, which have 2-norm smaller than one. We will now calculate the mean width of K_0^s in order to apply Corrolary 4.3.2 or 4.3.3.

Note that we can decompose the set K_0^s into smaller sets $K_J \cap B_2^n$ with $K_J = \{x\colon \operatorname{Jsupp}(x) \subset J\}$, $|J| = s$ and $B_2^n = \{x \in \mathbb{R}^n\colon \|x\|_2 \leq 1\}$. As we can't add any jumps within the set K_J, it is a subspace of $\mathbb{R}^n$. We can even quite easily find an orthonormal basis for it, if we define

$$(e_{[i,j]})_k := \frac{1}{\sqrt{j-i+1}} \begin{cases} 1, & \text{if } k \in [i,j] \\ 0, & \text{else} \end{cases} .$$

As we can align all elements of $J = \{j_1, j_2, \ldots, j_s\}$ with $1 \leq j_1 < j_2 < \ldots < j_s = n$, we see that $\{e_{[1,j_1]}, e_{[j_1+1,j_2]}, e_{[j_2+1,j_3]}, \ldots, e_{[j_{s-1}+1,j_s]}\}$ forms an ONB of K_J. Now, we can write all elements $x \in K_J \cap B_2^n$ as $x = \sum_{i=1}^s \alpha_i e_{[j_{i-1}+1,j_i]}$ by setting $j_0 := 0$. The property that $x \in B_2^n$ now enforces (ONB) that $\|\alpha\|_2 \leq 1$. Now, note that $K_0^s = -K_0^s$, so we have

$$w(K_0^s) = \mathbb{E} \sup_{x \in K_0^s - K_0^s} \langle g, x\rangle = 2\mathbb{E} \sup_{x \in K_0^s} \langle g, x\rangle.$$

Using the decomposition $K_0^s = \bigcup_{|J|=s} (K_J \cap B_2^n)$, we get

$$w(K_0^s) = 2\mathbb{E} \sup_{|J|=s} \sup_{x \in K_J \cap B_2^n} \langle g, x\rangle.$$

Now

$$\sup_{x \in K_J \cap B_2^n} \langle g, x\rangle \leq \sup_{\alpha \in B_2^s} \sum_{i=1}^s \alpha_i \langle g, e_{[j_{i-1}+1,j_i]}\rangle = \sup_{\alpha \in B_2^s} \sum_{i=1}^s \alpha_i \underbrace{\sum_{k=j_{i-1}+1}^{j_i} \frac{g_k}{\sqrt{j_i - j_{i-1}}}}_{=:G_i^J} .$$

Note that G_i^J is again a Gaussian random variable with mean 0 and variance 1. Furthermore, the supremum over α is attained, if α is parallel to G^J, so we have $\sup_{x \in K_J \cap B_2^n} \langle g, x\rangle = \|G^J\|_2$. Also note that G^J has i.i.d. entries, but for different J_1, J_2, the random vectors G^{J_1} and G^{J_2} may be dependent. Our task is now to calculate $\mathbb{E} \sup_{|J|=s} \|G^J\|_2$. As it has been shown for example in [FR13], we have that

$$\sqrt{\frac{2}{\pi}}\sqrt{s} \leq \mathbb{E}\|G^J\|_2 \leq \sqrt{s}$$

and from standard results for Gaussian concentration (cf. [PV13]), we get

$$\mathbb{P}(\|G^J\|_2 \geq \sqrt{s} + t) \leq \mathbb{P}(\|G^J\|_2 \geq \mathbb{E}\|G_J\|_2 + t) \leq \mathrm{e}^{-t^2/2}.$$

By noting that $|\{J \subset [n] \colon |J| = s\}| = \binom{n}{s}$, we see by a union bound that

$$\mathbb{P}(\sup_{|J|=s} \|G^J\|_2 \geq \sqrt{s} + t) \leq \binom{n}{s} \mathbb{P}(\|G^J\|_2 \geq \sqrt{s} + t) \leq \binom{n}{s} \mathrm{e}^{-t^2/2}.$$

For the following calculation, set $X := \sup_{|J|=s} \|G^J\|_2$. By Jensen's inequality and rewriting the expectation, we have that

$$\mathrm{e}^{\lambda \mathbb{E} X} \leq \mathbb{E}\mathrm{e}^{\lambda X} = \int_0^\infty \mathbb{P}(\mathrm{e}^{\lambda X} \geq \tau) \mathrm{d}\tau.$$

Now, the previous consideration showed, that

$$\mathbb{P}(\mathrm{e}^{\lambda X} \geq \underbrace{\mathrm{e}^{\lambda(\sqrt{s}+t)}}_{=:\tau}) = \mathbb{P}(X \geq \sqrt{s} + t) \leq \binom{n}{s} \mathrm{e}^{-t^2/2} = \binom{n}{s} \mathrm{e}^{-(\log(\tau)/\lambda - \sqrt{s})^2/2},$$

Computing the resulting integrals yields

$$\mathrm{e}^{\lambda \mathbb{E} X} \leq \binom{n}{s} \mathrm{e}^{-s/2} \lambda \sqrt{2\pi} \mathrm{e}^{(\sqrt{s}+\lambda)^2/2}.$$

Using a standard bound for the binomial coefficients, namely $\binom{n}{s} \leq \mathrm{e}^{s \log(\mathrm{e}n/s)}$, we see

$$\mathrm{e}^{\lambda \mathbb{E} X} \leq \mathrm{e}^{s \log(\mathrm{e}n/s) - s/2 + (\sqrt{s}+\lambda)^2/2 + \log(\lambda) + \log(\sqrt{2\pi})},$$

or equivalently

$$\lambda \mathbb{E} X \leq s \log(\mathrm{e}n/s) - s/2 + (\sqrt{s} + \lambda)^2/2 + \log(\lambda) + \log(\sqrt{2\pi})$$

By setting $\lambda = \sqrt{s \log(\mathrm{e}n/s)}$ and assuming (reasonably) large n, we thus get

$$\mathbb{E} X \leq 5\sqrt{s \log(\mathrm{e}n/s)}.$$

From this, we see that

$$w(K_0^s) \leq 10\sqrt{s \log(\mathrm{e}n/s)}.$$

It follows that the Gaussian mean width of the set of gradient sparse vectors is the same as the mean width of sparse vectors due to the similar structure. If we want to obtain accuracy δ for our reconstruction, according to Theorem 4.3.2, we need to take

$$m = \mathcal{O}\left(\frac{s \log(\mathrm{e}n/s)}{\delta^2}\right)$$

measurements.

In Compressed Sensing, the squared mean width of the set of s-sparse vectors (its so called *statistical dimension*) already determines the number of required measurements in order to recover a sparse signal with basis pursuit. This is the case because the convex hull of the set of sparse vectors can be embedded into the ℓ_1-ball inflated by a constant factor. In the case of TV minimization, as we will see in the following section, this embedding yields a (rather large) constant depending on the dimension.

4.3.3 The extension to gradient compressible vectors needs a new approach

In the previous subsection, we considered exactly gradient sparse vectors. However searching all such vectors x that satisfy $Ax = y$ is certainly not a feasible task. Instead, we want to solve the convex program

$$\min \|z\|_{TV} \text{ with } Az = y,$$

with $\|z\|_{TV} = \|\nabla z\|_1$ the total variation seminorm. Now if we have that $x \in K_0^s$, we get that

$$\|x\|_{TV} \leq 2\|\alpha\|_1 \leq 2\sqrt{s}\|\alpha\|_2 = 2\sqrt{s},$$

with α as in section 4.3.2, so $K_0^s \subset K_{TV}^{2\sqrt{s}} := \{x \in B_2^n \colon \|x\|_{TV} \leq 2\sqrt{s}\}$. As $K_{TV}^{2\sqrt{s}}$ is convex, we even have $\mathrm{ch}(K_0^s) \subset K_{TV}^{2\sqrt{s}}$. We can think of the set $K_{TV}^{2\sqrt{s}}$ as 'gradient- compressible' vectors.

In the proof of Theorem 3.3 in [CX15], the Gaussian width of the set $K_{TV}^{4\sqrt{s}}$ has been calculated via a wavelet-based argument. One obtains that $w(K_{TV}^{2\sqrt{s}}) \leq C\sqrt{\sqrt{ns}\log(2n)}$ with $C \leq 20$ being an absolute constant. In this section we illustrate, that proof techniques different from the ones used in the case of synthesis sparsity are indeed necessary in order to obtain useful results. In the synthesis case, the 1-norm ball of radius $\sqrt{s}$ is contained in the set of s-sparse vectors inflated by a constant factor. This in turn implies that the mean width of the compressible vectors is bounded by a constant times the mean width of the s-sparse vectors.

We will attempt a similar computation, that is to find a constant, such that the set $K_{TV}^{2\sqrt{s}}$ is contained in the 'inflated' set $c_{n,s}\mathrm{ch}(K_0^s)$. Then $w(K_{TV}^{2\sqrt{s}}) \leq c_{n,s}w(K_0^s)$. Although this technique works well for sparse recovery, where $c_{n,s} = 2$, it pityably fails in the case of TV recovery as we will see below.

Let us start with $x \in K_{TV}^{2\sqrt{s}}$. Now we can decompose $J := \mathrm{Jsupp}(x) = J_1 \uplus J_2 \uplus \dots J_p$ with $|J_k| \leq s$ in an ascending manner, i.e., for all $k \in J_i, l \in J_{i+1}$, we have that $\alpha_k < \alpha_l$.

Note that the number p of such sets satisfies $p \leq \frac{n}{s}$. Similarly as above, we now write $x = \sum_{i=1}^{|J|} \alpha_i e_{[j_{i-1}+1,j_i]} = \sum_{k=1}^{p} \sum_{i \in J_k} \alpha_i e_{[j_{i-1}+1,j_i]}$. From this, we see that

$$x = \sum_{k=1}^{p} \|\alpha_{J_k}\|_2 \underbrace{\sum_{i \in J_k} \frac{\alpha_i}{\|\alpha_{J_k}\|_2} e_{[j_{i-1}+1,j_i]}}_{\in K_0^s} .$$

The necessary factor $c_{n,s}$ can be found by bounding the size of $\|\alpha_{J_k}\|_2$, namely

$$\max(\|\alpha_{J_k}\|_2) \leq \sum_{k=1}^{p} \|\alpha_{J_k}\|_2 \overset{C-S}{\leq} \underbrace{\|\alpha\|_2}_{\leq 1} \sqrt{p} \leq \sqrt{\frac{n}{s}}.$$

From this, we see that $K_{TV}^{2\sqrt{s}} \subset \sqrt{\frac{n}{s}} \mathrm{ch}(K_0^s)$. To see that this embedding constant is optimal, we construct a vector, for which it is needed.

To simplify the discussion, suppose that n and s are even and $s|n$. For even n, the vector $x_1 = (\sqrt{\frac{1-(-1)^k \varepsilon}{n}})_k$ has unity norm, lies in $K_{TV}^{2\sqrt{s}}$ for $\varepsilon < \frac{2\sqrt{s}}{n}$ and has jump support on all of $[n]$!

For a vector $x \in \mathbb{R}^n$ and an index set $I \subset [n]$, we define the restriction of x to I by

$$(x|_I)_j := \begin{cases} x_j, \text{ if } j \in I \\ 0, \text{ else.} \end{cases}$$

By splitting $\mathrm{Jsupp}(x_1)$ into sets $J_1, \ldots, J_{n/s}$ and setting $a_k = \sqrt{\frac{n}{s}} x_1|_{J_k} \in K_0^s$, we see that $x_1 = \sum_{k=1}^{n/s} \sqrt{\frac{s}{n}} a_k$ and in order for this to be elements of $c_{n,s}\mathrm{ch}(K_0^s)$, we have to set $c_{n,s} = \sqrt{\frac{n}{s}}$. This follows from

$$x_1 = \sum_{k=1}^{n/s} x_1|_{J_k} = \sum_{k=1}^{n/s} \sqrt{\frac{s}{n} \frac{p}{p}} a_k = \sum_{k=1}^{n/s} \frac{1}{p} \underbrace{\left(\sqrt{\frac{n}{s}} a_k \right)}_{\in \sqrt{\frac{n}{s}} K_0^s} \in \sqrt{\frac{n}{s}} \mathrm{ch}(K_0^s)$$

and no smaller inflation factor than $\sqrt{\frac{n}{s}}$ can suffice.

So from the previous discussion, we get

Lemma 4.3.4. *We have the series of inclusions*

$$\mathrm{ch}(K_0^s) \subset K_{TV}^{2\sqrt{s}} \subset \sqrt{\frac{n}{s}} \mathrm{ch}(K_0^s).$$

In view of the results obtainable for sparse vectors and the ℓ_1-ball, this is very disappointing, because Lemma 4.3.4 now implies that the width of $K_{TV}^{2\sqrt{s}}$ satisfies

$$w(K_{TV}^{2\sqrt{s}}) \leq w\left(\sqrt{\frac{n}{s}}\operatorname{ch}(K_0^s)\right) = \sqrt{\frac{n}{s}}w(K_0^s) \leq 10\sqrt{n\log(\mathrm{e}(n-1)/s)},$$

which is highly suboptimal.

Luckily, the results in [CX15] suggest, that the factor n in the previous equation can be replaced by $\sqrt{sn}$. However, they have to resort to a direct calculation of the Gaussian width of $K_{TV}^{2\sqrt{s}}$. The intuition why the Gaussian mean width can be significantly smaller than the bound given in Lemma 4.3.4 stems from the fact, that in order to obtain an inclusion we need to capure all 'outliers' of the set - no matter how small their measure is.

4.3.4 Exact recovery

For exact recovery, the M^*-bound is not suitable anymore and, as suggested in [Ver15], we will use 'Gordon's escape through the mesh' in order to find conditions on exact recovery. Exact recovery for TV minimization via this approach has first been considered in [CX15].

Suppose, we want to recover $x \in K_0^s$ from Gaussian measurements $Ax = y$. Given, that we want our estimator $\hat{x}$ to lie in a set K, exact recovery is achieved, if $K \cap \{z\colon Az = y\} = \{x\}$. This is equivalent to requiring

$$(K - x) \cap \underbrace{\{z - x\colon Az = y\}}_{=\ker(A)} = \{0\}.$$

With the descent cone $D(K, x) = \{t(z - x)\colon t \geq 0, z \in K\}$, we can rewrite this condition as

$$D(K, x) \cap \ker(A) = \{0\},$$

by introducing the set $S(K, x) = D(K, x) \cap B_2^n$, we see that if

$$S(K, x) \cap \ker(A) = \emptyset,$$

we get exact recovery. The question, when a section of a subset of the sphere with a random hyperplane is empty is answered by Gordon's escape through a mesh.

Theorem 4.3.5 ([Gor88])**.** *Let $S \subset \mathbb{S}^{n-1}$ be fixed and $E \in G_{n,n-m}$ be drawn at random according to the Haar measure. Assume that $\hat{w}(S) = \mathbb{E}\sup_{u\in S}\langle g, u\rangle < \sqrt{m}$, then $S \cap E = \emptyset$*

with probability exceeding

$$1 - 2.5 \exp\left(-\frac{(m/\sqrt{m+1} - \hat{w}(S))^2}{18}\right).$$

So we get exact recovery with high probability from a program given in Theorem 4.3.2 or 4.3.3, provided that $m > \hat{w}(S(K, x_0))^2$.

Let's see how this applies to TV minimization. Suppose, we are given $x \in K_0^s$ and Gaussian measurements $Ax = y$. Solving

$$\min \|z\|_{TV} \text{ with } Az = y,$$

amounts to using the Minkowski functional of the set $K = \{z \in \mathbb{R}^n \colon \|z\|_{TV} \leq \|x\|_{TV}\}$, which is a scaled TV-Ball.

In [CX15], the null space property for TV minimization given in Corollary 4.2.1 has been used in order to obtain recovery guarantees.

They consider the set, where this condition is not met

$$\mathcal{S} := \{x' \in B_2^n \colon \exists J \subset [n], |J| \leq s, \|(\nabla x')_J\|_1 \geq \|(\nabla x')_{J^c}\|_1\},$$

and apply Gordon's escape through the mesh to see that with high probability, its intersection with the kernel of A is empty, thus proving exact recovery with high probability. Their estimate to the mean width of the set $\mathcal{S}$,

$$\hat{w}(\mathcal{S}) \leq c\sqrt[4]{ns}\sqrt{\log(2n)}$$

with $c < 19$ is essentially optimal (up to logarithmic factors), as they also show that $w(\mathcal{S}) \geq C\sqrt[4]{ns}$. So uniform exact recovery can only be expected for $m = \mathcal{O}(\sqrt{sn}\log n)$ measurements.

Let us examine some connections to the previous discussion about the descent cone.

Lemma 4.3.6. *We have that for $K = \{z \in \mathbb{R}^n \colon \|z\|_{TV} \leq \|x\|_{TV}\}$ defined as above and $x \in K_0^s$, it holds that $S(K, x) \subset \mathcal{S}$.*

Proof. Let $y \in S(K, x)$. Then there exists a $x \neq z \in K$, such that $y = \frac{z-x}{\|z-x\|_2}$. Set $J = \operatorname{Jsupp}(x)$, then, as $z \in K$, we have that $\|z\|_{TV} \leq \|x\|_{TV}$, or

$$\sum_{i \in J} |(\nabla x)_i| \geq \sum_{i \in J} |(\nabla z)_i| + \sum_{i \notin J} |(\nabla z)_i|$$

Now, by the triangle inequality and this observation, we have

$$\sum_{i\in J} |(\nabla x)_i - (\nabla z)_i| \geq \sum_{i\in J} |(\nabla x)_i| - |(\nabla z)_i| \geq \sum_{i\notin J} |(\nabla z)_i| = \sum_{i\notin J} |(\nabla x)_i - (\nabla z)_i|.$$

The last equality follows from the fact that ∇x is zero outside of the gradient support of x. Multiplying both sides with $\frac{1}{\|z-x\|_2}$ gives the desired result

$$\|(\nabla y)_J\|_1 = \frac{1}{\|z-x\|_2} \sum_{i\in J} |(\nabla x)_i - (\nabla z)_i| \geq \ \geq \frac{1}{\|z-x\|_2} \sum_{i\notin J} |(\nabla x)_i - (\nabla z)_i| = \|(\nabla y)_{J^c}\|_1.$$

□

The previous lemma shows that the recovery guarantees derived from the null space property and via the descent cone are actually connected in a very simple way.

