Compressive Sensing

Massimo Fornasier and Holger Rauhut

Austrian Academy of Sciences
Johann Radon Institute for
Computational and Applied
Mathematics (RICAM)
Altenbergerstrasse 69
A-4040, Linz, Austria
massimo.fornasier@oeaw.ac.at

Hausdorff Center for Mathematics,
Institute for Numerical Simulation
University of Bonn
Endenicher Allee 60
D-53115 Bonn, Germany
rauhut@hcm.uni-bonn.de

April 18, 2010

Abstract

Compressive sensing is a new type of sampling theory, which predicts that sparse signals and images can be reconstructed from what was previously believed to be incomplete information. As a main feature, efficient algorithms such as ℓ_1-minimization can be used for recovery. The theory has many potential applications in signal processing and imaging. This chapter gives an introduction and overview on both theoretical and numerical aspects of compressive sensing.

1 Introduction

The traditional approach of reconstructing signals or images from measured data follows the well-known Shannon sampling theorem [94], which states that the sampling rate must be twice the highest frequency. Similarly, the fundamental theorem of linear algebra suggests that the number of collected samples (measurements) of a discrete finite-dimensional signal should be at least as large as its length (its dimension) in order to ensure reconstruction. This principle underlies most devices of current technology, such as analog to digital conversion, medical imaging or audio and video electronics. The novel theory of compressive sensing (CS) — also known under the terminology of compressed sensing, compressive sampling or sparse recovery — provides a fundamentally new approach to data acquisition which overcomes this

common wisdom. It predicts that certain signals or images can be recovered from what was previously believed to be highly incomplete measurements (information). This chapter gives an introduction to this new field. Both fundamental theoretical and algorithmic aspects are presented, with the awareness that it is impossible to retrace in a few pages all the current developments of this field, which was growing very rapidly in the past few years and undergoes significant advances on an almost daily basis.

CS relies on the empirical observation that many types of signals or images can be well-approximated by a sparse expansion in terms of a suitable basis, that is, by only a small number of non-zero coefficients. This is the key to the efficiency of many lossy compression techniques such as JPEG, MP3 etc. A compression is obtained by simply storing only the largest basis coefficients. When reconstructing the signal the non-stored coefficients are simply set to zero. This is certainly a reasonable strategy when full information of the signal is available. However, when the signal first has to be acquired by a somewhat costly, lengthy or otherwise difficult measurement (sensing) procedure, this seems to be a waste of resources: First, large efforts are spent in order to obtain full information on the signal, and afterwards most of the information is thrown away at the compression stage. One might ask whether there is a clever way of obtaining the compressed version of the signal more directly, by taking only a small number of measurements of the signal. It is not obvious at all whether this is possible since measuring directly the large coefficients requires to know *a priori* their location. Quite surprisingly, compressive sensing provides nevertheless a way of reconstructing a compressed version of the original signal by taking only a small amount of *linear* and *non-adaptive* measurements. The precise number of required measurements is comparable to the compressed size of the signal. Clearly, the measurements have to be suitably designed. It is a remarkable fact that all provably good measurement matrices designed so far are random matrices. It is for this reason that the theory of compressive sensing uses a lot of tools from probability theory.

It is another important feature of compressive sensing that practical reconstruction can be performed by using efficient algorithms. Since the interest is in the vastly undersampled case, the linear system describing the measurements is underdetermined and therefore has infinitely many solution. The key idea is that the sparsity helps in isolating the original vector. The first naive approach to a reconstruction algorithm consists in searching for the sparsest vector that is consistent with the linear measurements. This leads to the combinatorial ℓ_0-problem, see (3.4) below, which unfortunately is NP-hard in general. There are essentially two approaches for

tractable alternative algorithms. The first is convex relaxation leading to ℓ_1-minimization — also known as basis pursuit, see (3.5) — while the second constructs greedy algorithms. This overview focuses on ℓ_1-minimization. By now basic properties of the measurement matrix which ensure sparse recovery by ℓ_1-minimization are known: the *null space property (NSP)* and the *restricted isometry property (RIP)*. The latter requires that all column submatrices of a certain size of the measurement matrix are well-conditioned. This is where probabilistic methods come into play because it is quite hard to analyze these properties for deterministic matrices with minimal amount of measurements. Among the provably good measurement matrices are Gaussian, Bernoulli random matrices, and partial random Fourier matrices.

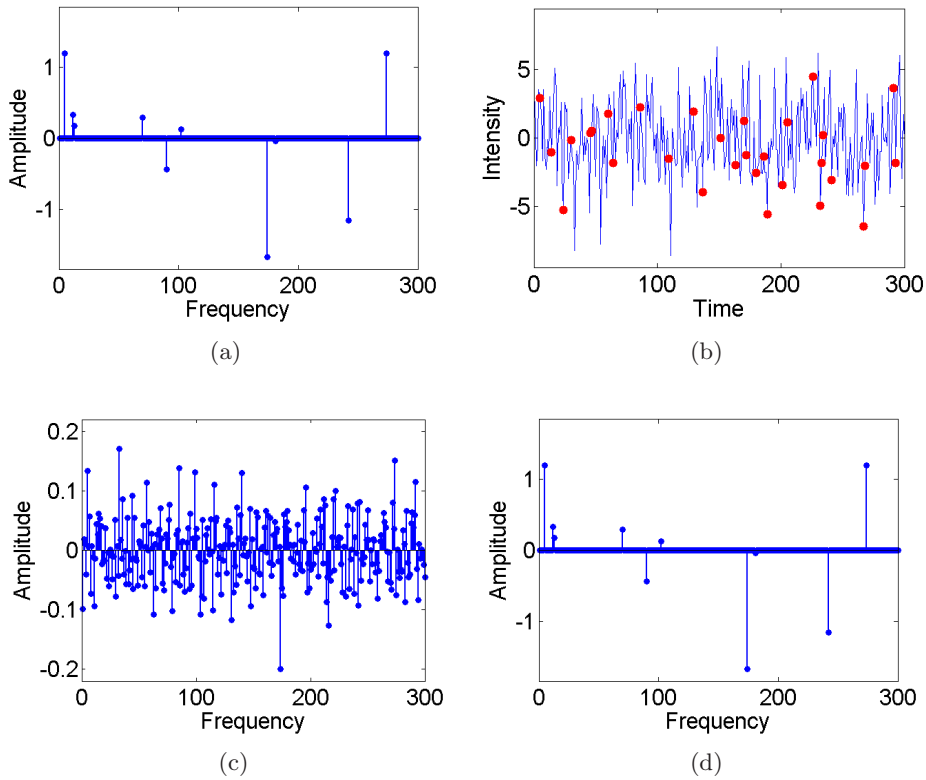

Figure 1: (a) 10-sparse Fourier spectrum, (b) time domain signal of length 300 with 30 samples, (c) reconstruction via ℓ_2-minimization, (d) exact reconstruction via ℓ_1-minimization