Clearly, now if we do not intersect the set $\mathcal{S}$, we also do not intersect the set $S(K,x)$, which yields exact recovery for example with the same upper bounds on m as for $\mathcal{S}$. Even more specifically, in the calculation of $\hat{w}(\mathcal{S})$ given in [CX15], an embedding into a slightly larger set $\tilde{\mathcal{S}} = \{x \in B_2^n \colon \|x\|_{TV} \leq 4\sqrt{s}\}$ is made. This embedding can also quite easily be done if we note that $\|x\|_{TV} \leq 2\sqrt{s}$, as we showed above and $\|z\|_{TV} \leq \|x\|_{TV}$.

Note that the same discussion also holds for higher dimensional signals, such that the improved numbers of measurements as given in Section 4.2.2 can be applied.

4.3.5 Subgaussian measurements

Up to this point, all our measurement matrices have been assumed to consist of i.i.d. Gaussian random variables. We will reduce this requirement in this section to be able to incorporate also subgaussian measurement matrices into our framework.

Definition 4.3.4. *A real valued random variable X is called* subgaussian*, if there exists a number $t > 0$, such that $\mathbb{E}e^{tX^2} < \infty$. A real valued random vector is called subgaussian, if all of its one dimensional marginals are subgaussian.*

An obvious example of subgaussian random variables are Gaussian random variables, as the expectation in Definition 4.3.4 exists for all $t < 1$. Also, all bounded random variables are subgaussian.

Here, we rely on results given by Tropp in [Tro15] using the results of Mendelson [KM15, Men14]. We will consider problems of the form

$$\min \|z\|_{TV} \text{ such that } \|Az - y\| \leq \varepsilon, \tag{4.12}$$

where A is supposed to be a matrix with independent subgaussian rows. Furthermore, we denote the exact solution by x_0, i.e., $Ax_0 = y$. We pose the following assumptions on the distribution of the rows of A.

(M1) $\mathbb{E}A_i = 0$,

(M2) There exists $\alpha > 0$, such that for all $u \in \mathbb{S}^{n-1}$ it holds that $\mathbb{E}|\langle A_i, u\rangle| \geq \alpha$,

(M3) There is a $\sigma > 0$, such that for all $u \in \mathbb{S}^{n-1}$ it holds that $\mathbb{P}(|\langle A_i, u\rangle| \geq t) \leq 2\exp(-t^2/(2\sigma^2))$,

(M4) The constant $\rho := \frac{\sigma}{\alpha}$ is small.

Then the small ball methods yields the following recovery guarantee (we present the version of [Tro15]).

Theorem 4.3.7. *Let $x_0 \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ be a subgaussian matrix satisfying (M1)-(M4) above. Furthermore, set $y = Ax_0 + e$, where $\|e\| \leq \varepsilon$ denotes the (bounded) error of the measurement. Then the solution $\hat{x}$ of* (4.12) *satisfies*

$$\|\hat{x} - x_0\| \leq \frac{2\varepsilon}{\max\{c\alpha\rho^{-2}\sqrt{m} - C\sigma w(S(K, x_0)) - \alpha t, 0\}}$$

with probability exceeding $1 - \mathrm{e}^{-ct^2}$. $D(K, x_0)$ denotes the descent cone of the set K at x_0, as defined in the previous section.

From this we see that, provided

$$m \geq \tilde{C}\rho^6 w^2(S(K, x_0)),$$

we obtain stable reconstruction of our original vector from (4.12). Note that the theorem is only meaningful for $t = \mathcal{O}(\sqrt{m})$, as otherwise the denominator vanishes.

In the previous section, we have shown the inclusion $S(K, x_0) \subset \mathcal{S}$ for $x_0 \in K_s^0$ and hence we have that

$$w(S(K, x_0) \leq w(\mathcal{S}) \leq c\sqrt[4]{ns}\sqrt{\log(2n)}.$$

So we see that for $m \geq \tilde{C}\rho^6\sqrt{ns}\log(2n)$, we obtain the bound

$$\begin{aligned}\|\hat{x} - x_0\| &\leq \frac{2\varepsilon}{\max\{c\alpha\rho^{-2}\sqrt{\tilde{C}}\rho^3\sqrt[4]{ns}\sqrt{\log(2n)} - C\sigma\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t, 0\}} \\ &= \frac{2\varepsilon}{\max\{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t, 0\}}\end{aligned}$$

with high probability. We conclude that, given the absolute constants c, C, we need to set $\tilde{C} \geq \frac{C^2}{c^2}$ in order to obtain a meaningful result. Combining all our previous discussions with Theorem 4.3.7, we get

Theorem 4.3.8. *Let $x_0 \in \mathbb{R}^n$, $m \geq \tilde{C}\rho^6\sqrt{ns}\log(2n)$ and $A \in \mathbb{R}^{m\times n}$ be a subgaussian matrix satisfying (M1)-(M4). Furthermore, set $y = Ax_0 + e$, where $\|e\| \leq \varepsilon$ denotes the (bounded) error of the measurement, constants $c, C, \tilde{C} > 0$ as above and $t \leq \frac{\sigma(c\sqrt{\tilde{C}}-C)\sqrt[4]{ns}\sqrt{\log(2n)}}{\alpha}$. Then the solution $\hat{x}$ of* (4.12) *satisfies*

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}\sqrt{\log(2n)} - \alpha t}\right) \leq \mathrm{e}^{-ct^2}.$$

We can for example set $t = \rho(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}$ (for $n \geq 2$) to obtain the bound

$$\mathbb{P}\left(\|\hat{x} - x_0\| > \frac{2\varepsilon}{\sigma(c\sqrt{\tilde{C}} - C)\sqrt[4]{ns}(\sqrt{\log(2n)} - 1)}\right) \leq \mathrm{e}^{-\tilde{c}\rho\sqrt{ns}}.$$

For example for i.i.d. standard Gaussian measurements, the constant $\rho = \sqrt{\frac{2}{\pi}}$.

Note that in the case of noisefree measurements $\varepsilon = 0$, Theorem 4.3.8 gives an exact recovery result for a wider class of measurement ensembles with high probability. Furthermore with a detailed computation of $w(S(K, x_0))$ one may be able to improve the number of measurements for nonuniform recovery. It also remains open, whether the lower bounds of Cai and Xu for the case of Gaussian measurements can be generalized to the subgaussian case. In fact, our numerical experiments summarized in Figure 4.2 suggest a better scaling in the ambient dimension, around $N^{1/4}$, in the average case. We consider it an interesting problem for future work to explore whether this is due to a difference between Rademacher and Gaussian matrix entries, between uniform and nonuniform recovery, or between the average and the worst case. Also, it is not clear whether the scaling is in fact $N^{1/4}$ or if the observed slope is just a linearization of, say, a logarithmic dependence.

4.4 Discussion and open problems

As the considerations in the previous sections illustrate, the mathematical properties of total variation minimization differ significantly from algorithms based on synthesis sparsity, especially in one dimension. For this reason, there are a number of questions that have been answered for synthesis sparsity, but which are still open for the framework of total variation minimization. For example, the analysis provided in [RRT12b, KMR14b] for deterministically subsampled partial

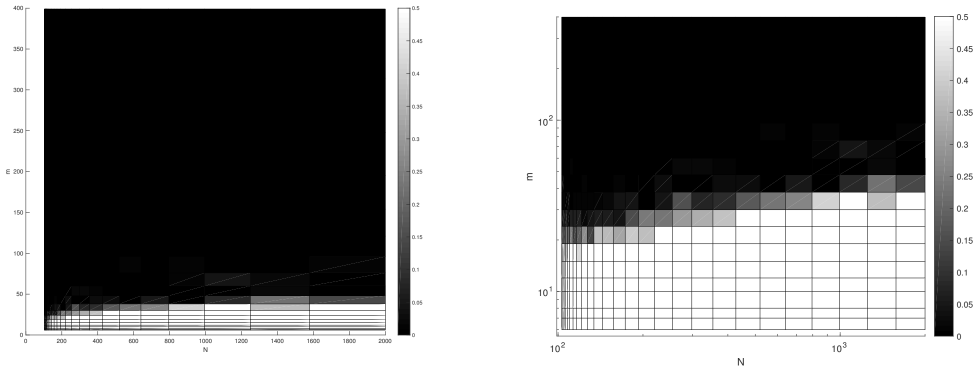

Figure 4.2: Average error of recovery from Rademacher measurements in $1d$ with m measurements and ambient dimension N for fixed cosparsity level $s = 5$. Left: linear axis scaling, Right: logarithmic axis scaling. The slope of the phase transition in the log-log plot is observed to be about $\frac{1}{4}$.

random circulant matrices, as they are used to model measurement setups appearing in remote sensing or coded aperture imaging, could not be generalized to total variation minimization. The difficulty in this setup is that the randomness is encoded by the convolution filter, so it is not clear what the analogy of variable density sampling would be.

Another case of practical interest is that of sparse $0/1$ measurement matrices. Recently it has been suggested that such meausurements increase efficiency in photoacoustic tomography, while at the same time, the signals to be recovered (after a suitable temporal transform) are approximately gradient sparse. This suggests the use of total variation minimization for recovery, and indeed empirically, this approaches yields good recovery results [SKB$^+$15]. Theoretical guarantees, however, (as they are known for synthesis sparse signals via an expander graph construction [BGI$^+$08]) are not available to date for this setup.

Acknowledgements

FK and MS acknowledge support by the Hausdorff Institute for Mathematics (HIM), where part of this work was completed in the context of the HIM trimester program ”Mathematics of Signal Processing”, FK and CK acknowledge support by the German Science Foundation in the context of the Emmy Noether Junior Research Group “Randomized Sensing and Quantization of Signals and Images” (KR 4512/1-1) and by the German Ministry of Research and Education in the context of the joint research initiative ZeMat. MS has been supported by the Austrian Science Fund (FWF) under Grant no. Y760 and the DFG SFB/TRR 109 ”Discretization in Geometry and Dynamics”.

5 Online and Stable Learning of Analysis Operators

This chapter is based on joint work with

K. SCHNASS[1].

It has been submitted to IEEE Transactions on Signal Processing. A preprint can be found on arXiv [SS17].

[1] Department of Mathematics, University of Innsbruck, Technikestraße 13, 6020 Innsbruck, Austria. E-mail: karin.schnass@uibk.ac.at

Interlude

The previous chapter was devoted to the study of total variation regularization, i.e. using the regularizer $\|\nabla \cdot\|_1$. This is a special case of analyisis-cosparse regularization, where one knows that there is some operator Ω, such that its application on a given signal produces a sparse vector. This means that regularization by $\|\Omega \cdot\|_1$ would be a good choice. However, typically one does not know, which operator Ω is suited for the given class of data. This is why in this chapter, four iterative algorithms for learning such analysis operators are presented. They are built upon the same optimization principle underlying both Analysis K-SVD and Analysis SimCO. The Forward and Sequential Analysis Operator Learning (FAOL and SAOL) algorithms are based on projected gradient descent with optimally chosen step size. The Implicit AOL (IAOL) algorithm is inspired by the implicit Euler scheme for solving ordinary differential equations and does not require to choose a step size. The fourth algorithm, Singular Value AOL (SVAOL), uses a similar strategy as Analysis K-SVD while avoiding its high computational cost. All algorithms are proven to decrease or preserve the target function in each step and a characterisation of their stationary points is provided. Further, they are tested on synthetic and image data, compared to Analysis SimCO and found to give better recovery rates and faster decay of the objective function respectively. In a final denoising experiment the presented algorithms are again shown to perform similar to or better than the state-of-the-art algorithm ASimCO.

5.1 Introduction

Many tasks in high dimensional signal processing, such as denoising or reconstruction from incomplete information, can be efficiently solved if the data at hand is known to have intrinsic low dimension. One popular model with intrinsic low dimension is the union of subspaces model, where every signal is assumed to lie in one of the low dimensional linear subspaces. However, as the number of subspaces increases, the model becomes more and more cumbersome to use unless the subspaces can be parametrised. Two examples of large unions of parametrised subspaces, that have been successfully employed, are sparsity in a dictionary and cosparsity in an analysis operator. In the sparse model the subspaces correspond to the linear span of just a few normalised columns, also known as atoms, from a $d \times K$ dictionary matrix, $\Phi = (\phi_1 \dots \phi_K)$ with $\|\phi_k\|_2 = 1$, meaning, any data point y can be approximately represented as superposition of $S \ll d$ dictionary elements. If we denote the restriction of the dictionary to the atoms/columns indexed by I as Φ_I, we have

$$y \in \bigcup_{|I| \leq S} \operatorname{colspan} \Phi_I, \quad \text{or} \quad y \approx \Phi x, \quad \text{with } x \text{ sparse.}$$

In the cosparse model the subspaces correspond to the orthogonal complement of the span of some normalised rows, also known as analysers, from a $K \times d$ analysis operator $\Omega = (\omega_1^\star \dots \omega_K^\star)^\star$ with $\|\omega_k\|_2 = 1$. This means that any data point y is orthogonal to ℓ analysers or in other words that the vector Ωy has ℓ zero entries and is sparse. If we denote the restriction of the analysis operator to the analysers/rows indexed by J as Ω_J, we have

$$y \in \bigcup_{|J| \geq \ell} (\operatorname{rowspan} \Omega_J)^\perp, \quad \text{or} \quad \Omega y \approx z, \quad \text{with } z \text{ sparse.}$$

Note that in this model it is not required that the signals lie in the span of $\Omega^\star$, in particular $\Omega^\star \Omega$ need not be invertible. Before being able to exploit these models for a given data class, it is necessary to identify the parametrising dictionary or analysis operator. This can be done either via a theoretical analysis or a learning approach. While dictionary learning is by now an established field, see [RBE10] for an introductory survey, results in analysis operator learning are still countable, [YNGD11, RB13, NDEG13, YNGD13, HKD13, RPE13, DWD14, EB14, GNE$^+$14, SWGK16, DWD$^+$16].

Most algorithms are based on the minimization of a target function together with various constraints on the analysis operator. In [YNGD11, YNGD13], an ℓ_1-norm target function together with the assumption that the operator is a unit norm tight frame was used. In transform learning [RB13], in addition to the sparsity of ΩY, a regularisation term amounting to the negative log-determinant of Ω is introduced to enforce full rank of the operator. Note that due to the nature of this penalty term overcomplete operators cannot be considered. The geometric analysis operator learning framework [HKD13] uses operators with unit norm rows, full rank and furthermore imposes the restriction that none of the rows are linearly dependent. The last assumption is an additional restriction only if the considered operators are overcomplete. These assumptions admit a manifold structure and a Riemannian gradient descent algorithm is used in order to find the operator. Analysis K-SVD [RPE13] uses an analysis sparse coding step to find the cosupports for the given data and then computes the singular vector corresponding to the smallest singular value of a matrix computed from the data. Finally, in [DWD14, DWD$^+$16] a projected gradient descent based algorithm with line search, called Analysis SimCO, is presented. There, the only restriction on the analysis operator is that its rows are normalised and the target function enforces sparsity of ΩY.

Contribution: In this work we will contribute to the development of the field by developing four algorithms for learning analysis operators, which improve over state of the art algorithms such as Analysis K-SVD, [RPE13] and Analysis SimCo, [DWD14, DWD$^+$16], in terms of convergence speed, memory complexity and performance.

Outline: The paper is organised as follows. After introducing the necessary notation, in the next section we will remotivate the optimization principle that is the starting point of A-KSVD and ASimCO and shortly discuss the advantages and disadvantages of the two algorithms. We then take a gradient descent approach similar to ASimCO, replacing the line search with an optimal choice for the step size, resulting in the Forward Analysis Operator Learning algorithm (FAOL). In order to obtain an online algorithm, which processes the training data sequentially, we devise an estimation procedure for the quantities involved in the step size calculation leading to the Sequential Analysis Operator Learning algorithm (SAOL) and test the presented algorithms both on synthetic and image data. Inspired by the implicit Euler scheme for solving ordinary differential equations and the analysis of some special solutions of these equations, in Section 5.3, we gradually invest in the memory requirements and computational complexity per iteration of our schemes in return for avoiding the stepsize altogether and overall faster convergence, leading to Implicit (IAOL) and Singular Vector Analysis Operator Learning (SVAOL). After testing the new algorithms on synthetic and image data demonstrating the improved recovery rates and convergence speed with respect to ASimCO, in Section 5.4 we apply them to image denoising again in comparison to ASimCO. Finally, in the last section we provide a short discussion of our results and point out future directions of research.