Figure 1 serves as a first illustration of the power of compressive sensing. It shows an example for recovery of a 10-sparse signal $x \in \mathbb{C}^{300}$ from only 30 samples (indicated by the red dots in Figure 1(b)). From a first look at the time-domain signal, one would rather believe that reconstruction should be impossible from only 30 samples. Indeed, the spectrum reconstructed by traditional ℓ_2-minimization is very different from the true spectrum. Quite surprisingly, ℓ_1-minimization performs nevertheless an exact reconstruction, that is, with no recovery error at all!

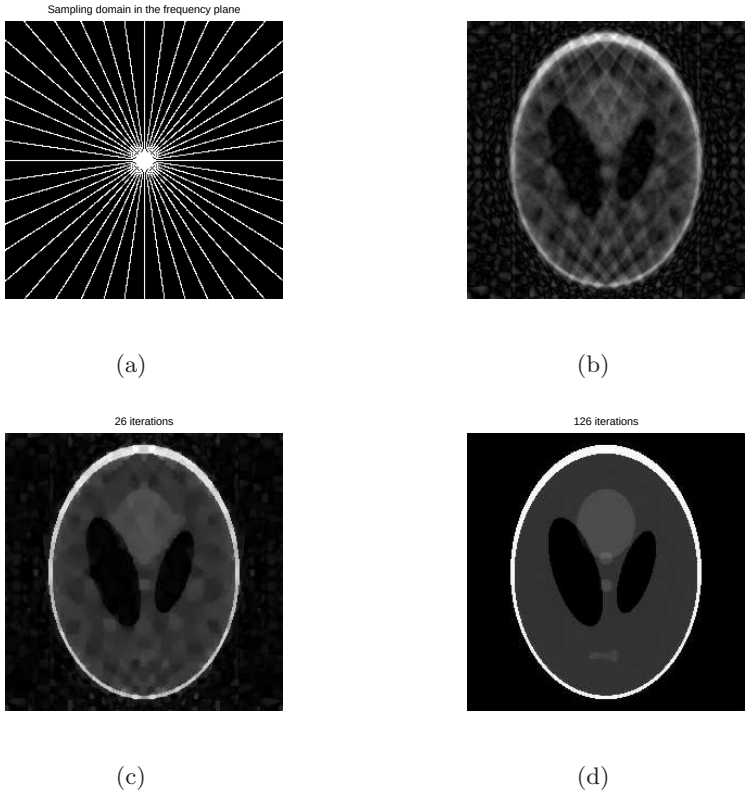

Figure 2: (a) Sampling data of the NMR image in the Fourier domain which corresponds to only 0.11% of all samples. (b) Reconstruction by backprojection. (c) Intermediate iteration of an efficient algorithm for large scale total variation minimization. (d) The final reconstruction is exact.

An example from nuclear magnetic resonance imaging serves as a second illustration. Here, the device scans a patient by taking 2D or 3D frequency measurements within a radial geometry. Figure 2(a) describes such a sam-

pling set of a 2D Fourier transform. Since a lengthy scanning procedure is very uncomfortable for the patient it is desired to take only a minimal amount of measurements. Total variation minimization, which is closely related to ℓ_1-minimization, is then considered as recovery method. For comparison, Figure 2(b) shows the recovery by a traditional backprojection algorithm. Figures 2(c), 2(d) display iterations of an algorithm, which was proposed and analyzed in [40] to perform efficient large scale total variation minimization. The reconstruction in Figure 2(d) is again exact!

2 Background

Although the term compressed sensing (compressive sensing) was coined only recently with the paper by Donoho [26], followed by a huge research activity, such a development did not start out of thin air. There were certain roots and predecessors in application areas such as image processing, geophysics, medical imaging, computer science as well as in pure mathematics. An attempt is made to put such roots and current developments into context below, although only a partial overview can be given due to the numerous and diverse connections and developments.

2.1 Early Developments in Applications

Presumably the first algorithm which can be connected to sparse recovery is due to the French mathematician de Prony [71]. The so-called Prony method, which has found numerous applications [62], estimates non-zero amplitudes and corresponding frequencies of a sparse trigonometric polynomial from a small number of equispaced samples by solving an eigenvalue problem. The use of ℓ_1-minimization appears already in the Ph.D. thesis of B. Logan [59] in connection with sparse frequency estimation, where he observed that L_1-minimization may recover exactly a frequency-sparse signal from undersampled data provided the sparsity is small enough. The paper by Donoho and Logan [25] is perhaps the earliest theoretical work on sparse recovery using L_1-minimization. Nevertheless, geophysicists observed in the late 1970's and 1980's that ℓ_1-minimization can be successfully employed in reflection seismology where a sparse reflection function indicating changes between subsurface layers is sought [87, 80]. In NMR spectroscopy the idea to recover sparse Fourier spectra from undersampled non-equispaced samples was first introduced in the 90's [96] and has seen a significant development since then. In image processing the use of total-variation minimization, which is closely connected to ℓ_1-minimization and compressive sensing, first

appears in the 1990's in the work of Rudin, Osher and Fatemi [79], and was widely applied later on. In statistics where the corresponding area is usually called *model selection* the use of ℓ_1-minimization and related methods was greatly popularized with the work of Tibshirani [88] on the so-called LASSO (Least Absolute Shrinkage and Selection Operator).

2.2 Sparse Approximation

Many lossy compression techniques such as JPEG, JPEG-2000, MPEG or MP3 rely on the empirical observation that audio signals and digital images have a sparse representation in terms of a suitable basis. Roughly speaking one compresses the signal by simply keeping only the largest coefficients. In certain scenarios such as audio signal processing one considers the generalized situation where sparsity appears in terms of a redundant system — a so called dictionary or frame [19] — rather than a basis. The problem of finding the sparsest representation / approximation in terms of the given dictionary turns out to be significantly harder than in the case of sparsity with respect to a basis where the expansion coefficients are unique. Indeed, in [61, 64] it was shown that the general ℓ_0-problem of finding the sparsest solution of an underdetermined system is NP-hard. Greedy strategies such as Matching Pursuit algorithms [61], FOCUSS [52] and ℓ_1-minimization [18] were subsequently introduced as tractable alternatives. The theoretical understanding under which conditions greedy methods and ℓ_1-minimization recover the sparsest solutions began to develop with the work in [30, 37, 29, 53, 49, 46, 91, 92].

2.3 Information Based Complexity and Gelfand Widths

Information based complexity (IBC) considers the general question of how well a function f belonging to a certain class $\mathcal{F}$ can be recovered from n sample values, or more generally, the evaluation of n linear or non-linear functionals applied to f [89]. The optimal recovery error which is defined as the maximal reconstruction error for the "best" sampling method and "best" recovery method (within a specified class of methods) over all functions in the class $\mathcal{F}$ is closely related to the so-called *Gelfand width* of $\mathcal{F}$ [66, 21, 26]. Of particular interest for compressive sensing is $\mathcal{F} = B_1^N$, the ℓ_1-ball in $\mathbb{R}^N$ since its elements can be well-approximated by sparse ones. A famous result due to Kashin [56], and Gluskin and Garnaev [47, 51] sharply bounds the Gelfand widths of B_1^N (as well as their duals, the *Kolmogorov widths*) from above and below, see also [44]. While the original interest of Kashin

was in the estimate of n-widths of Sobolev classes, these results give precise performance bounds in compressive sensing on how well any method may recover (approximately) sparse vectors from linear measurements [26, 21]. The upper bounds on Gelfand widths were derived in [56] and [47] using (Bernoulli and Gaussian) random matrices, see also [60], and in fact such type of matrices have become very useful also in compressive sensing [26, 16].

2.4 Compressive Sensing

The numerous developments in compressive sensing began with the seminal work [15] and [26]. Although key ingredients were already in the air at that time, as mentioned above, the major contribution of these papers was to realize that one can combine the power of ℓ_1-minimization and random matrices in order to show *optimal* results on the ability of ℓ_1-minimization of recovering (approximately) sparse vectors. Moreover, the authors made very clear that such ideas have strong potential for numerous application areas. In their work [16, 15] Candès, Romberg and Tao introduced the *restricted isometry property* (which they initially called the *uniform uncertainty principle*) which is a key property of compressive sensing matrices. It was shown that Gaussian, Bernoulli, and partial random Fourier matrices [16, 78, 73] possess this important property. These results require many tools from probability theory and finite dimensional Banach space geometry, which have been developed for a rather long time now, see e.g. [58, 55].

Donoho [28] developed a different path and approached the problem of characterizing sparse recovery by ℓ_1-minimization via polytope geometry, more precisely, via the notion of k-neighborliness. In several papers sharp phase transition curves were shown for Gaussian random matrices separating regions where recovery fails or succeeds with high probability [31, 28, 32]. These results build on previous work in pure mathematics by Affentranger and Schneider [2] on randomly projected polytopes.

2.5 Developments in Computer Science

In computer science the related area is usually addressed as the *heavy hitters* detection or *sketching*. Here one is interested not only in recovering signals (such as huge data streams on the internet) from vastly undersampled data, but one requires sublinear runtime in the signal length N of the recovery algorithm. This is no impossibility as one only has to report the locations and values of the non-zero (most significant) coefficients of the sparse vector. Quite remarkably sublinear algorithms are available for sparse Fourier re-

covery [48]. Such algorithms use ideas from *group testing* which date back to World War II, when Dorfman [34] invented an efficient method for detecting draftees with syphilis.

In sketching algorithms from computer science one actually designs the matrix and the fast algorithm simultaneously [22, 50]. More recently, *bipartite expander graphs* have been successfully used in order to construct good compressed sensing matrices together with associated fast reconstruction algorithms [5].

3 Mathematical Modelling and Analysis

This section introduces the concept of sparsity and the recovery of sparse vectors from incomplete linear and non-adaptive measurements. In particular, an analysis of ℓ_1-minimization as a recovery method is provided. The *null-space property* and the *restricted isometry property* are introduced and it is shown that they ensure robust sparse recovery. It is actually difficult to show these properties for deterministic matrices and the optimal number m of measurements, and the major breakthrough in compressive sensing results is obtained for random matrices. Examples of several types of random matrices which ensure sparse recovery are given, such as Gaussian, Bernoulli and partial random Fourier matrices.

3.1 Preliminaries and Notation

This exposition mostly treats complex vectors in $\mathbb{C}^N$ although sometimes the considerations will be restricted to the real case $\mathbb{R}^N$. The ℓ_p-norm of a vector $x \in \mathbb{C}^N$ is defined as

$$\begin{aligned} \|x\|_p &:= \left(\sum_{j=1}^{N} |x_j|^p \right)^{1/p}, \quad 0 < p < \infty, \\ \|x\|_\infty &:= \max_{j=1,\dots,N} |x_j|. \end{aligned} \tag{3.1}$$

For $1 \le p \le \infty$, it is indeed a norm while for $0 < p < 1$ it is only a quasi-norm. When emphasizing the norm the term ℓ_p^N is used instead of $\mathbb{C}^N$ or $\mathbb{R}^N$. The unit ball in ℓ_p^N is $B_p^N = \{x \in \mathbb{C}^N, \|x\|_p \le 1\}$. The operator norm of a matrix $A \in \mathbb{C}^{m \times N}$ from ℓ_p^N to ℓ_p^m is denoted

$$\|A\|_{p \to p} = \max_{\|x\|_p = 1} \|Ax\|_p. \tag{3.2}$$

In the important special case $p = 2$, the operator norm is the maximal singular value $\sigma_{\max}(A)$ of A.