Notation: Before hitting the slopes, we summarise the notational conventions used throughout this paper. The operators Ω and Γ will always denote matrices in $\mathbb{R}^{K\times d}$ and for a matrix A we denote its transpose by $A^\star$. More specifically, we will mostly consider matrices in the manifold

$\mathcal{A} := \{\Gamma \in \mathbb{R}^{K\times d} \colon \forall k \in [K] \colon \|\gamma_k\|_2 = 1\}$, where γ_k denotes the k-th row of the matrix Γ. By $[n]$, we denote the set $\{1, 2, \ldots, n\}$ and we adopt the standard notation $|M|$ for the cardinality of a set M. By Γ_J with $J \subset [K]$ we denote the restriction of Γ to the rows indexed by J.

A vector $y \in \mathbb{R}^d$ is called ℓ-cosparse with respect to Ω, if there is an index set $\Lambda \subset [K]$ with $|\Lambda| = \ell$, such that $\Omega_\Lambda y = 0$. The support of a vector $x \in \mathbb{R}^K$ is defined as $\operatorname{supp}(x) = \{k \in [K] \colon x_k \neq 0\}$ and the cosupport accordingly as $\operatorname{cosupp}(x) = \{k \in [K] \colon x_k = 0\}$. Note that by definition we have $\operatorname{supp}(x) \cup \operatorname{cosupp}(x) = [K]$. For the runtime complexity $R(n)$, we adopt standard Landau notation, i.e. $R(n) = \mathcal{O}(f(n))$ means, there is a constant $C > 0$, such that for large n, the runtime $R(n)$ satisfies $R(n) \leq Cf(n)$.

Finally, the Frobenius norm of a matrix A is defined by $\|A\|_F^2 := \operatorname{tr}(A^\star A)$.

5.2 Two explicit analysis operator learning algorithms - FAOL and SAOL

Since optimization principles have already successfully led to online algorithms for dictionary learning, [Sch15a, Sch16], we will start our quest for an online algorithm by motivating a suitable optimization principle for analysis operator learning. Suppose, we are given signals $y_n \in \mathbb{R}^d$ that are perfectly cosparse in an operator Ω, i.e. Ωy_n has ℓ zero entries or equivalently $\Omega y_n - x_n = 0$ for some x_n which has $K - \ell$ non-zero entries. If we collect the signals y_n as columns in the matrix $Y = (y_1 \ldots y_N)$, then by construction we have $\Omega Y - X = 0$ for some $X \in \mathcal{X}_\ell$ with $\mathcal{X}_\ell := \{(x_1, x_2, \ldots, x_N) \in \mathbb{R}^{K\times N} \colon |\operatorname{supp}(x_n)| = K - \ell\}$. In the more realistic scenario, where the signals are not perfectly cosparse, we should still have $\Omega Y - X \approx 0$, which naturally leads to the following minimization program to recover Ω,

$$\operatorname*{arg\,min}_{\Gamma\in\mathcal{A}, X\in\mathcal{X}_\ell} \|\Gamma Y - X\|_F^2. \tag{5.1}$$

Apart from additional side constraints on Γ, such as incoherence, the optimization program above has already been used successfully as starting point for the development of two analysis operator learning algorithms, Analysis K-SVD [RPE13] and Analysis SimCO [DWD14, DWD$^+$16]. AKSVD is an alternating minimization algorithm, which alternates between finding the best $X \in \mathcal{X}_\ell$ for the current Γ and updating Γ based on the current X. The cosparse approximation scheme used there is quite cumbersome and costly, which means that the algorithm soon becomes intractable as d increases. ASimCO is a (gradient) descent algorithm with line search. It produces results similar to AKSVD and has the advantage that it does so with a fraction of the computational cost. Still, at closer inspection we see that the algorithm has some problematic aspects. The line search cannot be realised resource efficiently, since in each step several evaluations of the target function are necessary, which take up a lot of computation time. Moreover, for each of these function evaluations we must either reuse the training data, thus incurring high storage costs, or use a new batch of data, thus needing a huge amount of training samples. Still, if we consider the speedup of ASimCO with respect to AKSVD we see that gradient descent is a promising approach if we can avoid the line search and its associated problems.

To see that a gradient descent based algorithm for our problem can also be sequential, let us rewrite our target function, $g_N(\Gamma) = \min_{X \in \mathcal{X}_\ell} \|\Gamma Y - X\|_F^2$. Abbreviating $\Lambda_n = \operatorname{supp}(x_n)$ and $\Lambda_n^c = \operatorname{cosupp}(x_n)$, we have

$$\begin{aligned} g_N(\Gamma) &= \sum_{n=1}^{N} \min_{x_n : |\Lambda_n| = K-\ell} \|\Gamma y_n - x_n\|_2^2 = \\ &= \sum_{n=1}^{N} \min_{x_n : |\Lambda_n| = K-\ell} (\|\Gamma_{\Lambda_n^c} y_n\|_2^2 + \underbrace{\|\Gamma_{\Lambda_n} y_n - x_n\|_2^2}_{=0}) \\ &= \sum_{n=1}^{N} \min_{|J|=\ell} \|\Gamma_J y_n\|_2^2 =: f_N(\Gamma). \end{aligned}$$

Since the gradient of a sum of functions is the sum of the gradients of these functions, from f_N we see that the gradient of our objective function can be calculated in an online fashion.

Before going into more details about how to avoid a line search and stay sequential, let us lose a few words about the uniqueness of the minima of our objective function.

If the signals are perfectly cosparse in Ω, clearly there is a global minimum of f_N at Ω. However, one can easily see that all permutations and sign flips of rows of Ω are also minimizers of f_N. We call these the *trivial ambiguities*. The more interesting question is whether there are other global or local minima?

This question is best answered with an example. Assume that all our training signals are (perfectly) ℓ-cosparse in Ω but lie in a subspace of $\mathbb{R}^d$. In this case we can construct a continuum of operators Γ, which also satisfy $f_N(\Gamma) = 0$ by choosing a vector v with $\|v\|_2 = 1$ in the orthogonal complement of this subspace, and by setting $\gamma_k = a_k\omega_k + b_k v$ for some $a_k^2 + b_k^2 = 1$. This example indicates that isotropy in the data is important for our problem to be well posed. On the other hand, in case the data has such a low dimensional structure, which can be found via a singular value decomposition of $Y^\star Y$, it is easy to transform the ill posed problem into a well posed one. Armed with the non-zero singular vectors, we just have to project our data onto the lower dimensional space spanned by these vectors and learn the analysis operator within this lower dimensional space. In the following, we assume for simplicity that any such preprocessing has already been done and that the data isotropically occupies the full ambient space $\mathbb{R}^d$ or equivalently that $Y^\star Y$ is well conditioned.

5.2.1 Minimzing f_N

As mentioned above in order to get an online algorithm we want to use a gradient descent approach but avoid the line search. Our strategy will be to use projected stochastic gradient-type descent with carefully chosen stepsize. Given the current estimate of the analysis operator Γ, one step of (standard) gradient descent takes the form

$$\bar{\Gamma} = \Gamma - \alpha \nabla f_N\left(\Gamma\right).$$

Let us calculate the gradient $\nabla f_N\left(\Gamma\right)$ wherever it exists. Denote by J_n the set[1] for which $\|\Gamma_{J_n} y_n\|_2^2 = \min_{|J|=\ell} \|\Gamma_J y_n\|_2^2$, then the derivative of f_N with respect to a row γ_k of Γ is

$$\frac{\partial f_N}{\partial \gamma_k}(\Gamma) = \sum_{n=1}^{N} \sum_{j \in J_n} \frac{\partial}{\partial \gamma_k} \langle \gamma_j, y_n \rangle^2 = \sum_{n=1}^{N} \sum_{j \in J_n} 2 \langle \gamma_j, y_n \rangle y_n^\star \delta_{kj} = \sum_{n\colon k \in J_n} 2 \langle \gamma_k, y_n \rangle y_n^\star =: 2 g_k. \tag{5.2}$$

Note that as expected the vectors g_k can be calculated online, that is given a continuous stream of data y_n, we compute J_n, update all g_k for $k \in J_n$, and forget the existence of y_n. After processing all signals, we set

$$\bar{\gamma}_k = \left(\gamma_k - \alpha_k g_k\right) \beta_k. \tag{5.3}$$

where $\beta_k = \|\gamma_k - \alpha_k g_k\|_2^{-1}$ is a factor ensuring normalisation of $\bar{\gamma}_k$. This normalisation corresponds to a projection onto the manifold $\mathcal{A}$ and is necessary, since a standard descent step will most likely take us out of the manifold. If we compare to dictionary learning, e.g. [Sch15a], it is interesting to observe that we cannot simply choose α_k by solving the linearised optimization problem with side constraints using Lagrange multipliers, since this would lead to a zero-update $\bar{\gamma}_k = 0$.

In order to find the correct descent parameter, note that the current value of the target function is given by

$$f_N(\Gamma) = \sum_{n=1}^{N} \sum_{k \in J_n} |\langle \gamma_k, y_n \rangle|^2 = \sum_{k=1}^{K} \sum_{n\colon k \in J_n} |\langle \gamma_k, y_n \rangle|^2.$$

[1] The careful reader will observe that the set J_n might not be unique for every y_n and Γ. If for a given Γ at least one J_n is not uniquely determined and $\min_{|J|=\ell} \|\Gamma_J y_n\|_2^2 > 0$, then the target function is not differentiable in Γ. For simplicity we will continue the presentation as if the J_n where uniquely determined, keeping in mind that the derived descent direction only coincides with the gradient where it exists.

Defining $A_k := \sum_{n\colon k\in J_n} y_n y_n^\star$, we see that $f_N(\Gamma) = \sum_{k=1}^K \gamma_k A_k \gamma_k^\star$ and we can optimally decrease the objective function by choosing α_k, such that it minimizes $\bar{\gamma}_k A_k \bar{\gamma}_k^\star$. Note also that with this definition, the descent directions g_k defined in Equation (5.2) are given by $g_k = \gamma_k A_k$. First assume that $g_k \neq 0$, or more generally $g_k \neq \lambda_k \gamma_k$. In case $g_k = 0$ the part of the objective function associated to γ_k is already zero and cannot be further reduced, while in case $g_k = \lambda_k \gamma_k$ any admissible stepsize not leading to the zero vector preserves the current analyser, that is $\bar{\gamma}_k = \gamma_k$. To optimally decrease the target function, we need to solve

$$\alpha_k = \arg\min_{\alpha} \frac{(\gamma_k - \alpha g_k) A_k (\gamma_k - \alpha g_k)^\star}{\|(\gamma_k - \alpha g_k)\|^2}. \tag{5.4}$$

Defining $a_k = \gamma_k A_k \gamma_k^\star$, $b_k = \gamma_k A_k^2 \gamma_k^\star$ and $c_k = \gamma_k A_k^3 \gamma_k^\star$ a short computation given in Appendix 5.6 shows that whenever $b_k^2 \neq a_k c_k$ the optimal stepsize has the form,

$$\alpha_k = \frac{a_k b_k - c_k + \sqrt{(c_k - a_k b_k)^2 - 4(b_k^2 - a_k c_k)(a_k^2 - b_k)}}{2(b_k^2 - a_k c_k)}.$$

If $b_k^2 = a_k c_k$ and $b_k \neq 0$, the optimal stepsize is $\alpha_k = \frac{a_k}{b_k}$. Finally, if $b_k = \|A_k \gamma_k^\star\|_2^2 = 0$ it follows that $A_k \gamma_k^\star = 0$ and therefore also $a_k = c_k = 0$. In this case we set $\alpha_k = 0$, as $\gamma_k A_k \gamma_k^\star$ is already minimal.

We summarise the first version of our derived algorithm, called Forward Analysis Operator Learning (FAOL) in Table 5.1. As input parameters, it takes the current estimate of the analysis operator $\Gamma \in \mathbb{R}^{K\times d}$, the cosparsity parameter ℓ and N training signals $Y = (y_1, y_2, \dots, y_N)$. As a result of the optimal stepsize choice we can prove the following theorem characterising the behaviour of the FAOL algorithm.

Theorem 5.2.1. *The FAOL algorithm decreases or preserves the value of the target function in each iteration.*

Preservation rather than decrease of the target function can only occur if all rows γ_k of the current iterate Γ are eigenvectors of the matrix $A_k(\Gamma)$.

Proof. To prove the first part of the theorem observe that

$$f(\Gamma) = \sum_{k=1}^K \sum_{n\colon k\in J_n} |\langle \gamma_k, y_n\rangle|^2 \geq \sum_{k=1}^K \sum_{n\colon k\in J_n} |\langle \bar{\gamma}_k, y_n\rangle|^2 \geq \sum_{k=1}^K \sum_{n\colon k\in \bar{J}_n} |\langle \bar{\gamma}_k, y_n\rangle|^2 = f(\bar{\Gamma}).$$

where $\bar{J}_n$ denotes the minimizing set for y_n based on $\bar{\Gamma}$. The first inequality follows from the choice of α_k and the second inequality follows from the definition of the sets J_n and $\bar{J}_n$.

The second part of the theorem is a direct consequence of the derivation of α_k given in Ap-

```
FAOL($\Gamma, \ell, Y$) - (one iteration)

• For all $n \in [N]$:
    – Find $J_n = \arg\min_{|J|=\ell} \|\Gamma_J y_n\|_2^2$.
    – For all $k \in [K]$ update $A_k = A_k + y_n y_n^\star$ if $k \in J_n$.
• For all $k \in [K]$:
    – Set $a = \gamma_k A_k \gamma_k^\star$, $b = \gamma_k A_k^2 \gamma_k^\star$ and $c = \gamma_k A_k^3 \gamma_k^\star$.
    – If $b^2 - ac \neq 0$, set $\alpha_k := \frac{ab-c+\sqrt{(c-ab)^2-4(b^2-ac)(a^2-b)}}{2(b^2-ac)}$.
    – If $b^2 - ac = 0$ and $b \neq 0$, set $\alpha_k := \frac{a}{b}$.
    – If $b^2 - ac = 0$ and $b = 0$, set $\alpha_k := 0$.
    – Set $\bar{\gamma}_k = \gamma_k(\mathbb{1} - \alpha_k A_k)$.
Output $\bar{\Gamma} = (\frac{\bar{\gamma}_1}{\|\bar{\gamma}_1\|_2}, \dots, \frac{\bar{\gamma}_K}{\|\bar{\gamma}_K\|_2})^\star$.
```

Table 5.1: The FAOL algorithm

pendix 5.6. □

Let us shortly discuss the implications of Theorem 5.2.1. It shows that the sequence of values of the target function $v_k = f(\Gamma^{(k)})$ converges. This, however, does not imply convergence of the algorithm as suggested in [DWD$^+$16], at least not in the sense that the sequence $\Gamma^{(k)}$ converges. Indeed the sequence $\Gamma^{(k)}$ could orbit around the set $\mathcal{L} = \{\Gamma \in \mathcal{A} \colon f(\Gamma) = v\}$, where $v = \lim_{k\to\infty} v_k$. If this set contains more than one element, there need not exist a limit point of the sequence $\Gamma^{(k)}$. Nevertheless, due to compactness of the manifold $\mathcal{A}$, we can always find a subsequence, that converges to an element $\Gamma \in \mathcal{L}$. In order to avoid getting trapped in such orbital trajectories, in numerical experiments we draw a fresh batch of signals $y_1, \dots, y_N$ in each iteration of the algorithm.

We proceed with an analysis of the runtime complexity of the FAOL algorithm. The cost of finding the support sets $S_k = \{n \colon k \in J_n\}$ of average size $N\ell/K$ in the FAOL algorithm is $\mathcal{O}(dKN)$ amounting to the multiplication of the current iterate Γ with the data matrix Y and subsequent thresholding. We can now either store the K matrices A_k of size $d \times d$ amounting to a memory complexity of $\mathcal{O}(Kd^2)$ or store the data matrix Y and the optimal cosupports S_k requiring memory on the order of $\mathcal{O}(dN)$ and $\mathcal{O}(\ell N)$, respectively. Setting up all matrices A_k takes $\mathcal{O}(d^2N)$ multiplications and $\mathcal{O}(\ell d^2 N)$ additions, if done sequentially, and dominates the cost of calculating a_k, b_k, c_k. Denote by Y_k the submatrix of the data matrix Y with columns indexed by S_k. Note that with this convention we have $A_k = Y_k Y_k^\star$. If we store the data matrix Y and the sets S_k, we can also compute all necessary quantities via $g_k = (\gamma_k Y_k)Y_k^\star$, $a = \langle g_k, \gamma_k \rangle$, $b = \langle g_k, g_k \rangle$ and $c = \langle g_k Y_k, g_k Y_k \rangle$ altogether amounting to $\mathcal{O}(\ell d N)$ floating point operations, as in this case only matrix-vector products have to be computed. So while the memory complexity of the first approach might be smaller depending on the amount of training data, the second approach has a reduced computational complexity.

If we are now given a continuous stream of high dimensional data, it is not desirable to store either the matrices A_k or the data matrix Y, so as a next step we will reformulate the FAOL algorithm in an online fashion. Note, that with the exception of c, all quantities in the FAOL algorithm can be computed in an online fashion. We will solve this issue by estimating c from part of the data stream. First, note that if we exchange the matrix A_k in the FAOL algorithm with the matrix $\tilde{A}_k := \frac{1}{|S_k|} \sum_{n \in S_k} y_n y_n^\star$, where $S_k := \{n \in [N] \colon k \in J_n\}$, we do not alter the algorithm. The numbers c_k can be computed from the gradients g_k and the matrix A_k via $c_k = g_k A_k g_k^\star = \frac{1}{|S_k|} \sum_{n \in S_k} |\langle g_k, y_n \rangle|^2$. If we want to estimate c_k, we need both, a good estimate of the gradients g_k, and a good estimate of $A_k g_k^\star$. We do this by splitting the datastream into two parts. The first part of the datastream is used to get a good estimate of the normalised gradient g_k. The second part is used to refine g_k as well as to estimate $\frac{1}{|S_k|} \sum_{n \in S_k} |\langle g_k, y_n \rangle|^2$. The parameter ε specifies the portion of the datastream used to estimate c_k and refine g_k. We summarise all our considerations leading to the algorithm, referred to as Sequential Analysis Operator Learning (SAOL), in Table 5.2.