For a subset $T \subset \{1, \dots, N\}$ we denote by $x_T \in \mathbb{C}^N$ the vector which coincides with $x \in \mathbb{C}^N$ on the entries in T and is zero outside T. Similarly, A_T denotes the column submatrix of A corresponding to the columns indexed by T. Further, $T^c = \{1, \dots, N\} \setminus T$ denotes the complement of T and $\#T$ or $|T|$ indicate the cardinality of T. The kernel of a matrix A is denoted by $\ker A = \{x, Ax = 0\}$.

3.2 Sparsity and Compression

Compressive Sensing is based on the empirical observation that many types of real-world signals and images have a sparse expansion in terms of a suitable basis or frame, for instance a wavelet expansion. This means that the expansion has only a small number of significant terms, or in other words, that the coefficient vector can be well-approximated with one having only a small number of nonvanishing entries.

The support of a vector x is denoted $\operatorname{supp}(x) = \{j : x_j \neq 0\}$, and

$$\|x\|_0 := |\operatorname{supp}(x)|.$$

It has become common to call $\|\cdot\|_0$ the ℓ_0-norm, although it is not even a quasi-norm. A vector x is called *k-sparse* if $\|x\|_0 \leq k$. For $k \in \{1, 2, \dots, N\}$,

$$\Sigma_k := \{x \in \mathbb{C}^N : \|x\|_0 \leq k\}$$

denotes the set of k-sparse vectors. Furthermore, the *best k-term approximation error* of a vector $x \in \mathbb{C}^N$ in ℓ_p is defined as

$$\sigma_k(x)_p = \inf_{z \in \Sigma_k} \|x - z\|_p.$$

If $\sigma_k(x)$ decays quickly in k then x is called *compressible*. Indeed, in order to compress x one may simply store only the k largest entries. When reconstructing x from its compressed version the nonstored entries are simply set to zero, and the reconstruction error is $\sigma_k(x)_p$. It is emphasized at this point that the procedure of obtaining the compressed version of x is *adaptive* and *nonlinear* since it requires the search of the largest entries of x in absolute value. In particular, the location of the non-zeros is a nonlinear type of information.

The *best k-term approximation* of x can be obtained using the nonincreasing rearrangement $r(x) = (|x_{i_1}|, \dots, |x_{i_N}|)^T$, where i_j denotes a permutation of the indexes such that $|x_{i_j}| \geq |x_{i_{j+1}}|$ for $j = 1, \dots, N-1$. Then

it is straightforward to check that

$$\sigma_k(x)_p := \left(\sum_{j=k+1}^{N} r_j(x)^p \right)^{1/p}, \quad 0 < p < \infty.$$

and the vector $x_{[k]}$ derived from x by setting to zero all the $N-k$ smallest entries in absolute value is the *best k-term approximation*,

$$x_{[k]} = \arg\min_{z \in \Sigma_k} \|x - z\|_p,$$

for any $0 < p \leq \infty$.

The next lemma states essentially that ℓ_q-balls with small q (ideally $q \leq 1$) are good models for compressible vectors.

Lemma 3.1. *Let $0 < q < p \leq \infty$ and set $r = \frac{1}{q} - \frac{1}{p}$. Then*

$$\sigma_k(x)_p \leq k^{-r}, \quad k = 1, 2, \ldots, N \quad \text{for all } x \in B_q^N.$$

Proof. Let T be the set of indeces of the k-largest entries of x in absolute value. The nonincreasing rearrangement satisfies $|r_k(x)| \leq |x_j|$ for all $j \in T$, and therefore

$$k r_k(x)^q \leq \sum_{j \in T} |x_j|^q \leq \|x\|_q^q \leq 1.$$

Hence, $r_k(x) \leq k^{-\frac{1}{q}}$. Therefore

$$\sigma_k(x)_p^p = \sum_{j \notin T} |x_j|^p \leq \sum_{j \notin T} r_k(x)^{p-q} |x_j|^q \leq k^{-\frac{p-q}{q}} \|x\|_q^q \leq k^{-\frac{p-q}{q}},$$

which implies $\sigma_k(x)_p \leq k^{-r}$. ■

3.3 Compressive Sensing

The above outlined adaptive strategy of compressing a signal x by only keeping its largest coefficients is certainly valid when full information on x is available. If, however, the signal first has to be acquired or measured by a somewhat costly or lengthy procedure then this seems to be a waste of resources: At first, large efforts are made to acquire the full signal and then most of the information is thrown away when compressing it. One may ask whether it is possible to obtain more directly a compressed version of the signal by taking only a small amount of *linear and nonadaptive* measurements. Since one does not know a priori the large coefficients, this seems a

daunting task at first sight. Quite surprisingly, compressive sensing nevertheless predicts that reconstruction from vastly undersampled nonadaptive measurements is possible — even by using efficient recovery algorithms.

Taking m linear measurements of a signal $x \in \mathbb{C}^N$ corresponds to applying a matrix $A \in \mathbb{C}^{m \times N}$ — the *measurement matrix* —

$$y = Ax. \tag{3.3}$$

The vector $y \in \mathbb{C}^m$ is called the *measurement vector*. The main interest is in the vastly undersampled case $m \ll N$. Without further information, it is, of course, impossible to recover x from y since the linear system (3.3) is highly underdetermined, and has therefore infinitely many solutions. However, if the additional assumption that the vector x is k-sparse is imposed, then the situation dramatically changes as will be outlined.

The approach for a recovery procedure that probably comes first to mind is to search for the sparsest vector x which is consistent with the measurement vector $y = Ax$. This leads to solving the *ℓ_0-miminization problem*

$$\min \|z\|_0 \quad \text{subject to } Az = y. \tag{3.4}$$

Unfortunately, this combinatorial minimization problem is NP–hard in general [61, 64]. In other words, an algorithm that solves (3.4) for *any* matrix A and *any* right hand side y is necessarily computationally intractable. Therefore, essentially two practical and tractable alternatives to (3.4) have been proposed in the literature: convex relaxation leading to ℓ_1-minimization — also called basis pursuit [18] — and greedy algorithms, such as various matching pursuits [91, 90]. Quite surprisingly for both types of approaches various recovery results are available, which provide conditions on the matrix A and on the sparsity $\|x\|_0$ such that the recovered solution coincides with the original x, and consequently also with the solution of (3.4). This is no contradiction to the NP–hardness of (3.4) since these results apply only to a subclass of matrices A and right-hand sides y.

The ℓ_1-minimization approach considers the solution of

$$\min \|z\|_1 \quad \text{subject to } Az = y, \tag{3.5}$$

which is a convex optimization problem and can be seen as a convex relaxation of (3.4). Various efficient convex optimization techniques apply for its solution [9]. In the real-valued case, (3.5) is equivalent to a linear program and in the complex-valued case it is equivalent to a second order cone program. Therefore standard software applies for its solution — although

algorithms which are specialized to (3.5) outperform such standard software, see Section 4.

The hope is, of course, that the solution of (3.5) coincides with the solution of (3.4) and with the original sparse vector x. Figure 3 provides an intuitive explanation why ℓ_1-minimization promotes sparse solutions. Here, $N = 2$ and $m = 1$, so one deals with a line of solutions $\mathcal{F}(y) = \{z : Az = y\}$ in $\mathbb{R}^2$. Except for pathological situations where $\ker A$ is parallel to one of the faces of the polytope B_1^2, there is a unique solution of the ℓ_1-minimization problem, which has minimal sparsity, i.e., only one nonzero entry.

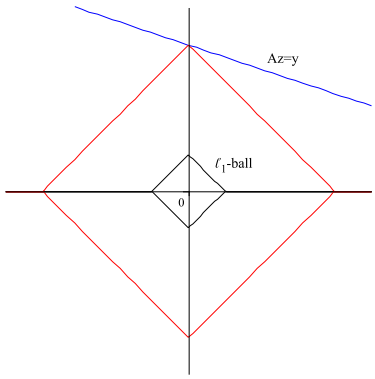

Figure 3: The ℓ_1-minimizer within the affine space of solutions of the linear system $Az = y$ coincides with a sparsest solution.

Recovery results in the next sections make rigorous the intuition that ℓ_1-minimization indeed promotes sparsity.

For sparse recovery via greedy algorithms we refer the reader to the literature [91, 90].

3.4 The Null Space Property

The null space property is fundamental in the analysis of ℓ_1-minimization.

Definition 3.1. *A matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the* null space property *(NSP) of order k with constant $\gamma \in (0, 1)$ if*

$$\|\eta_T\|_1 \leq \gamma \|\eta_{T^c}\|_1,$$

for all sets $T \subset \{1, \ldots, N\}$, $\#T \leq k$ and for all $\eta \in \ker A$.

The following sparse recovery result is based on this notion.

Theorem 3.2. *Let $A \in \mathbb{C}^{m\times N}$ be a matrix that satisfies the NSP of order k with constant $\gamma \in (0,1)$. Let $x \in \mathbb{C}^N$ and $y = Ax$ and let x^* be a solution of the ℓ_1-minimization problem* (3.5)*. Then*

$$\|x - x^*\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_k(x)_1. \tag{3.6}$$

In particular, if x is k-sparse then $x^* = x$.

Proof. Let $\eta = x^* - x$. Then $\eta \in \ker A$ and

$$\|x^*\|_1 \leq \|x\|_1$$

because x^* is a solution of the ℓ_1-minimization problem (3.5). Let T be the set of the k-largest entries of x in absolute value. One has

$$\|x^*_T\|_1 + \|x^*_{T^c}\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

It follows immediately from the triangle inequality that

$$\|x_T\|_1 - \|\eta_T\|_1 + \|\eta_{T^c}\|_1 - \|x_{T^c}\|_1 \leq \|x_T\|_1 + \|x_{T^c}\|_1.$$

Hence,

$$\|\eta_{T^c}\|_1 \leq \|\eta_T\|_1 + 2\|x_{T^c}\|_1 \leq \gamma\|\eta_{T^c}\|_1 + 2\sigma_k(x)_1,$$

or, equivalently,

$$\|\eta_{T^c}\|_1 \leq \frac{2}{1-\gamma}\sigma_k(x)_1. \tag{3.7}$$

Finally,

$$\|x - x^*\|_1 = \|\eta_T\|_1 + \|\eta_{T^c}\|_1 \leq (\gamma+1)\|\eta_{T^c}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_k(x)_1$$

and the proof is completed. ■

One can also show that if all k-sparse x can be recovered from $y = Ax$ using ℓ_1-minimization then necessarily A satisfies the NSP of order k with some constant $\gamma \in (0,1)$ [53, 21]. Therefore, the NSP is actually equivalent to sparse ℓ_1-recovery.

3.5 The Restricted Isometry Property

The NSP is somewhat difficult to show directly. The *restricted isometry property* (RIP) is easier to handle and it also implies stability under noise as stated below.

Definition 3.2. *The restricted isometry constant δ_k of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest number such that*

$$(1-\delta_k)\|z\|_2^2 \le \|Az\|_2^2 \le (1+\delta_k)\|z\|_2^2, \tag{3.8}$$

for all $z \in \Sigma_k$.

A matrix A is said to satisfy the *restricted isometry property* of order k with constant δ_k if $\delta_k \in (0,1)$. It is easily seen that δ_k can be equivalently defined as

$$\delta_k = \max_{T\subset\{1,\dots,N\},\#T\le k} \|A_T^*A_T - \mathrm{Id}\,\|_{2\to 2},$$

which means that *all* column submatrices of A with at most k columns are required to be well-conditioned. The RIP implies the NSP as shown in the following lemma.