Concerning the computation and storage costs, we see that, as for FAOL, the computationally expensive task is determining the sets J_n. This has to be done for each of our N sample vectors via determining the ℓ smallest entries in the product Γy_n. The matrix-vector product takes $(2d - 1)K$ operations and searching can be done in one run through the K resulting entries, yielding an overall runtime complexity of $\mathcal{O}(dKN)$. However, compared to FAOL, the sequential version has much lower memory requirements on the order of $\mathcal{O}(dK)$, corresponding to the gradients g_k and the current version of the operator Γ. In order to see how the two algorithms perform, we will next conduct some experiments both on synthetic and image data.

5.2.2 Experiments on synthetic data

In the first set of experiments, we use synthetic data generated from a given (target) analysis operator Ω. A data vector y is generated by choosing a vector z from the unit sphere and a random subset Λ of ℓ analysers. We then project z onto the orthogonal complement of the chosen analysers, contaminate it with Gaussian noise and normalise it, see Table 5.3. The cosparse signals generated according to this model are very isotropic and thus do not exhibit the pathologies we described in the counterexample at the beginning of the section.

```
SAOL($\Gamma, \ell, Y, \varepsilon$) - (one iteration)

Initialize $I_k, C_k, c_k = 0$ and $g_k = 0$ for $k \in [K]$.

• For all $n \in [N]$:
    – Find $J_n = \arg\min_{|J|=\ell} \|\Gamma_J y_n\|_2^2$.
    – For all $k \in J_n$ update $I_k \to I_k + 1$ and
          $$g_k \to \frac{I_k - 1}{I_k} g_k + \frac{1}{I_k} \langle \gamma_k, y_n \rangle y_n^\star.$$
    – If $n > (1-\varepsilon)N$ and $k \in J_n$ update $C_k \to C_k + 1$ and
          $$c_k \to \frac{C_k - 1}{C_k} c_k + \frac{1}{C_k} |\langle g_k, y_n \rangle|^2.$$
• For all $k \in [K]$:
    – Set $a = \langle \gamma_k, g_k \rangle$, $b = \langle g_k, g_k \rangle$ and $c = c_k$,
    – Set $\alpha_k = \frac{ab-c+\sqrt{(c-ab)^2-4(b^2-ac)(a^2-b)}}{2(b^2-ac)}$.
    – Set $\bar{\gamma}_k = (\gamma_k - \alpha_k g_k)$.
• Output $\bar{\Gamma} = (\frac{\bar{\gamma}_1}{\|\bar{\gamma}_1\|_2}, \dots, \frac{\bar{\gamma}_K}{\|\bar{\gamma}_K\|_2})^\star$.
```

Table 5.2: The SAOL algorithm

Signal model(Ω, ℓ, ρ)
Input:

- $\Omega \in \mathbb{R}^{K \times d}$ - target analysis Operator,
- ℓ - cosparsity level of the signals w.r.t. Ω,
- ρ - noise level.

Generation of the signals is done in the following way:

- Draw $z \sim \mathcal{N}(0, I_d), r \sim \mathcal{N}(0, \rho^2 I_d)$ and $\Lambda \sim \mathcal{U}(\binom{[K]}{\ell})$.
- Set

$$y = \frac{(\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda)z + r}{\|(\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda)z + r\|}. \tag{5.5}$$

The matrix $(\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda)$ is a projector onto the space of all cosparse signals with cosupport Λ, so generating our signals in this way makes sure that they are (up to some noise) cosparse.

Table 5.3: Signal model

Target operator: As target operator for our experiments with synthetic data, we used a random operator of size 128×64 consisting of rows drawn i.i.d. from the unit sphere $\mathbb{S}^{63}$.
Training signals: Unless specified otherwise, in each iteration of the algorithm, we use $2^{17} = 131072$ signals drawn according to the signal model in Table 5.3 with cosparsity level $\ell = 55$ and noiselevel $\rho = 0$ for noisefree resp. $\rho = 0.2/\sqrt{d}$ for noisy data. We also conducted experiments with cosparsity level $\ell = 60$, but the results are virtually indistinguishable from the results for $\ell = 55$, so we chose not to present them here. We refer the interested reader to the AOL toolbox on the homepage of the authors[2], which can be used to reproduce the experiments.
Initialisation & setup: We use both a closeby and a random initialisation of the correct size. For the closeby initialisation, we mix the target operator 1:1 with a random operator and normalise the rows, that is, our initialisation operator is given by $\Gamma_0 = D_n(\Omega + R)$, where R is a $K \times d$ matrix with rows drawn uniformly at random from the unit sphere $\mathbb{S}^{d-1}$ and D_n is a diagonal matrix ensuring that the rows of Γ_0 are normalised. For the random initialisation we simply set $\Gamma_0 = R$. The correct cosparsity level ℓ is given to the algorithm and the results have been averaged over 5 runs with different initialisations.

[2] All experiments can be reproduced using the AOL Matlab toolbox available at `https://www.uibk.ac.at/mathematik/personal/schnass/code/aol.zip`.

Recovery threshold: We use the convention that an analyser ω_k is recovered if $\max_j |\langle \omega_k, \gamma_j \rangle| \geq 0.99$.

Our first experiment is designed to determine the proportion of signals $L = \varepsilon N$ that SAOL should use to estimate the values of c_k. We make an exploratory run for FAOL and SAOL with several choices of ε, using 16384 noiseless, 60-cosparse signals per iteration and a random initialisation.

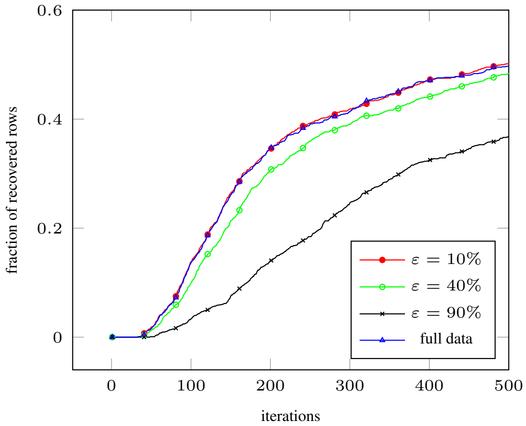

Figure 5.1: Recovery rates of an exploratory run using FAOL and SAOL with different epsilons for the recovery of a random 128×64 operator using 16384 samples in each iteration.

The recovery rates in Figure 5.1 indicate that in order for the SAOL algorithm to achieve the best possible performance, ε should be chosen small, meaning one should first get a good estimate of the gradients g_k. This allocation of resources also seems natural since for small ε a large portion of the data is invested into estimating the d-dimensional vectors g_k, while only a small portion is used to subsequently estimate the numbers c_k. Based on these findings we from now on set $\varepsilon = 10\%$ for the SAOL-algorithm.

In the next experiment we compare the recovery rates of FAOL, SAOL and Analysis-SimCO [DWD$^+$16] from random and closeby initialisations in a noiseless setting.

The first good news of the results, plotted in Figure 5.2, is that the sequential algorithm SAOL with estimated stepsize performs as well as the one with explicitly calculated optimal stepsize. We also see that with a closeby initialisation both algorithms recover the target operator (almost) perfectly for both cosparsity levels, which indicates that locally our algorithms perform as expected. With a random initialisation the algorithms tend to saturate well below full recovery. This is not surprising, as the nonconvex optimization we perform depends heavily on the initialisation. In case of the closeby initialisation, we set each row of the starting operator near the desired row of the target operator. In contrast, for the random initialisation it is very likely that two rows of the initialised operator lie close to the same row of the target operator. Our algorithms then tend

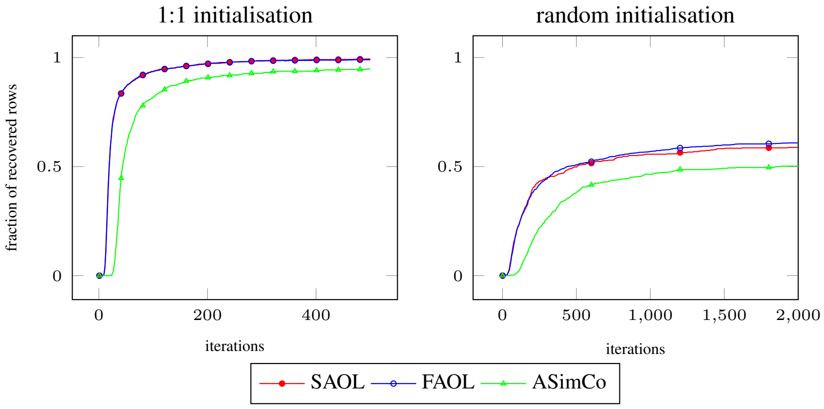

Figure 5.2: Recovery rates of SAOL, FAOL and ASimCo from signals with various cosparsity levels ℓ in a noiseless setting, using closeby (left) and random (right) initialisations for cosparsity level $\ell = 55$.

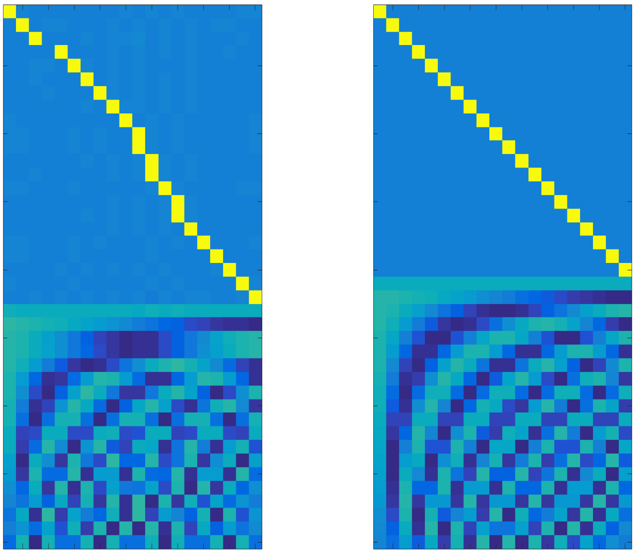

Figure 5.3: Operator learned with FAOL from a random initialisation (left) vs the original Dirac-DCT operator (right). The rows of the learned operator have been reordered and the signs have been matched with the original operator for easier comparison. For the learning 300 iterations with 8192 noiseless 12-cosparse signals, constructed according to the model in Table 5.3 were used.

to find the nearest row of the target operator and thus we get multiple recovery of the same row. As we have prescribed a fixed number of rows, another row must be left out, which leads to the observed stagnation of the recovery rates and means that we are trapped in a local minimum of our target function. Figure 5.3 illustrates this effect for the Dirac-DCT operator in $\mathbb{R}^{40\times 20}$.

Since the phenomenon of recovering duplicates is not only as old as analysis operator learning but as old as dictionary learning, [AEB06], there is also a known solution to the problem, which is the replacement of coherent analysers or atoms.

5.2.3 Replacement

A straightforward way to avoid learning analysis operators with duplicate rows is to check after each iteration, whether two analysers of the current iterate Γ are very coherent. Under the assumption that the coherence of the target operator $\mu(\Omega) = \max_{i\neq j\in[K]} |\langle\omega_i, \omega_j\rangle|$ is smaller than some threshold $\mu(\Omega) \leq \mu_0$, we know that two rows of γ_i, γ_j are likely to converge to the same target analyser, whenever we have $|\langle\gamma_i, \gamma_j\rangle| > \mu_0$.

In this case, we perform the following decorrelation procedure. During the algorithm, we monitor the activation of the individual rows of the operator, that is, if the $k-$th row of the operator is used for a set J_n, we increment a counter v_k. If now two rows γ_i and γ_j have overlap larger than μ_0, we compare the numbers v_i and v_j and keep the row with larger counter. Without loss of generality suppose $v_i > v_j$. We then subtract the component in direction of γ_i from γ_j, namely $\tilde{\gamma}_j = \gamma_j - \langle\gamma_i, \gamma_j\rangle\gamma_i$ and renormalise.

This decorrelation is different from the one that is performed in [DWD$^+$16], but has the merit that correct rows that have already been found do not get discarded or perturbed. This is especially useful in the case we consider most likely, where one row already has large overlap with a row of the target operator and another row slowly converges towards the same row. Then our decorrelation procedure simply subtracts the component pointing in this direction.

Since, unlike dictionaries, analysis operators can be quite coherent and still perform very well, for real data it is recommendable to be conservative and set the coherence threshold μ_0 rather high.

Figure 5.4 shows the recovery results of our algorithm with the added replacement step for $\mu_0 = 0.8$, when using a random initialisation and the same settings as described at the beginning of the section.

We see that in the noiseless case, after 10000 iterations almost 90% of the signals have been recovered. If we introduce a small amount of noise, however, significantly fewer rows are recovered. To avoid repetition we postpone a thorough comparison of FAOL/SAOL to ASimCo on synthetic data to Section 5.3.3 after the introduction of our other two algorithms in Section 5.3,

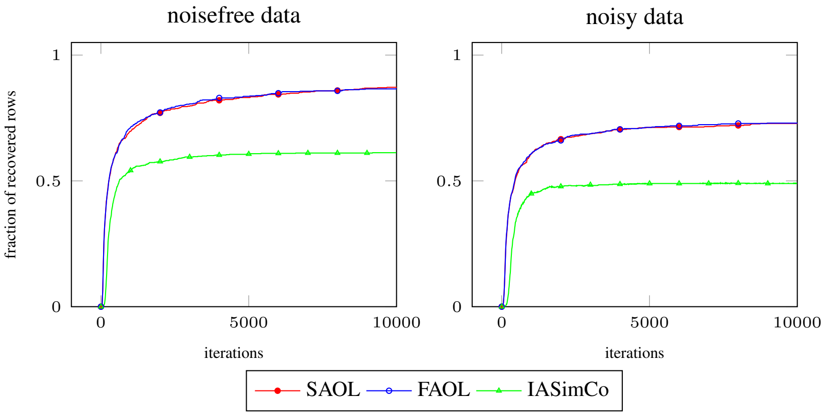

Figure 5.4: Recovery rates of SAOL and FAOL with replacement from signals with cosparsity level $\ell = 55$ in a noiseless (left) and a noisy setting (right), using a random initialisation.

however we can already observe now that both SAOL and FAOL perform better than IASimCO. Next we take a look at how the optimal stepsize affects learning on image data.

5.2.4 Experiments on image data

To get an indication how our algorithms perform on real data, we will them to learn a quadratic analysis operator on all 8×8 patches of the 256×256 Shepp Logan phantom, cf. Figure 5.11. We initialise the analysis operator $\Gamma \in \mathbb{R}^{64 \times 64}$ randomly as for the synthetic data and set the cosparsity level $\ell = 57$, the parameter $\varepsilon = 10\%$ and the replacement threshold $\mu_0 = 0.99$. For each iteration we choose 16384 out of the available 62001 patches uniformly at random as training signals. Since we do not have a reference operator for comparison this time, we compare the value of target function after each iteration, as plotted in Figure 5.5. We can see that the target function is only decreased very slowly by all algorithms, where ASimCO saturates at a seemingly suboptimal value, which is also illustrated by the recovered operator shown in Figure 5.5.

As we choose the optimal stepsize for FAOL, we further cannot hope to increase convergence speed significantly using an explicit descent algorithm.

Still, if we look at the learned operator, we can see the merit of our method. After 100000 iterations, the learned operator seems to consist of pooled edge detectors, which are known to cosparsify piecewise constant grayscale images. Note also that the $d \times d$ analysis operator is naturally very different from any $d \times d$ dictionary we could have learned with corresponding sparsity level $S = d - \ell$, see e.g [Sch16]. This is due to the fact that image patches are not isotropic, but have their energy concentrated in the low frequency ranges. So while both the

$d \times d$ dictionary and the analysis operator will not have (stable) full rank, the dictionary atoms will tend to be in the low frequency ranges, and the analysers will - as can be seen - tend to be in the high frequency ranges.

We also want to mention that for image data the replacement strategy for $\mu_0 = 0.99$ is hardly ever activated. Lowering the threshold results in continuous replacement and refinding of the same analysers. This phenomenon is again explained by the lack of isotropy and the shift invariant structure of the patch data, for which translated and thus coherent edge detectors, as seen in Figure 5.5, naturally provide good cosparsity.

Encouraged by the learned operator we will explore in the next section how to stabilise the algorithm and accelerate its convergence on image data.

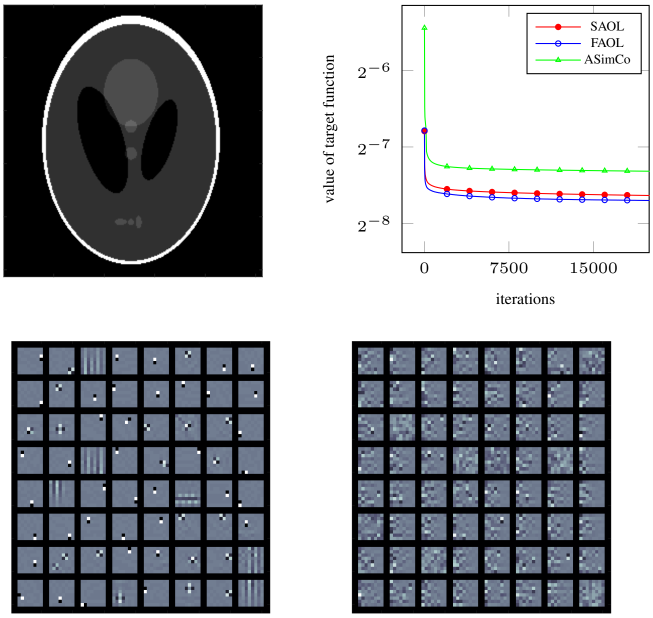

Figure 5.5: Shepp Logan phantom (top left), value of the target function for both algorithms (top right), operator obtained by FAOL (bottom left) and by ASimCo (bottom right) after 100000(!) iterations.