Lemma 3.3. *Assume that $A \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $K = k+h$ with constant $\delta_K \in (0,1)$. Then A has the NSP of order k with constant $\gamma = \sqrt{\frac{k}{h}\frac{1+\delta_K}{1-\delta_K}}$.*

Proof. Let $\eta \in \mathcal{N} = \ker A$ and $T \subset \{1,\dots,N\}$, $\#T \le k$. Define $T_0 = T$ and $T_1, T_2, \dots, T_s$ to be disjoint sets of indexes of size at most h, associated to a nonincreasing rearrangement of the entries of $\eta \in \mathcal{N}$, i.e.,

$$|\eta_j| \le |\eta_i| \quad \text{for all } j \in T_\ell, i \in T_{\ell'}, \ell \ge \ell' \ge 1. \tag{3.9}$$

Note that $A\eta = 0$ implies $A\eta_{T_0\cup T_1} = -\sum_{j=2}^s A\eta_{T_j}$. Then, from the Cauchy–Schwarz inequality, the RIP, and the triangle inequality, the following sequence of inequalities is deduced,

$$\begin{aligned}
\|\eta_T\|_1 &\le \sqrt{k}\|\eta_T\|_2 \le \sqrt{k}\|\eta_{T_0\cup T_1}\|_2 \\
&\le \sqrt{\frac{k}{1-\delta_K}}\|A\eta_{T_0\cup T_1}\|_2 = \sqrt{\frac{k}{1-\delta_K}}\|A\eta_{T_2\cup T_3\cup\cdots\cup T_s}\|_2 \\
&\le \sqrt{\frac{k}{1-\delta_K}}\sum_{j=2}^s \|A\eta_{T_j}\|_2 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}}\sqrt{k}\sum_{j=2}^s \|\eta_{T_j}\|_2.
\end{aligned} \tag{3.10}$$

It follows from (3.9) that $|\eta_i| \le |\eta_\ell|$ for all $i \in T_{j+1}$ and $\ell \in T_j$. Taking the sum over $\ell \in T_j$ first and then the ℓ_2-norm over $i \in T_{j+1}$ yields

$$|\eta_i| \le h^{-1}\|\eta_{T_j}\|_1, \quad \text{and} \quad \|\eta_{T_{j+1}}\|_2 \le h^{-1/2}\|\eta_{T_j}\|_1.$$

Using the latter estimates in (3.10) gives

$$\|\eta_T\|_1 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}\frac{k}{h}}\sum_{j=1}^{s-1}\|\eta_{T_j}\|_1 \le \sqrt{\frac{1+\delta_K}{1-\delta_K}\frac{k}{h}}\|\eta_{T^c}\|_1, \tag{3.11}$$

and the proof is finished. ■

Taking $h = 2k$ above shows that $\delta_{3k} < 1/3$ implies $\gamma < 1$. By Theorem 3.2, recovery of all k-sparse vectors by ℓ_1-minimization is then guaranteed. Additionally, stability in ℓ_1 is also ensured. The next theorem shows that RIP implies also a bound on the reconstruction error in ℓ_2.

Theorem 3.4. *Assume $A \in \mathbb{C}^{m\times N}$ satisfies the RIP of order $3k$ with $\delta_{3k} < 1/3$. For $x \in \mathbb{C}^N$, let $y = Ax$ and x^* be the solution of the ℓ_1-minimization problem* (3.5)*. Then*

$$\|x - x^*\|_2 \le C\frac{\sigma_k(x)_1}{\sqrt{k}}$$

with $C = \frac{2}{1-\gamma}\left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right)$, $\gamma = \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}}$.

Proof. Similarly as in the proof of Lemma 3.3, denote $\eta = x^* - x \in \mathcal{N} = \ker A$, $T_0 = T$ the set of the $2k$-largest entries of η in absolute value, and T_j's of size at most k corresponding to the nonincreasing rearrangement of η. Then, using (3.10) and (3.11) with $h = 2k$ of the previous proof,

$$\|\eta_T\|_2 \le \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}}k^{-1/2}\|\eta_{T^c}\|_1.$$

From the assumption $\delta_{3k} < 1/3$ it follows that $\gamma := \sqrt{\frac{1+\delta_{3k}}{2(1-\delta_{3k})}} < 1$. Lemma 3.1 and Lemma 3.3 yield

$$\begin{aligned}\|\eta_{T^c}\|_2 &= \sigma_{2k}(\eta)_2 \le (2k)^{-\frac{1}{2}}\|\eta\|_1 = (2k)^{-1/2}\left(\|\eta_T\|_1 + \|\eta_{T^c}\|_1\right)\\ &\le (2k)^{-1/2}\left(\gamma\|\eta_{T^c}\|_1 + \|\eta_{T^c}\|_1\right) \le \frac{\gamma+1}{\sqrt{2}}k^{-1/2}\|\eta_{T^c}\|_1.\end{aligned}$$

Since T is the set of $2k$-largest entries of η in absolute value, it holds

$$\|\eta_{T^c}\|_1 \le \|\eta_{(\operatorname{supp} x_{[2k]})^c}\|_1 \le \|\eta_{(\operatorname{supp} x_{[k]})^c}\|_1, \tag{3.12}$$

where $x_{[k]}$ is the best k-term approximation to x. The use of this latter estimate, combined with inequality (3.7), finally gives

$$\begin{aligned} \|x - x^*\|_2 &\leq \|\eta_T\|_2 + \|\eta_{T^c}\|_2 \\ &\leq \left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right) k^{-1/2} \|\eta_{T^c}\|_1 \\ &\leq \frac{2}{1-\gamma}\left(\frac{\gamma+1}{\sqrt{2}} + \gamma\right) k^{-1/2} \sigma_k(x)_1. \end{aligned}$$

This concludes the proof. ■

The restricted isometry property implies also robustness under noise on the measurements. This fact was first noted in [16, 15]. We present the so far best known result [43, 45] concerning recovery using a noise aware variant of ℓ_1-minimization without proof.