5.3 Two implicit operator learning algorithms - IAOL and SVAOL

Due to the small optimal stepsize that has to be chosen on real data and the resulting slow convergence, we need to rethink our approach and enforce stability of the algorithm even with larger stepsizes.

5.3.1 The IAOL algorithm

In standard gradient descent, for each row of Γ, we have the iteration

$$\bar{\gamma}_k = \gamma_k - \alpha \nabla f_N(\Gamma)_k. \tag{5.6}$$

Rewriting yields

$$\frac{\bar{\gamma}_k - \gamma_k}{\alpha} = -\nabla f_N(\Gamma)_k, \tag{5.7}$$

which can be interpreted as an explicit Euler step for the system of ordinary differential equations

$$\dot{\gamma}_k = -\nabla f_N(\Gamma)_k, \ k \in [K]. \tag{5.8}$$

The explicit Euler scheme is the simplest integration scheme for ordinary differential equations and known to have a very limited region of convergence with respect to the stepsize. In our case, this means that we have to choose extremely small values for the descent parameter α in order to achieve convergence.

The tried and tested strategy to overcome stability issues when numerically solving differential equations is to use an implicit scheme for the integration, [HNW93, HW10]. We will use this as an inspiration to obtain a more stable learning algorithm.

We briefly sketch the ideas behind an implicit integration scheme. Suppose we want to solve the differential equation $\dot{x} = f(x)$. If we discretise $x(t)$ and approximate the derivative by $\dot{x}(t_n) \approx \frac{x(t_n)-x(t_{n-1})}{t_n - t_{n-1}}$, we have to choose whether we use the approximation $\dot{x}(t_n) = f(x(t_n))$ or $\dot{x}(t_n) = f(x(t_{n-1}))$. Choosing $f(x(t_{n-1}))$ yields the explicit Euler scheme, which in our setting corresponds to the FAOL algorithm. If we choose $f(x(t_n))$ we obtain the implicit Euler scheme and need to solve

$$\frac{x(t_n) - x(t_{n-1})}{t_n - t_{n-1}} = f(x(t_n)). \tag{5.9}$$

```
IAOL($\Gamma, \ell, Y, \alpha$) - (one iteration)

  • For all n:
      – Find $J_n = \arg\min_{|J|=\ell} \|\Gamma_J y_n\|_2^2$.
      – For all $k \in J_n$ update $A_k = A_k + y_n y_n^\star$.
  • For all $k \in [K]$ set $\bar{\gamma}_k = \gamma_k \left(\mathbb{1} + \alpha A_k\right)^{-1}$.
  • Output $\bar{\Gamma} = (\frac{\bar{\gamma}_1}{\|\bar{\gamma}_1\|_2}, \dots, \frac{\bar{\gamma}_K}{\|\bar{\gamma}_K\|_2})^\star$.
```

Table 5.4: The IAOL algorithm

If $f(x) = Ax$ is linear, this leads to the recursion

$$x(t_n) = (\mathbb{1} - (t_n - t_{n-1})A)^{-1} x(t_{n-1}), \tag{5.10}$$

and in each step we need to solve a system of linear equations. This makes implicit integration schemes inherently more expensive than explicit schemes. However, in return we get additional stability with respect to the possible stepsizes. If f is a nonlinear function, the inversion is more difficult and can often only be approximated for example via a Newton method.
Mapping everything to our setting, we observe that the gradient $\nabla f_N(\Gamma)$ is nonlinear because the sets J_n depend on Γ. Still, due to the special structure of the gradient $\nabla f_N(\Gamma)$, it has a simple linearisation, $\nabla f_N(\Gamma)_k = 2\gamma_k \sum_{n\colon k \in J_n} y_n y_n^\star$. We can now use the current iterate of Γ to compute the matrix $A_k(\Gamma) := \sum_{n\colon k \in J_n} y_n y_n^\star$ and to linearise the equation. For our operator learning problem, we get the following linearised variant of the implicit Euler scheme

$$\frac{\bar{\gamma}_k - \gamma_k}{\alpha} = -\bar{\gamma}_k A_k(\Gamma), \tag{5.11}$$

leading to the recursion

$$\bar{\gamma}_k = \gamma_k (\mathbb{1} + \alpha A_k(\Gamma))^{-1} \tag{5.12}$$

Due to the unconditional stability of the implicit Euler scheme, [HW10], we can take α considerably larger than in case of FAOL or SAOL. We only need to make sure that one step of the algorithm does not take us too close to zero, which is a stable attractor of the unconstrained system. In order to stay within the manifold $\mathcal{A}$, we again have to renormalise after each step. The final algorithm is summarised in Table 5.4.

Let us take a short look at the computational complexity of the implicit algorithm and the price we have to pay for increased stability. As in the previous section, we need to compute all products of the vectors y_n with the current iterate Γ, costing $\mathcal{O}(NKd)$. Further, in each step we need to solve K linear systems of size $d \times d$, amounting to an additional cost of $\mathcal{O}(Kd^2)$. So, altogether for one step, we arrive at $\mathcal{O}(NKd + Kd^2) = \mathcal{O}(NKd)$. However, in contrast to FAOL, the IAOL algorithm cannot be made sequential, unless we are willing to store the K matrices A_k in each step. This amounts to an additional spatial complexity of $\mathcal{O}(Kd^2)$, and only pays off for $N > Kd$, since the storage cost of the data matrix is $\mathcal{O}(Nd)$. Note that in contrast to FAOL, the explicit calculation of the matrices A_k is necessary, since we need to solve a system of linear equations. As for the FAOL algorithm, we can guarantee decrease or preservation of the target function by the IAOL algorithm.

Theorem 5.3.1. *The IAOL algorithm decreases or preserves the value of the target function in each step regardless of the choice of $\alpha > 0$. Preservation rather than decrease of the objective function can only occur if all rows γ_k of the current iterate Γ are eigenvectors of the matrices $A_k(\Gamma)$.*

Proof. In order to simplify notation, we drop the indices and write Γ for the current iterate. As we will do the computations only for a fixed row γ_k of Γ, we denote it by γ. We further write A for the matrix corresponding to γ and $\bar{\gamma}$ for the next iterate. We want to show

$$\bar{\gamma} A \bar{\gamma}^\star - \gamma A \gamma^\star = \frac{\gamma A (\mathbb{1} + \alpha A)^{-2} \gamma^\star}{\gamma (\mathbb{1} + \alpha A)^{-2} \gamma^\star} - \gamma A \gamma^\star \leq 0,$$

which is implied by

$$\gamma (A(\mathbb{1} + \alpha A)^{-2} - A \gamma^\star \gamma (\mathbb{1} + \alpha A)^{-2}) \gamma^\star = \gamma (A(\mathbb{1} - \gamma^\star \gamma)(\mathbb{1} + \alpha A)^{-2}) \gamma^\star \leq 0,$$

as the denominator is positive. We now use the eigendecomposition of the symmetric, positive semidefinite matrix A, that is $A = \sum_i \lambda_i u_i u_i^\star$, where $\lambda_i \geq 0$ for all i and $(u_i)_{i \in [d]}$ is an orthonormal basis.

Inserting this, a short computation shows that

$$\gamma (A(\mathbb{1} - \gamma^\star \gamma)(\mathbb{1} + \alpha A)^{-2}) \gamma^\star = \sum_i \sum_{l \neq i} \lambda_i \underbrace{\frac{\alpha (2 + \alpha(\lambda_i + \lambda_l)) |\langle \gamma, u_i \rangle|^2 |\langle \gamma, u_l \rangle|^2}{(1 + \alpha \lambda_i)^2 (1 + \alpha \lambda_l)^2}}_{=: a_{il}} (\lambda_l - \lambda_i).$$

Note that $a_{il} = a_{li} \geq 0$. Further, we can drop the condition $l \neq i$ in the sums above, as the term corresponding to the case $i = l$ is zero.

In order to show that the sum $S_1 := \sum_{i,l} \lambda_i a_{il} (\lambda_l - \lambda_i)$ is never positive, define a second sum $S_2 := \sum_{i,l} \lambda_l a_{il}(\lambda_l - \lambda_i)$. Then, by antisymmetry, we have that $S_1 + S_2 = \sum_{i,l}(\lambda_i + \lambda_l)a_{il}(\lambda_l - \lambda_i) = 0$. Further, $S_2 - S_1 = \sum_{i,l} a_{il}(\lambda_l - \lambda_i)^2 \geq 0$, from which follows that $S_1 \leq 0$. The whole discussion is independent of $\alpha > 0$, so any viable choice of α results in a decrease of the objective function.

Assume now that $\bar{\gamma} A \bar{\gamma}^\star - \gamma A \gamma^\star = 0$. Then also $\gamma(A(\mathbb{1} - \gamma^\star\gamma)(\mathbb{1} + \alpha A)^{-2})\gamma^\star = S_1 = 0$. This in turn implies that $S_2 = 0$ and $S_2 - S_1 = 0$. As every term in $\sum_{i,l} a_{il}(\lambda_l - \lambda_i)^2$ is positive or zero, we have that for all $i \neq l$ also $a_{il}(\lambda_l - \lambda_i)^2$ must be zero. If all eigenvalues of the matrix A are distinct this implies that $a_{il} = 0$ for all $i \neq l$. This in turn implies that for all $i \neq l$ the product $|\langle \gamma, u_i \rangle|^2 |\langle \gamma, u_l \rangle|^2 = 0$, so either the overlap of γ_k with u_i or u_l is zero. But this means that γ_k must be equal to one of the eigenvectors. If not all eigenvalues of the matrix A are distinct, then the previous discussion still holds for the eigenvalues which are distinct. Assume that $i \neq l$ and $\lambda_i = \lambda_j$. Then $a_{il}(\lambda_l - \lambda_i)^2 = 0$ regardless of the value of a_{il}, so if $\gamma \in \operatorname{span}\{u_i, u_l\}$, we still have that $S_2 - S_1 = 0$. This shows that in all cases, where the target function does not decrease, γ needs to be an eigenvector of A. □

Note that the IAOL algorithm essentially performs a single step of an inverse iteration to compute the eigenvectors corresponding to the smallest eigenvalues of the matrices A_k. We will use this fact in the next section to introduce our last algorithm to learn analysis operators.

5.3.2 The SVAOL algorithm

Revisiting condition (5.4) suggests another algorithm for learning analysis operators. The step-size choice essentially amounts to

$$\bar{\gamma}_k = \operatorname*{arg\,min}_{v \in \mathcal{K}_2(\gamma_k, A_k) \cap (\mathbb{S}^{d-1})^\star} \frac{v A_k v^\star}{v v^\star},$$

where $\mathcal{K}_2(\gamma_k, A_k) = \operatorname{span}\{\gamma_k, \gamma_k A_k\}$, the Krylov space of order 2. Removing the restriction that $\bar{\gamma}_k$ must lie in the Krylov space $\mathcal{K}_2(\gamma_k, A_k)$ yields the update step

$$\bar{\gamma}_k = \operatorname*{arg\,min}_{v \in (\mathbb{S}^{d-1})^\star} v A_k v^\star,$$

which means that $\bar{\gamma}_k$ is the eigenvector corresponding to the smallest eigenvalue of the matrix A_k. The resulting algorithm, called SVAOL, is summarised in Table 5.5.

The obtained SVAOL algorithm bears close resemblance to the 'Sequential Minimal Eigenvalues' algorithm devised in [OEBP11]. However, a key difference is that the computation of the rows of the target operator is not done sequentially in the SVAOL algorithm. Furthermore, the following theorem concerning decrease of the target function can be established.

```
**SVAOL($\Gamma, \ell, Y$) - (one iteration)**
For all $k \in [K]$, set $A_k = 0$.
  • For all n:
      – Find $J_n = \arg\min_{|J|=\ell} \|\Gamma_J y_n\|_2^2$.
      – For all $k \in [K]$ update $A_k = A_k + y_n y_n^\star$ if $k \in J_n$.
  • For all $k \in [K]$, set $\bar{\gamma}_k = \arg\min_{v \in (\mathbb{S}^{d-1})^\star} v A_k v^\star$.
  • Output $\bar{\Gamma} = (\bar{\gamma}_1, \dots, \bar{\gamma}_K)^\star$.
```

Table 5.5: The SVAOL algorithm

Theorem 5.3.2. *The SVAOL algorithm decreases or preserves the value of the target function in each step. The only case when it preserves the value of the target function is when the rows γ_k are already eigenvectors corresponding to the smallest eigenvalues of the matrices A_k.*

The results for SVAOL given in Theorem 5.3.2 improve the results obtained for IAOL in Theorem 5.3.1. Now the decrease of the target function can be guaranteed if not all rows of the current iterate Γ are already the eigenvectors corresponding to the smallest eigenvalues of the matrices A_k.[3]

Proof. To show this, denote by $(A_k)_{k\in[K]}$ and $(\bar{A}_k)_{k\in[K]}$ the matrices defined in Table 5.5 for the operators Γ and $\bar{\Gamma}$, respectively. Further denote by $(\sigma_k)_{k\in[K]}$ and $(\bar{\sigma}_k)_{k\in[K]}$ their smallest singular values.

Then

$$
\begin{aligned}
f(\Gamma) &= \sum_{k=1}^{K} \gamma_k A_k \gamma_k^\star \geq \sum_{k=1}^{K} \sigma_k = \sum_{k=1}^{K} \bar{\gamma}_k A_k \bar{\gamma}_k^\star = \\
&= \sum_{k=1}^{K} \sum_{n\colon k\in J_n} |\langle \bar{\gamma}_k, y_n \rangle|^2 \geq \sum_{k=1}^{K} \sum_{n\colon k\in \bar{J}_n} |\langle \bar{\gamma}_k, y_n \rangle|^2 = \\
&= \sum_{k=1}^{K} \bar{\gamma}_k \bar{A}_k \bar{\gamma}_k^\star = f(\bar{\Gamma})
\end{aligned}
$$

[3] This means that if the target function is differentiable in Γ and cannot be decreased by the SVAOL algorithm, we have already arrived at a local minimum. As we have stated previously, however, the target is not differentiable everywhere and thus this cannot be used to derive a local optimality result.

due to the definition of the sets J_n and $\bar{J}_n$. The first inequality is strict, except in the case when γ_k are already eigenvectors of A_k corresponding to the smallest eigenvalues. □

Finding the eigenvectors corresponding to the smallest eigenvalues of the matrices A_k is indeed the desired outcome, which can be seen as follows. First, note that the matrices $A_k(\Gamma)$ are (up to a constant) empirical estimators of the matrices $\mathbb{A}_k(\Gamma) := \mathbb{E} yy^\star \chi_{\{y:\, k \in J^\Gamma(y)\}}$, where $J^\Gamma(y) = \arg\min_{|J|=\ell} \|\Gamma_J y\|_2^2$. The rows ω_k of the target operator Ω are (in the noisefree setting) always eigenvectors to the eigenvalue zero for the matrix $\mathbb{A}_k(\Omega)$, since according to the signal model given in Table 5.3, we have

$$\begin{aligned}
\mathbb{A}_k(\Omega) &= \mathbb{E} yy^\star \chi_{\{y:\, k \in J^\Omega(y)\}} \\
&= \mathbb{E}(\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) zz^\star (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) \chi_{\{(\Lambda,z):\, k \in J_{(\Lambda,z)}\}} \\
&= \mathbb{E}_\Lambda (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) \mathbb{E}_z zz^\star (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) \chi_{\{\Lambda:\, k \in \Lambda\}} \\
&= \mathbb{E}_\Lambda (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) \chi_{\{\Lambda:\, k \in \Lambda\}} \\
&= \binom{K}{\ell}^{-1} \sum_{\Lambda:\, k \in \Lambda} (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda),
\end{aligned}$$

where $J_{(\Lambda,z)} = \arg\min_{|J|=\ell} \|\Omega_J (\mathbb{1} - \Omega_\Lambda^\dagger \Omega_\Lambda) z\|_2^2$.

Multiplying this matrix with ω_k yields zero, as k always lies in Λ and so every term in the sum maps ω_k to zero.

The Analysis K-SVD algorithm [RPE13] takes a similar approach as the SVAOL algorithm. At first, cosupport sets are estimated and then the singular vector to the smallest singular value of the resulting data matrix is computed. The notable difference, however, is how the cosupport set is estimated. We use a hard thresholding approach, whereas for Analysis K-SVD an analysis sparse-coding step is employed, which uses significantly more computational resources.

We see that the computation and storage complexity for the first steps (i.e. setting up and storing the matrices A_k) of the SVAOL algorithm are the same as for IAOL. This means that the spatial complexity of SVAOL is $\mathcal{O}(Kd^2)$. For the runtime complexity, which is $\mathcal{O}(NKd)$ for the setup of the matrices, we need to include the cost of computing the K smallest singular values of the matrices A_k. This can be done for example using the SSVII algorithm presented in [SS03], in which for each iteration a system of linear equations of size $d+1$ has to be solved. We observed that typically 10 iterations of the algorithm are sufficient, so the computational complexity for this step is 10 times higher than for IAOL.

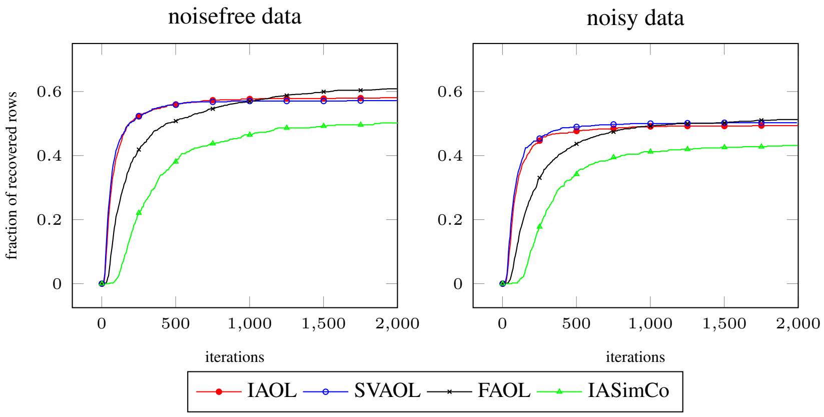

Figure 5.6: Recovery rates of IAOL and SVAOL in a noisefree (left) and a noisy (right) setting compared to FAOL and ASimCO using cosparsity level $\ell = 55$ using a random initialisation.

5.3.3 Experiments on synthetic data

As for the explicit algorithms presented above, we first try our new algorithm on synthetic data. For this, we again learn an operator from data generated from a random operator with normalised rows in $\mathbb{R}^{128\times 64}$. The setup is the same as in Section 5.2.2 and the results are shown in Figure 5.6. We use a large stepsize $\alpha = 100$ in order to achieve fast convergence.

Note that IAOL and SVAOL saturate faster than FAOL, cf. Figure 5.2. However, IAOL and SVAOL without replacement recover slightly fewer rows as FAOL, which is probably a result of the faster convergence speed.

Finally, since the implicit algorithms per se, like FAOL, do not penalise the recovery of two identical rows, cf. Figure 5.3, we again need to use the replacement strategy introduced in Section 5.2.3.

The simulation results, using replacement with $\mu_0 = 0.8$ and the usual setup are shown in Figure 5.7. We see that IAOL and SVAOL come closer to full recovery than their explicit counterpart FAOL within the considered 10000 iterations. Again, for noisy data the algorithms saturate well below full recovery.

5.3.4 Experiments on image data

Finally, we want to see how the stabilised algorithms perform on real data. We use the same image (Shepp Logan) and setup as in Section 5.2.4 to learn a square analysis operator for 8×8 patches, cf. Figure 5.5. We will not use the SAOL algorithm in the simulations from now on, as the execution time in Matlab is considerably higher due to the required for-loops. However, as

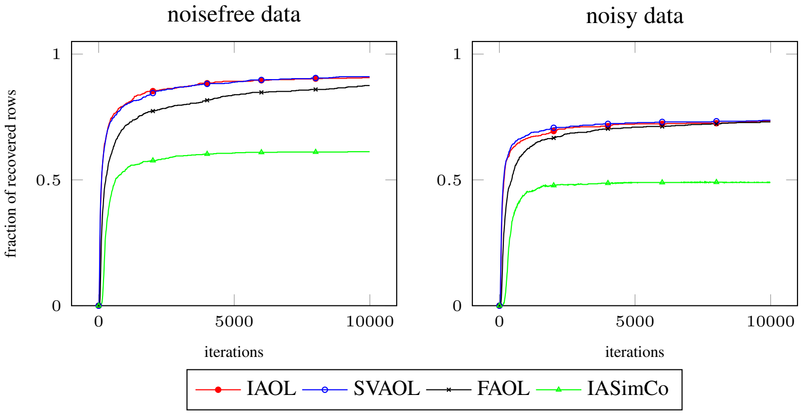

Figure 5.7: Recovery rates of IAOL, SVAOL, FAOL and IASimCo with replacement from signals with cosparsity level $\ell = 55$ in a noiseless (left) and a noisy (right) setting, using a random initialisation.

we have seen, it performs mostly like the FAOL algorithm.

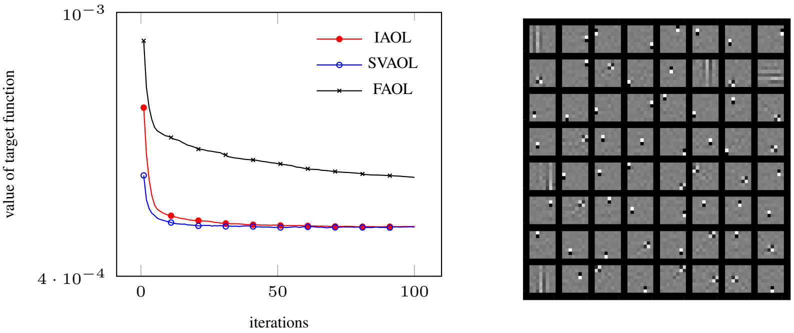

Figure 5.8: Decay of the target function using FAOL, IAOL and SVAOL for the Shepp Logan phantom (left) and the operator recovered by IAOL after 100(!) iterations (right).

As can be seen in Figure 5.8, the training is much faster now, because the stepsize does not have to be chosen as small as in the previous section. The decrease in the objective function is very fast compared to FAOL, and we see that already after a few iterations the algorithm stabilises and we, as expected, obtain combinations of discrete gradients as anlysers. As for FAOL we observe that the replacement strategy for $\mu_0 = 0.99$ is hardly ever activated and that lowering the threshold results in finding and replacing the same translated edge detectors.

In the remainder of this section, we will investigate the time complexity of the presented algorithms numerically. Naturally, the explicit algorithms use significantly fewer computational resources per iteration. We compare the average calculation times per iteration on a 3.1 GHz In-

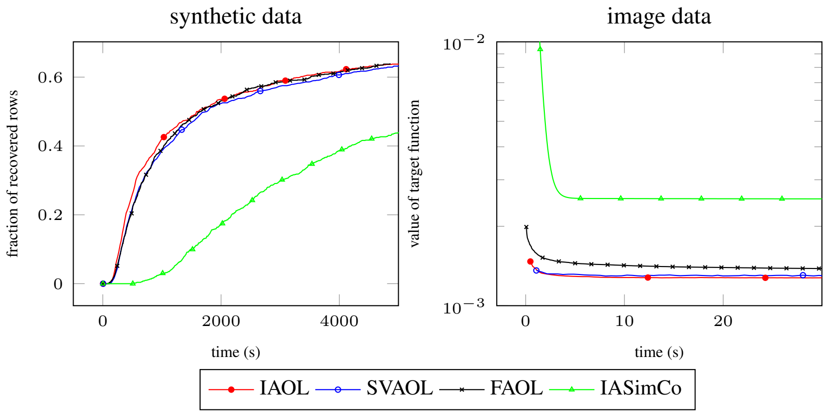

Figure 5.9: Recovery rates of FAOL, IAOL, SVAOL and IASimCo from 55-cosparse signals in a noisy setting (left). Decay of the target function using IASimCO, FAOL, IAOL and SVAOL to learn a 128×64 operator for the Shepp Logan image (right). The figures depict the execution time of the algorithm on the x-axis. SAOL is omitted, as the execution time is not comparable due to the Matlab implementation.

tel Core i7 Processor. For IASimCO the out-of-the-box version on the homepage of the authors of [DWD$^+$16] is used.

Figure 5.9 (left) shows the recovery rates of the four algorithms plotted against the cumulative execution time. We can see that on synthetic data, the IAOL and SVAOL algorithms show a similar performance to FAOL, followed by IASimCo. Using FAOL as a baseline, we see that one iteration of ASimCo takes about twice as long, one iteration of IAOL takes about four times as long and one iteration of SVAOL takes 5-6 times longer, but this significantly depends on the number of iterations of the inverse iteration to find the smallest singular value.

In the experiment on image data, we learn an overcomplete operator with 128 rows from the 8×8 patches of the 256×256 (unnormalised) Shepp Logan phantom contamined with Gaussian noise with PSNR ≈ 20. We choose as cosparsity level $\ell = 120$, initialise randomly and in each iteration use 20000 randomly selected patches out of the available 62001. Since for image data our replacement strategy is hardly ever activated, we directly omit it to save computation time. IASimCo is again used in its out-of-the-box version. In Figure 5.9, one can clearly observe that the IAOL and SVAOL algorithms indeed minimise the target function in a fraction of the iterations necessary for the FAOL algorithm, which in turn is much faster than IASimCO. Already after 10 seconds, IAOL and SVAOL have essentially finished minimizing the objective function, whereas FAOL needs, as seen in the previous section, about 100000 iterations to get to approximately the same value of the objective function. IASimCo lags behind severely and, as indicated by the shape of the curve, saturates at a highly suboptimal value of

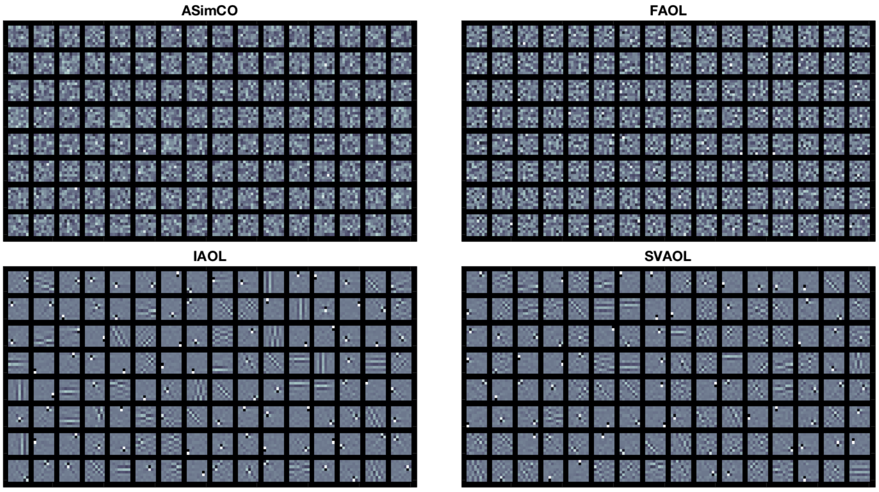

Figure 5.10: Analysis operators learned in 120 seconds with data drawn from the Shepp-Logan phantom contaminated with Gaussian noise, by ASimco (top left), FAOL (top right), IAOL (bottom left) and SVAOL (bottom right).

the target function. This fact can also be observed by looking at the learned analysis operators in Figure 5.10.

Encouraged by this good performance we will in the next section apply our algorithms to image denoising.

5.4 Image denoising

In this section we will compare the performance of analysis operators learned by the FAOL, IAOL and SVAOL algorithms presented in this paper in combination with Tikhonov regularisation for image denoising to the performance of operators learned by (I)ASimCO. For easy comparison we use the same setup as in [DWD$^+$16], where (I)ASimCo is compared to several other major algorithms for analysis operator learning, [RB13, YNGD13, HKD13, RPE13, EB14], and found to give the best performance.

Learning setup: We follow the setup for the Shepp-Logan image in the last section. Our training data consists of all 8×8 patches of one of the 256×256 images from Figure 5.11 corrupted with Gaussian white noise of standard deviation $\sigma = 12.8$ and $\sigma = 45$ leading to a PSNR of approximately 25dB and 15dB, respectively. The analysis operators of size 128×64 are initialised by drawing each row uniformly at random from the unit sphere, and then updated using in each step 20000 randomly selected patches of the available 62001 and a cosparsity level $\ell \in \{70, 80, 90, 100, 110, 120\}$. The same initialisation is used for all algorithms. For

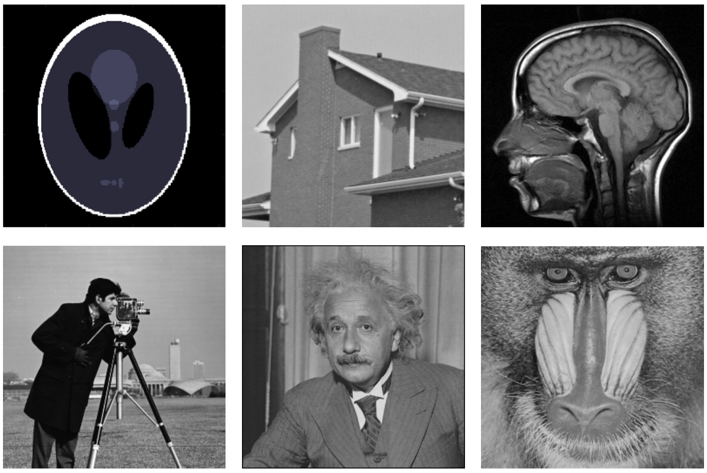

Figure 5.11: Images used for learning and denoising. Top: Shepp-Logan, House, MRI; Bottom: Cameraman, Einstein, Mandrill.

(I)ASimCo and FAOL we use 2000 and for IAOL and SVAOL 500 iterations. We perform the optimization without replacement for FAOL, IAOL and SVAOL.

Denoising setup: For the denoising step we use a standard approach via Tikhonov regularisation based on the learned analysis operator Γ, [EA06, EMR07]. For each noisy patch y we solve,

$$\hat{y} = \arg\min_{z} \lambda \|\Gamma z\|_1 + \|z - y\|_2 \tag{5.13}$$

for a regularisation parameter $\lambda \in \{0.002, 0.01, 0.05, 0.1, 0.3, 0.5\}$. We then reassemble the denoised patches $\hat{y}$ to the denoised image, by averaging each pixel in the full image over the denoised patches in which it is contained. To measure the quality of the reconstruction for each cosparsity level ℓ and regularisation parameter λ we average the PSNR of the denoised image over 5 different noise realisations and initialisations. Table 5.6 shows the PSNR for optimal choice of ℓ and λ for each of the algorithms. We can see that all five algorithms provide a comparable denoising performance, mostly staying with 0.1dB of each other. However, while FAOL, IAOL, SVAOL never lag more than 0.14dB behind, they do improve upon (I)ASimCo for more than 1dB twice. The denoising results for one of these cases, that is the Shepp-Logan phantom in the low-noise regime, are shown in Figure 5.12 .

After confirming that our algorithms indeed learn useful operators also on real data, we now turn to a discussion of our results.

	Algorithm	SL	Cam	Ein	MRI	Hou	Man
$\sigma = 12.8$	ASimCO	32.57	**30.36**	31.30	31.69	31.78	28.52
	IASimCO	32.49	30.32	31.19	**31.77**	31.60	28.28
	FAOL	**33.91**	30.33	**31.62**	31.70	**32.86**	**28.71**
	IAOL	33.39	30.24	31.54	31.71	32.66	28.64
	SVAOL	33.49	30.22	31.57	31.73	32.77	28.70
$\sigma = 45$	ASimCO	27.65	24.16	**25.87**	**25.66**	27.19	23.46
	IASimCO	27.50	23.87	25.78	25.52	27.12	23.33
	FAOL	28.33	24.17	25.82	**25.66**	**27.21**	23.50
	IAOL	**28.38**	24.18	25.83	25.65	27.19	**23.52**
	SVAOL	28.36	**24.20**	25.84	25.62	27.18	23.51

Table 5.6: Performance of FAOL, IAOL, SVAOL and (I)ASimCO for denoising for different pictures and noise levels σ.

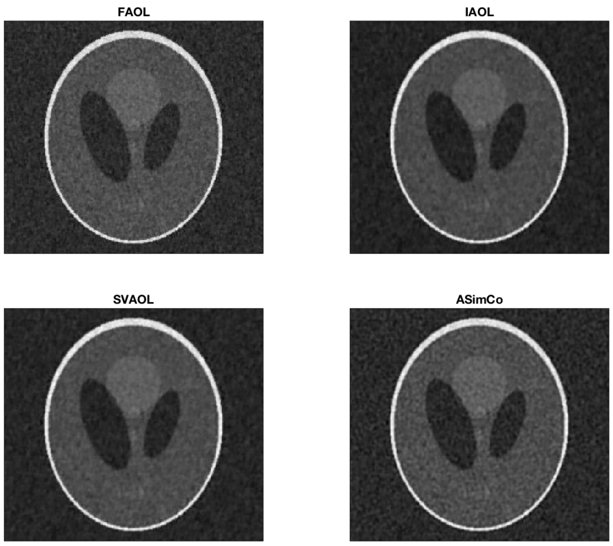

Figure 5.12: Denoised Shepp-Logan phantom using the optimal parameters for the various presented algorithms for noise level $\sigma = 12.8$.

5.5 Discussion

We have developed four algorithms for analysis operator learning based on projected stochastic gradient-like descent, SAOL, FAOL, IAOL and SVAOL. The algorithms perform better than the state-of-the-art algorithms (I)ASimCO, [DWD$^+$16], which are similarly gradient descent based and have slightly higher but comparable computational complexity per iteration, in terms of recovery rates resp. reduction of the objective function. Another advantage of SAOL is that it is sequential with a memory requirement corresponding to the size of the operator, $\mathcal{O}(dK)$. In contrast ASimCO either is non-sequential with a memory requirement of the order of the data matrix, $\mathcal{O}(dN)$, or in a sequential setting needs $O(LN)$ training samples corresponding to the L evaluations of the objective function necessary for the line search. IAOL and SVAOL, which are more stable than SAOL and FAOL, are sequential when accepting a memory requirement $\mathcal{O}(d^2K)$ and in a non-sequential setting have again memory requirement $\mathcal{O}(dN)$.

On synthetic data, the recovery of the target operator using the algorithms presented here is significantly faster than with (I)ASimCo. On real data the implicit algorithms IAOL and SVAOL minimize the objective function in a fraction of the time that is needed by the explicit algorithms FAOL and ASimCo. Considering image denoising via Tikhonov regularisation as application of analysis operator learning, we see that the operators presented in this paper give similar or better results as the (I)ASimCo operators in the considered denoising setups.