Theorem 3.5. *Assume that the restricted isometry constant δ_{2k} of the matrix $A \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{2k} < \frac{2}{3+\sqrt{7/4}} \approx 0.4627. \tag{3.13}$$

Then the following holds for all $x \in \mathbb{C}^N$. Let noisy measurements $y = Ax + e$ be given with $\|e\|_2 \leq \eta$. Let x^* be the solution of

$$\min \|z\|_1 \quad \text{subject to } \|Az - y\|_2 \leq \eta. \tag{3.14}$$

Then

$$\|x - x^*\|_2 \leq C_1 \eta + C_2 \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for some constants $C_1, C_2 > 0$ that depend only on δ_{2k}.

3.6 Coherence

The *coherence* is a by now classical way of analyzing the recovery abilities of a measurement matrix [29, 91]. For a matrix $A = (a_1|a_2|\cdots|a_N) \in \mathbb{C}^{m\times N}$ with normalized columns, $\|a_\ell\|_2 = 1$, it is defined as

$$\mu := \max_{\ell \neq k} |\langle a_\ell, a_k \rangle|.$$

Applying Gershgorin's disc theorem [54] to $A_T^* A_T - I$ with $\#T = k$ shows that

$$\delta_k \leq (k-1)\mu. \tag{3.15}$$

Several explicit examples of matrices are known which have small coherence $\mu = \mathcal{O}(1/\sqrt{m})$. A simple one is the concatenation $A = (I|F) \in \mathbb{C}^{m\times 2m}$ of the identity matrix and the unitary Fourier matrix $F \in \mathbb{C}^{m\times m}$ with entries $F_{j,k} = m^{-1/2}e^{2\pi ijk/m}$. It is easily seen that $\mu = 1/\sqrt{m}$ in this case. Furthermore, [82] gives several matrices $A \in \mathbb{C}^{m\times m^2}$ with coherence $\mu = 1/\sqrt{m}$. In all these cases, $\delta_k \leq C\frac{k}{\sqrt{m}}$. Combining this estimate with the recovery results for ℓ_1-minimization above shows that all k-sparse vectors x can be (stably) recovered from $y = Ax$ via ℓ_1-minimization provided

$$m \geq C'k^2. \tag{3.16}$$

At first sight one might be satisfied with this condition since if k is very small compared to N then still m might be chosen smaller than N and all k-sparse vectors can be recovered from the undersampled measurements $y = Ax$. Although this is great news for a start, one might nevertheless hope that (3.16) can be improved. In particular, one may expect that actually a linear scaling of m in k should be enough to guarantee sparse recovery by ℓ_1-minimization. The existence of matrices, which indeed provide recovery conditions of the form $m \geq Ck\log^\alpha(N)$ (or similar) with some $\alpha \geq 1$, is shown in the next section. Unfortunately, such results cannot be shown using simply the coherence because of the general lower bound [82]

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}} \sim \frac{1}{\sqrt{m}} \qquad (N \text{ sufficiently large}).$$

In particular, it is not possible to overcome the "quadratic bottleneck" in (3.16) by using Gershgorin's theorem or Riesz-Thorin interpolation between $\|\cdot\|_{1\to 1}$ and $\|\cdot\|_{\infty\to\infty}$, see also [75, 81]. In order to improve on (3.16) one has to take into account also cancellations in the Gramian $A_T^*A_T - I$, and this task seems to be quite difficult using deterministic methods. Therefore, it will not come as a surprise that the major breakthrough in compressive sensing was obtained with random matrices. It is indeed easier to deal with cancellations in the Gramian using probabilistic techniques.

3.7 RIP for Gaussian and Bernoulli Random Matrices

Optimal estimates for the RIP constants in terms of the number m of measurement matrices can be obtained for Gaussian, Bernoulli or more general subgaussian random matrices.

Let X be a random variable. Then one defines a random matrix $A = A(\omega)$, $\omega \in \Omega$, as the matrix whose entries are independent realizations of

X, where $(\Omega, \Sigma, \mathbb{P})$ is their common probability space. One assumes further that for any $x \in \mathbb{R}^N$ we have the identity $\mathbb{E}\|Ax\|_2^2 = \|x\|_2^2$, $\mathbb{E}$ denoting expectation.

The starting point for the simple approach in [4] is a concentration inequality of the form

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \delta\|x\|_2^2\right) \leq 2e^{-c_0\delta^2 m}, \quad 0 < \delta < 1, \tag{3.17}$$

where $c_0 > 0$ is some constant.

The two most relevant examples of random matrices which satisfy the above concentration are the following.

1. **Gaussian Matrices.** Here the entries of A are chosen as i.i.d. Gaussian random variables with expectation 0 and variance $1/m$. As shown in [1] Gaussian matrices satisfy (3.17).

2. **Bernoulli Matrices** The entries of a Bernoulli matrices are independent realizations of $\pm 1/\sqrt{m}$ Bernoulli random variables, that is, each entry takes the value $+1/\sqrt{m}$ or $-1/\sqrt{m}$ with equal probability. Bernoulli matrices also satisfy the concentration inequality (3.17) [1].

Based on the concentration inequality (3.17) the following estimate on RIP constants can be shown [4, 16, 63].

Theorem 3.6. *Assume $A \in \mathbb{R}^{m\times N}$ to be a random matrix satisfying the concentration property* (3.17)*. Then there exists a constant C depending only on c_0 such that the restricted isometry constant of A satisfies $\delta_k \leq \delta$ with probability exceeding $1 - \varepsilon$ provided*

$$m \geq C\delta^{-2}(k\log(N/m) + \log(\varepsilon^{-1})).$$

Combining this RIP estimate with the recovery results for ℓ_1-minimization shows that all k-sparse vectors $x \in \mathbb{C}^N$ can be stably recovered from a random draw of A satisfying (3.17) with high probability provided

$$m \geq Ck\log(N/m). \tag{3.18}$$

Up to the log-factor this provides the desired linear scaling of the number m of measurements with respect to the sparsity k. Furthermore, as shown in Section 3.9 below, condition (3.18) cannot be further improved; in particular, the log-factor cannot be removed.