A Matlab toolbox to reproduce all the experiments reported in this paper can be found at `http://homepage.uibk.ac.at/~c7021041/code/AOL.zip`.

While the good performance of the developed algorithms certainly justified the effort, one of our main motivations for considering a projected gradient descent approach to analysis operator learning was to derive convergence results similar to those for dictionary learning, [Sch16]. However, even a local convergence analysis, turns out to be quite different and much more complicated than for dictionary learning. The main reason for this is that sparsity is more robust to perturbations than cosparsity. So for an S-sparse signal $y = \Phi_I x_I$ and a perturbed dictionary Ψ with $\|\psi_k - \phi_k\|_2 < \varepsilon$ for balanced x_I the best S-term approximation in Ψ will still use the same support I. In contrast, if y is ℓ-cosparse with respect to an analysis operator Ω, $\Omega_\Lambda y = 0$, then for a perturbed operator Γ with $\|\gamma_k - \omega_k\|_2 < \varepsilon$ the smallest ℓ entries of Γy will not all be located in Λ. In order to get a local convergence result one has to deal with the fact that only part of the cosupport is preserved. We expect that for most signals containing k in the cosupport with respect to Ω, k will also be in the cosupport with respect to Γ. Unfortunately the mathematical tools necessary to quantify these statements are much more involved than the comparatively simple results necessary for the convergence of dictionary learning and so the local convergence analysis remains on our agenda for future research. It is also possible that the analysis can be carried out from a dynamical systems perspective using the differential equations described in the beginning of Section 5.3.

Another research direction we are currently pursuing is inspired by the shape of the analysis operators learned on noiseless images. The translation invariance of the edge detector like analysers suggests to directly assume translation invariance of the analysis operator. Such an operator has two advantages, first, learning it will require less training samples and second, since it can be reduced to several translated mother functions, it will be cost efficient to store and apply.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y760. In addition the computational results presented have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck. Part of this work has been carried out while M.S. was supported by the trimester program ’Mathematics of Signal Processing’ at the Hausdorff Research Institute for Mathematics. The authors would also like to thank Marie Pali for proofreading the manuscript and her valuable comments.

5.6 Computation of the optimal stepsize

To find the optimal stepsize, we first compute the derivative of

$$F(\alpha) = \frac{(\gamma_k - \alpha g_k) A_k (\gamma_k - \alpha g_k)^\star}{\|(\gamma_k - \alpha g_k)\|^2},$$

which is given by

$$F'(\alpha) = -2\frac{(b-a^2) + \alpha(ab-c) + \alpha^2(ac-b^2)}{(1-2\alpha a + \alpha^2 b^2)^2},$$

where we used the notation $a = \gamma_k A_k \gamma_k^\star$, $b = \gamma_k A_k^2 \gamma_k^\star$ and $c = \gamma_k A_k^3 \gamma_k^\star$.

First, suppose that $b^2 \neq ac$. Setting the first derivative equal to zero and solving a quadratic equation gives the results

$$\alpha_\pm = \frac{ab - c \pm \sqrt{(c-ab)^2 - 4(b^2-ac)(a^2-b)}}{2(b^2-ac)}. \tag{5.14}$$

The discriminant is always larger or equal than zero, as

$$(c-ab)^2 - 4(b^2-ac)(a^2-b) = (c + 2a^3 - 3ab)^2 - 4(a^2-b)^3$$

and $a^2 - b = (\gamma_k A_k \gamma_k^\star)^2 - \gamma_k A_k^2 \gamma_k^\star = \gamma_k A_k (\gamma_k^\star \gamma_k - \mathbb{1}) A_k \gamma_k^\star \le 0$, because the matrix $\gamma_k^\star \gamma_k - \mathbb{1}$ is negative semidefinite.

We can verify that α_+ indeed minimizes F, by substituting it into the second derivative

$$F''(\alpha) = -2\tfrac{(ab-c+2\alpha(ac-b^2))(1-2\alpha a+\alpha^2 b^2)^2 - 4(1-2\alpha a+\alpha^2 b^2)(\alpha b-a)(b-a^2+\alpha(ab-c)+\alpha^2(ac-b^2))}{(1-2\alpha a+\alpha^2 b^2)^4}.$$

We see that

$$F''(\alpha_+) = -2\tfrac{(ab-c+2\alpha_+(ac-b^2))(1-2\alpha_+ a+\alpha_+^2 b^2)^2}{(1-2\alpha_+ a+\alpha_+^2 b^2)^4} = -2\tfrac{ab-c+2\alpha_+(ac-b^2)}{(1-2\alpha_+ a+\alpha_+^2 b^2)^2}.$$

The denominator of the fraction above is positive, so we need to show that the numerator is negative. Inserting the expression for α_+ from Equation (5.14) into the numerator yields

$$\begin{aligned} ab - c - 2(b^2 - ac)\tfrac{ab-c+\sqrt{(c-ab)^2-4(b^2-ac)(a^2-b)}}{2(b^2-ac)} &= \\ = -\sqrt{(c-ab)^2 - 4(b^2-ac)(a^2-b)} &\le 0, \end{aligned}$$

so $F''(\alpha_+) \ge 0$ and α_+ is indeed the desired minimum.

If $F''(\alpha_+)$ is zero, then α_+ need not be a minimum. However, this is only the case if

$$\sqrt{(c-ab)^2 - 4(b^2-ac)(a^2-b)} = 0.$$

The computation showing the positivity of the discriminant suggests that in this case $(c-ab)^2 - 4(b^2-ac)(a^2-b) = (c+2a^3-3ab)^2 + 4(b-a^2)^3 = 0$. This is a sum of two nonnegative numbers, so both numbers must be zero. However, $b - a^2$ has been shown to vanish only if γ_k is an eigenvector of A_k. In this case also $c + 2a^3 - 3ab = 0$, which shows that for $b^2 \ne ac$, we have that $F''(\alpha_+) > 0$, unless γ_k is an eigenvector of A_k.

Now suppose that $b^2 = ac$. If $b = 0$, it follows that $g_k = 0$ and hence $F(\alpha)$ is zero everywhere, so regardless of the choice of α, we cannot further decrease the objective function. In this case we choose $\alpha_+ = 0$. If $b \ne 0$, we have

$$F'(\alpha) = -2\frac{b - a^2 + \alpha(ab-c)}{(1-2\alpha a+\alpha^2 b^2)^2},$$

which vanishes for $\alpha_+ = \frac{a}{b}$. The second derivative in α_+ is given by

$$F''(\alpha_+) = -2\frac{ab-c}{(1-2\frac{a^2}{b}+a^2)^2}.$$

Again, the denominator is positive and the numerator can be shown to be negative using a similar symmetrisation argument as in the proof of Theorem 5.3.1. This argument also shows that $F''(\alpha_+) = 0$ if and only if γ_k is an eigenvector of A_k.

6 Outlook

> I may not have gone where I intended to go, but I think I have ended up where I needed to be.
>
> \- Douglas Adams

Wrapping up the thesis, in Chapters 2 and 3, we proposed two different methods to reduce the number of needed measurements in photoacoustic imaging using a sparsity prior. Due to the fact that the data itself was not sparse, sparsifying transformations have been derived in order to be able to use Compressed Sensing reconstruction guarantees. In Chapter 4, we examined the geometric structure of total variation minimization in more detail and derived recovery results for subgaussian measurement ensembles. Finally, in Chapter 5, we presented four optimization-based algorithms to learn operators that sparsify a given class of data. In addition, theorems were presented that show that these algorithms indeed decrease the given target function.

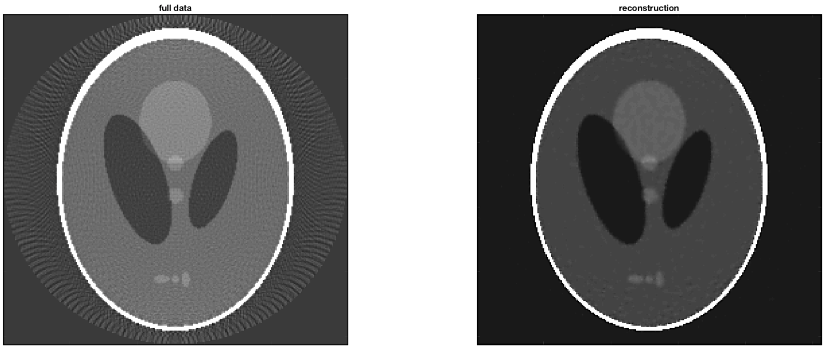

Figure 6.1: Reconstruction via FBP from 200 sensors (left) and reconstruction from 66 Bernoulli measurements after 1000 iterations using the proposed method (right).

From the presented results, several research topics emerge. First, consider a question raised in Chapter 3, namely the problem stated in Remark 3.3.1. As we have mentioned, instead of solving the joint reconstruction problem given in Equation (3.11), we can instead solve

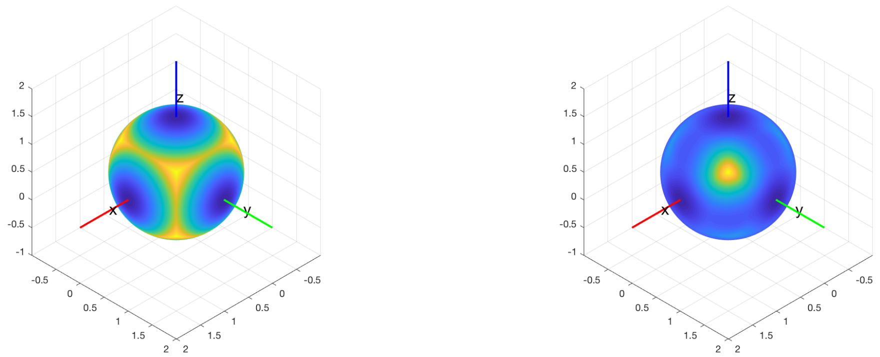

Figure 6.2: The target function generated from the first standard basis vector as target for $d = 3$ with cosparsity level $\ell = 2$ (left) and $\ell = 1$ (right). One can clearly see how for low cosparsity levels local minima emerge.

$$\begin{aligned} &\min_f \|c\Delta_{\mathbf{r}} f\|_1 + I_C(f) \\ &\text{such that } \mathbf{M}f = y\,. \end{aligned}$$

A relaxed formulation also incorporating possible noise is given by

$$\min_f \|c\Delta_{\mathbf{r}} f\|_1 + I_C(f) + \tfrac{\lambda}{2}\|\mathbf{M}f - y\|_2^2\,.$$

For this problem, one of the simplest approaches is a proximal subgradient algorithm. The subgradient of $\|c\Delta_{\mathbf{r}} f\|_1$ with respect to f is $c\Delta \sum_i \operatorname{sign}(c\Delta f)_i$, the gradient of $\frac{\lambda}{2}\|\mathbf{M}f - y\|_2^2$ is given by $\lambda\mathbf{M}^*(\mathbf{M}f - y)$. The proximal projection on the positive cone C is componentwise given by $\operatorname{prox}_C(y)_i = \max(y_i, 0)$. So one step of the proximal subgradient algorithm is given by

$$\bar{f} = \max\left(0, f - \mu\left(c\Delta\sum_i \operatorname{sign}(c\Delta f)_i + \lambda\mathbf{M}^*(\mathbf{M}f - y)\right)\right),$$

where $\bar{f}$ is the update and μ is some stepsize. Preliminary experiments suggest that this program also works very well at least for synthetic data, cf. Figure 6.1.

Another possible research direction was pointed out in Chapter 5. In images, translation invariant structure appears quite naturally. Such a structure is for example exhibited in convolutional analysis operators, which consist of few atoms, which are convolved with the data. It can again be seen as the cosparse analogon to convolutional dictionary learning [PRSE17, GCW17].

This bears close resemblance to blind deconvolution and blind demixing problems, where one is given a single signal, which can be decomposed in a sum of one or more unknown sparse vectors convolved with unknown atoms and possibly corrupted with noise. For blind deconvolution and blind demixing several theoretical results are known, cf. [JKS17, LS17a, LS17b]. In our setting, we are not just given a single signal but a sample of several different signals or perhaps even an online stream of data. So in a similar spirit as in Chapter 5, we can solve the empirical minimization problem

$$\text{minimize} \sum_{n=1}^{N} \min_{|J|=\ell} \|((x_1 * y_n)^*, (x_2 * y_n)^*, \dots, (x_K * y_n)^*)_J^*\|_2^2$$

with respect to $x_1, \dots, x_K \in \mathbb{S}^{d-1}$. For $K = 1$, this problem is again dual to the problem of finding a convolutional dictionary. It seems furthermore that in this case, the optimization problem has, up to the trivial ambiguities, no spurious local minima if the cosparsity level is chosen large enough.

Note that the application of these operators is now significantly cheaper as the convolutions can be implemented efficiently via an FFT.

For the minimization again as in Chapter 5, explicit or implicit projected gradient methods can be used.

List of Figures

Bibliography

[ABB$^+$16] S. Arridge, P. Beard, M. Betcke, B. Cox, N. Huynh, F. Lucka, O. Ogunlade, and E. Zhang. Accelerated high-resolution photoacoustic tomography via compressed sensing. *Phys. Med. Biol.*, 61(24):8908, 2016.

[AEB06] M. Aharon, M. Elad, and A.M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing.*, 54(11):4311–4322, November 2006.

[BBMG$^+$07] P. Burgholzer, J. Bauer-Marschallinger, H. Grün, M. Haltmeier, and G. Paltauf. Temporal back-projection algorithms for photoacoustic tomography with integrating line detectors. *Inverse Probl.*, 23(6):S65–S80, 2007.

[BCH$^+$17] M. M. Betcke, B. T. Cox, N. Huynh, E. Z. Zhang, P. C. Beard, and S. R. Arridge. Acoustic wave field reconstruction from compressed measurements with application in photoacoustic tomography. *IEEE Trans. Comput. Imaging*, 3:710–721, 2017.

[BDDW08] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin. A simple proof of the Restricted Isometry Property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[Bea11] P. Beard. Biomedical photoacoustic imaging. *Interface focus*, 1(4):602–631, 2011.

[Ber11] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

[BGI$^+$08] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *46th Annual Allerton Conference on Communication, Control, and Computing, 2008*, pages 798–805, 2008.

[BHP$^+$05] P. Burgholzer, C. Hofer, G. Paltauf, M. Haltmeier, and O. Scherzer. Thermoacoustic tomography with integrating area and line detectors. *IEEE Trans. Ultrason., Ferroeletr., Freq. Control*, 52(9):1577–1583, 2005.

[BIR08] R. Berinde, P. Indyk, and M. Ruzic. Practical near-optimal sparse recovery in the l_1 norm. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 198–205, 2008.

[BMFB$^+$15] J. Bauer-Marschallinger, K. Felbermayer, K.-D. Bouchal, I. A. Veres, H. Grün, P. Burgholzer, and T. Berer. Photoacoustic projection imaging using a 64-channel fiber optic detector array. In *Proc. SPIE*, volume 9323, 2015.

[BMFB17] J. Bauer-Marschallinger, K. Felbermayer, and T. Berer. All-optical photoacoustic projection imaging. *Biomed. Opt. Express*, 8(9):3938–3951, 2017.

[BMFH$^+$13] J. Bauer-Marschallinger, K. Felbermayer, A. Hochreiner, H. Grün, G. Paltauf, P. Burgholzer, and T. Berer. Low-cost parallelization of optical fiber based detectors for photoacoustic imaging. In *Proc. SPIE*, volume 8581, pages 85812M–85812M–8, 2013.

[BMHP07] P. Burgholzer, G. J. Matt, M. Haltmeier, and G. Paltauf. Exact and approximate imaging methods for photoacoustic tomography using an arbitrary detection surface. *Phys. Rev. E*, 75(4):046706, 2007.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[BVG$^+$12] T. Berer, I. A. Veres, H. Grün, J. Bauer-Marschallinger, K. Felbermayer, and P. Burgholzer. Characterization of broadband fiber optic line detectors for photoacoustic tomography. *J. Biophotonics*, 5(7):518–528, 2012.

[Can08] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9):589–592, 2008.

[CDDD03] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Harmonic analysis of the space bv. *Rev. Mat. Iberoam.*, 19(1):235–263, 2003.

[CENR10] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59–73, 2010.

[CL97] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997.

[CN17] J. Chung and L. Nguyen. Motion estimation and correction in photoacoustic tomographic reconstruction. *SIAM J. Imaging Sci. 10(2): 535–557, 2017 [arxiv.org]*, 10(1):216–242, 2017.

[CP11] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[CR07] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Probl.*, 23(3):969, 2007.

[CRT06a] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[CRT06b] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[CT06] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12), 2006.

[CTL08] G.-H. Chen, J. Tang, and S. Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Med. Phys.*, 35(2):660–663, 2008.

[CX15] J.-F. Cai and W. Xu. Guarantees of total variation minimization for signal recovery. *Information and Inference*, 4(4):328–353, 2015.

[CZ14] T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inf. Theory*, 60:122 – 132, 2014.

[DDT$^+$08] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, 2008.

[DE03] D. L. Donoho and M. Elad. Maximal sparsity representation via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, March 2003.

[DK01] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Stat.*, pages 1–48, 2001.

[Don04] D. Donoho. High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension. Technical report, Department of Statistics, Stanford University, 2004.

[Don06a] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[Don06b] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. *Commun. Pure Appl. Anal.*, 59(6):797–829, 2006.

[DWD14] J. Dong, W. Wang, and W. Dai. Analysis SimCO: A new algorithm for analysis dictionary learning. In *ICASSP14*, 2014.

[DWD$^+$16] J. Dong, W. Wang, W. Dai, M.D. Plumbley, Z. Han, and J. Chambers. Analysis SimCO algorithms for sparse analysis model based dictionary learning. *IEEE Transactions on Signal Processing*, 64(2):417–431, 2016.

[EA06] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[EB14] E. M. Eksioglu and O. Bayir. K-SVD meets transform learning: Transform K-SVD. *IEEE Signal Processing Letters*, 21(3):347–351, 2014.

[EHN96] H.-W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[EMR07] M. Elad, R. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23(3):947–968, 2007.

[FHR07] D. Finch, M. Haltmeier, and Rakesh. Inversion of spherical means and the wave equation in even dimensions. *SIAM J. Appl. Math.*, 68(2):392–412, 2007.

[FPR04] D. Finch, S. K. Patch, and Rakesh. Determining a function from its mean values over a family of spheres. *SIAM J. Math. Anal.*, 35(5):1213–1240, 2004.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[FRon] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Appl. Numer. Harmon. Anal. Birkhäuser, Boston, in preparation.

[Fuc04] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6), 2004.

[GBB$^+$10] H. Grün, T. Berer, P. Burgholzer, R. Nuster, and G. Paltauf. Three-dimensional photoacoustic imaging using fiber-based line detectors. *J. Biomed. Optics*, 15(2):021306–021306–8, 2010.

[GCW17] C. Garcia-Cardona and B. Wohlberg. Convolutional dictionary learning. *arXiv preprint arXiv:1709.02893*, 2017.

[GHS08] M. Grasmair, M. Haltmeier, and O. Scherzer. Sparse regularization with l^q penalty term. *Inverse Probl.*, 24(5):055020, 13, 2008.

[GHS11] M. Grasmair, M. Haltmeier, and O. Scherzer. Necessary and sufficient conditions for linear convergence of ℓ^1-regularization. *Comm. Pure Appl. Math.*, 64(2):161–182, 2011.

[GI10] A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proc. IEEE*, 98(6):937–947, 2010.

[GLSW10] Z. Guo, C. Li, L. Song, and L. V. Wang. Compressed sensing in photoacoustic tomography in vivo. *J. Biomed. Opt.*, 15(2):021311–021311, 2010.

[GNE$^+$14] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M.E. Davies. Greedy-like algorithms for the cosparse analysis model. *Linear Algebra and its Applications*, 441:22–60, 2014.

[GNW$^+$14] S. Gratt, R. Nuster, G. Wurzinger, M. Bugl, and G. Paltauf. 64-line-sensor array: fast imaging system for photoacoustic tomography. *Proc. SPIE*, 8943:894365–894365–6, 2014.

[GO08] M. Grasmair and A. Obereder. Generalizations of the taut string method. *Numer. Funct. Anal. Optim.*, 29(3-4):346–361, 2008.

[Gor88] Y. Gordon. *On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$*. Springer, 1988.

[GV91] R. Gorenflo and S. Vessella. *Abel integral equations*, volume 1461 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1991. Analysis and applications.

[Hal13] M. Haltmeier. Inversion of circular means and the wave equation on convex planar domains. *Computers & Mathematics with Applications. An International Journal*, 65(7):1025–1036, 2013.

[Hal14] M. Haltmeier. Universal inversion formulas for recovering a function from spherical means. *SIAM J. Math. Anal.*, 46(1):214–232, 2014.

[Hal16] M. Haltmeier. Sampling conditions for the circular radon transform. *IEEE Trans. Image Process.*, 25(6):2910–2919, 2016.

[HBMB16] M. Haltmeier, T. Berer, S. Moon, and P. Burgholzer. Compressed sensing and sparsity in photoacoustic tomography. *J. Opt.*, 18(11):114004–12pp, 2016.

[HKD13] S. Hawe, M. Kleinsteuber, and K. Diepold. Analysis operator learning and its application to image reconstruction. *IEEE Transactions on Image Processing*, 22(6):2138–2150, 2013.

[HKN08] Y. Hristova, P. Kuchment, and L. Nguyen. Reconstruction and time reversal in thermoacoustic tomography in acoustically homogeneous and inhomogeneous media. *Inverse Probl.*, 24(5):055006 (25pp), 2008.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc.*, 43(4):439–561, 2006.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[HSB$^+$07] M. Haltmeier, O. Scherzer, P. Burgholzer, R. Nuster, and G. Paltauf. Thermoacoustic tomography and the circular radon transform: exact inversion formula. *Mathematical Models and Methods in Applied Sciences*, 17(04):635–655, 2007.

[HSB$^+$17] M. Haltmeier, M. Sandbichler, T. Berer, J. Bauer-Marschallinger, P. Burgholzer, and L. Nguyen. A new sparsification and reconstruction strategy for compressed sensing photoacoustic tomography. *arXiv preprint arXiv:1801.00117*, 2017.

[HSS05] M. Haltmeier, T. Schuster, and O. Scherzer. Filtered backprojection for thermoacoustic computed tomography in spherical geometry. *Mathematical methods in the applied sciences*, 28(16):1919–1937, 2005.

[HW10] E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition, paperback.

[HWNW13] C. Huang, K. Wang, L. Nie, and M. A. Wang, L. V.and Anastasio. Full-wave iterative image reconstruction in photoacoustic tomography with acoustically inhomogeneous media. *IEEE Trans. Med. Imag.*, 32(6):1097–1110, 2013.

[HZB$^+$14] N. Huynh, E. Zhang, M. Betcke, S. Arridge, P. Beard, and B. Cox. Patterned interrogation scheme for compressed sensing photoacoustic imaging using a fabry perot planar sensor. In *Proc. SPIE*, pages 894327–894327, 2014.

[HZB$^+$16] N. Huynh, E. Zhang, M. Betcke, S. Arridge, P. Beard, and B. Cox. Single-pixel optical camera for video rate ultrasonic imaging. *Optica*, 3(1):26–29, 2016.

[IR08] P. Indyk and M. Ruzic. Near-optimal sparse recovery in the l_1 norm. In *49th Annual IEEE Symposium on Foundations of Computer Science*, pages 199–207, 2008.

[JKL15] J. Jørgensen, C. Kruschel, and D. Lorenz. Testable uniqueness conditions for empirical assessment of undersampling levels in total variation-regularized x-ray ct. *Inverse Probl. Sci. En.*, 23(8):1283–1305, 2015.

[JKS17] P. Jung, F. Krahmer, and D. Stöger. Blind demixing and deconvolution at near-optimal rate. *arXiv preprint arXiv:1704.04178*, 2017.

[Joh82] F. John. *Partial Differential Equations*, volume 1 of *Applied Mathematical Sciences*. Springer Verlag, New York, fourth edition, 1982.

[JXHC09] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inf. Theory*, 55(9):4299–4308, 2009.

[KK08] P. Kuchment and L. A. Kunyansky. Mathematics of thermoacoustic and photoacoustic tomography. *Eur. J. Appl. Math.*, 19:191–224, 2008.

[KKS17] F. Krahmer, C. Kruschel, and M. Sandbichler. Total variation minimization in compressed sensing. In *Compressed Sensing and its Applications*, pages 333–358. Springer, 2017.

[Kla84] M. Klawe. Limitations on explicit constructions of expanding graphs. *SIAM J. Comput.*, 13:156–166, 1984.

[KM15] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Notices*, 2015(23):12991–13008, 2015.

[KMR14a] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.

[KMR14b] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.

[KNW15] F. Krahmer, D. Needell, and R Ward. Compressive sensing with redundant dictionaries and structured measurements. *SIAM J. Math. Anal.*, 47(6):4606–4629, 2015.

[KR14] F. Krahmer and H. Rauhut. Structured random measurements in signal processing. *GAMM-Mitteilungen*, 37(2):217–238, 2014.

[KR15] M. Kabanava and H. Rauhut. Analysis ℓ_1-recovery with frames and gaussian measurements. *Acta Appl. Math.*, 140(1):173–195, 2015.

[Kru15] C. Kruschel. *Geometrical Interpretations and Algorithmic Verification of Exact Solutions in Compressed Sensing*. PhD thesis, TU Braunschweig, 2015.

[KRZ15] M. Kabanava, H. Rauhut, and H. Zhang. Robust analysis ℓ_1-recovery from gaussian measurements and total variation minimization. *European J. Appl. Math.*, 26(06):917–929, 2015.

[Kun07] L. A. Kunyansky. Explicit inversion formulae for the spherical mean Radon transform. *Inverse Probl.*, 23(1):373–383, 2007.

[KW11] F. Krahmer and R. Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

[KW14a] F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *IEEE Trans. Image Proc.*, 23(2):612–622, 2014.

[KW14b] F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *IEEE Trans. Image Proc.*, 23(2):612–622, 2014.

[LDP07] M. Lustig, D. Donoho, and J. M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine*, 58(6):1182–1195, 2007.

[LDSP08] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compressed sensing mri. *IEEE Sig. Proc. Mag.*, 25(2):72–82, 2008.

[LS17a] S. Ling and T. Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Trans. Inf. Theory*, 63(7):4497–4520, 2017.

[LS17b] S. Ling and T. Strohmer. Regularized gradient descent: A nonconvex recipe for fast joint blind deconvolution and demixing. *arXiv preprint arXiv:1703.08642*, 2017.

[Lub94] A. Lubotzky. *Discrete groups, expanding graphs and invariant measures*, volume 125. Springer, 1994.

[Lub12] A. Lubotzky. Expander graphs in pure and applied mathematics. *Bull. Amer. Math. Soc.*, 49(1):113–162, 2012.

[Men14] S. Mendelson. Learning without concentration. In *COLT*, pages 25–39, 2014.

[MPTJ07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[MvdG$^+$97] E. Mammen, S. van de Geer, et al. Locally adaptive regression splines. *Ann. Stat.*, 25(1):387–413, 1997.

[NDEG13] S. Nam, M. Davies, M. Elad, and R. Gribonval. The cosparse analysis model and algorithms. *Appl. Comput. Harmon. Anal.*, 34(1):30–56, 2013.

[Ngu09] Linh V. Nguyen. A family of inversion formulas in thermoacoustic tomography. *Inverse Probl. Imaging*, 3(4):649–675, 2009.

[NHK$^+$10] R. Nuster, M. Holotta, C. Kremser, H. Grossauer, P. Burgholzer, and G. Paltauf. Photoacoustic microtomography using optical interferometric detection. *J. Biomed. Optics*, 15(2):021307–021307–6, 2010.

[NW13a] D. Needell and R. Ward. Near-optimal compressed sensing guarantees for total variation minimization. *IEEE Trans. Image Proc.*, 22(10):3941–3949, 2013.

[NW13b] D. Needell and R. Ward. Stable image reconstruction using total variation minimization. *SIAM J. Imaging Sci.*, 6(2):1035–1058, 2013.

[OEBP11] B. Ophir, M. Elad, N. Bertin, and M. D. Plumbley. Sequential minimal eigenvalues-an approach to analysis dictionary learning. In *Signal Processing Conference, 2011 19th European*, pages 1465–1469. IEEE, 2011.

[PHKN17] G. Paltauf, P. Hartmair, G. Kovachev, and R. Nuster. Piezoelectric line detector array for photoacoustic tomography. *Photoacoustics*, 8:28–36, 2017.

[PL09] J. Provost and F. Lesage. The application of compressed sensing for photo-acoustic tomography. *IEEE Trans. Med. Imag.*, 28(4):585–594, 2009.

[PMG$^+$12] G. Puy, J. Marques, R. Gruetter, J.-P. Thiran, D. Van De Ville, P. Vandergheynst, and Y. Wiaux. Spread spectrum magnetic resonance imaging. *IEEE Trans. Med. Imaging*, 31(3):586–598, 2012.

[PNHB07a] G. Paltauf, R. Nuster, M. Haltmeier, and P. Burgholzer. Experimental evaluation of reconstruction algorithms for limited view photoacoustic tomography with line detectors. *Inverse Probl.*, 23(6):S81–S94, 2007.

[PNHB07b] G. Paltauf, R. Nuster, M. Haltmeier, and P. Burgholzer. Photoacoustic tomography using a Mach-Zehnder interferometer as an acoustic line detector. *Appl. Opt.*, 46(16):3352–3358, 2007.

[PNHB09] G. Paltauf, R. Nuster, M. Haltmeier, and P. Burgholzer. Photoacoustic tomography with integrating area and line detectors. In *Photoacoustic imaging and spectroscopy*, chapter 20, pages 251–263. CRC Press, 2009.

[Poo15] C. Poon. On the role of total variation in compressed sensing. *SIAM J. Imag. Sci*, 8(1):682–720, 2015.

[PRSE17] V. Papyan, Y. Romano, J. Sulam, and M. Elad. Convolutional dictionary learning via local processing. *arXiv preprint arXiv:1705.03239*, 2017.

[PV13] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory*, 59(1):482–494, 2013.

[PVW11] G. Puy, P. Vandergheynst, and Y. Wiaux. On variable density compressive sampling. *IEEE Signal Proc. Let.*, 18:595–598, 2011.

[Rau07] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comp. Harm. Anal.*, 22(1):16–42, 2007.

[RB13] S. Ravishankar and Y. Bresler. Learning sparsifying transforms. *IEEE Transactions on Signal Processing*, 61(5):1072–1086, 2013.

[RBE10] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[Roc72] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1972.

[ROF92] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1–4):259–268, 1992.

[RPE13] R. Rubinstein, T. Peleg, and M. Elad. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model. *IEEE Transactions on Signal Processing*, 61(3):661–677, 2013.

[RRT12a] H. Rauhut, J. Romberg, and J. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmon. Anal.*, 32(2):242–254, 2012.

[RRT12b] H. Rauhut, J. Romberg, and J.A. Tropp. Restricted isometries for partial random circulant matrices. *Appl. Comput. Harmon. Anal.*, 32(2):242–254, 2012.

[RSV08] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, 54(5):2210–2219, 2008.

[RV08] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[Sch15a] K. Schnass. Local identification of overcomplete dictionaries. *Journal of Machine Learning Research (arXiv:1401.6354)*, 16(Jun):1211–1242, 2015.

[Sch15b] K. Schnass. A personal introduction to theoretical dictionary learning. *Internationale Mathematische Nachrichten*, 228:5–15, 2015.

[Sch16] K. Schnass. Convergence radius and sample complexity of ITKM algorithms for dictionary learning. *accepted to Applied and Computational Harmonic Analysis (arXiv:1503.07027)*, 2016.

[Sho10] Ralph E Showalter. *Hilbert space methods in partial differential equations*. Courier Corporation, 2010.

[SKB$^+$15] M. Sandbichler, F. Krahmer, T. Berer, P. Burgholzer, and M. Haltmeier. A novel compressed sensing scheme for photoacoustic tomography. *SIAM J. Appl. Math.*, 75(6):2475–2494, 2015.

[SQW15] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. In *ICML 2015 (arXiv:1504.06785)*, 2015.

[SS03] H. Schwetlick and U. Schnabel. Iterative computation of the smallest singular value and the corresponding singular vectors of a matrix. *Linear algebra and its applications*, 371:1–30, 2003.

[SS17] M. Sandbichler and K. Schnass. Online and stable learning of analysis operators. *arXiv preprint arXiv:1704.00227*, 2017.

[SWGK16] M. Seibert, J. Wörmann, R. Gribonval, and M. Kleinsteuber. Learning co-sparse analysis operators with separable structures. *IEEE Transactions on Signal Processing*, 64(1):120–130, 2016.

[TC10] B. E. Treeby and B. T. Cox. k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave-fields. *J. Biomed. Opt.*, 15:021314, 2010.

[TP14] A. Tillmann and M. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, 60(2):1248–1259, 2014.

[Tro15] J. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[Ver15] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling theory, a renaissance*, pages 3–66. Springer, 2015.

[WA11] K. Wang and M. A. Anastasio. Photoacoustic and thermoacoustic tomography: image formation principles. In *Handbook of Mathematical Methods in Imaging*, pages 781–815. Springer, 2011.

[Wan09] L. V. Wang. Multiscale photoacoustic microscopy and computed tomography. *Nat. Photonics*, 3(9):503–509, 2009.

[XH07] W. Xu and B. Hassibi. Efficient compressive sensing with deterministic guarantees using expander graphs. In *IEEE Information Theory Workshop*, pages 414–419, 2007.

[XW05] M. Xu and L. V. Wang. Universal back-projection algorithm for photoacoustic computed tomography. *Phys. Rev. E*, 71(1):0167061–0167067, 2005.

[XW06a] M. Xu and L. V. Wang. Photoacoustic imaging in biomedicine. *Rev. Sci. Instrum.*, 77(4):041101 (22pp), 2006.

[XW06b] M. Xu and L. V. Wang. Photoacoustic imaging in biomedicine. *Rev. Sci. Instruments*, 77(4):041101, 2006.

[YNGD11] M. Yaghoobi, S. Nam, R. Gribonval, and M.E. Davies. Analysis operator learning for overcomplete cosparse representations. In *EUSIPCO11*, pages 1470–1474, 2011.

[YNGD13] M. Yaghoobi, S. Nam, R. Gribonval, and M.E. Davies. Constrained overcomplete analysis operator learning for cosparse signal modelling. *IEEE Transactions on Signal Processing*, 61(9):2341–2355, 2013.

[ZMY16] H. Zhang, Y. Ming, and W. Yin. One condition for solution uniqueness and robustness of both ℓ_1-synthesis and ℓ_1-analysis minimizations. *Adv. Comput. Math.*, 42(6):1381–1399, 2016.