It is useful to observe that the concentration inequality is invariant under unitary transforms. Indeed, suppose that z is not sparse with respect to the

canonical basis but with respect to a different orthonormal basis. Then $z = Ux$ for a sparse x and a unitary matrix $U \in \mathbb{C}^{N\times N}$. Applying the measurement matrix A yields

$$Az = AUx,$$

so that this situation is equivalent to working with the new measurement matrix $A' = AU$ and again sparsity with respect to the canonical basis. The crucial point is that A' satisfies again the concentration inequality (3.17) once A does. Indeed, choosing $x = U^{-1}x'$ and using unitarity gives

$$\begin{aligned}&\mathbb{P}\left(\left|\|AUx\|_2^2 - \|x\|_2^2\right| \geq \delta\|x\|_{\ell_2^N}^2\right) = \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|U^{-1}x'\|_2^2\right| \geq \delta\|U^{-1}x'\|_{\ell_2^N}^2\right)\\ &= \mathbb{P}\left(\left|\|Ax'\|_2^2 - \|x'\|_2^2\right| \geq \delta\|x'\|_{\ell_2^N}^2\right) \leq 2e^{-c_0\delta^{-2}m}.\end{aligned}$$

Hence, Theorem 3.6 also applies to $A' = AU$. This fact is sometimes referred to as the *universality* of the Gaussian or Bernoulli random matrices. It does not matter in which basis the signal x is actually sparse. At the coding stage, where one takes random measurements $y = Az$, knowledge of this basis is not even required. Only the decoding procedure needs to know U.

3.8 Random Partial Fourier Matrices

While Gaussian and Bernoulli matrices provide optimal conditions for the minimal number of required samples for sparse recovery, they are of somewhat limited use for practical applications for several reasons. Often the application imposes physical or other constraints on the measurement matrix, so that assuming A to be Gaussian may not be justifiable in practice. One usually has only limited freedom to inject randomness in the measurements. Furthermore, Gaussian or Bernoulli matrices are not structured so there is no fast matrix-vector multiplication available which may speed up recovery algorithms, such as the ones described in Section 4. Thus, Gaussian random matrices are not applicable in large scale problems.

A very important class of structured random matrices that overcomes these drawbacks are random partial Fourier matrices, which were also the object of study in the very first papers on compressive sensing [13, 16, 72, 73]. A random partial Fourier matrix $A \in \mathbb{C}^{m\times N}$ is derived from the discrete Fourier matrix $F \in \mathbb{C}^{N\times N}$ with entries

$$F_{j,k} = \frac{1}{\sqrt{N}} e^{2\pi jk/N},$$

by selecting m rows uniformly at random among all N rows. Taking measurements of a sparse $x \in \mathbb{C}^N$ corresponds then to observing m of the entries of its discrete Fourier transform $\hat{x} = Fx$. It is important to note that the fast Fourier transform may be used to compute matrix-vector multiplications with A and A^* with complexity $\mathcal{O}(N \log(N))$. The following theorem concerning the RIP constant was proven in [75], and improves slightly on the results in [78, 16, 73].

Theorem 3.7. *Let $A \in \mathbb{C}^{m \times N}$ be the random partial Fourier matrix as just described. Then the restricted isometry constant of the rescaled matrix $\sqrt{\frac{N}{m}}A$ satisfy $\delta_k \leq \delta$ with probability at least $1 - N^{-\gamma \log^3(N)}$ provided*

$$m \geq C\delta^{-2} k \log^4(N). \tag{3.19}$$

The constants $C, \gamma > 1$ are universal.

Combining this estimate with the ℓ_1-minimization results above shows that recovery with high probability can be ensured for all k-sparse x provided

$$m \geq Ck \log^4(N).$$

The plots in Figure 1 illustrate an example of successful recovery from partial Fourier measurements.

The proof of the above theorem is not straightforward and involves Dudley's inequality as a main tool [78, 75]. Compared to the recovery condition (3.18) for Gaussian matrices, we suffer a higher exponent at the log-factor, but the linear scaling of m in k is preserved. Also a nonuniform recovery result for ℓ_1-minimization is available [13, 72, 75], which states that each k-sparse x can be recovered using a random draw of the random partial Fourier matrix A with probability at least $1 - \varepsilon$ provided $m \geq Ck \log(N/\varepsilon)$. The difference to the statement in Theorem 3.7 is that, for each sparse x, recovery is ensured with high probability for a new random draw of A. It does not imply the existence of a matrix which allows recovery of *all* k-sparse x simultaneously. The proof of such recovery results do not make use of the restricted isometry property or the null space property.

One may generalize the above results to a much broader class of structured random matrices which arise from random sampling in bounded orthonormal systems. The interested reader is referred to [72, 73, 75].

Another class of structured random matrices, for which recovery results are known, consist of partial random circulant and Toeplitz matrices. These correspond to subsampling the convolution of x with a random vector b at

m fixed (deterministic) entries. The reader is referred to [74, 75] for detailed information. It is only noted that a good estimate for the RIP constants for such types of random matrices is still an open problem. Further types of random measurement matrices are discussed in [69, 93].

3.9 Compressive Sensing and Gelfand Widths

In this section a quite general viewpoint is taken. The question is investigated how well any measurement matrix and any reconstruction method — in this context usually called the *decoder* — may perform. This leads to the study of *Gelfand widths*, already mentioned in Section 2.3. The corresponding analysis will allow to draw the conclusion that Gaussian random matrices in connection with ℓ_1-minimization provide optimal performance guarantees.

Following the tradition of the literature in this context, only the real-valued case will be treated. The complex-valued case is easily deduced from the real-case by identifying $\mathbb{C}^N$ with $\mathbb{R}^{2N}$ and by corresponding norm equivalences of ℓ_p-norms.

The measurement matrix $A \in \mathbb{R}^{m\times N}$ is here also referred to as the *encoder*. The set $\mathcal{A}_{m,N}$ denotes all possible encoder / decoder pairs (A, Δ) where $A \in \mathbb{R}^{m\times N}$ and $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ is any (nonlinear) function. Then, for $1 \leq k \leq N$, the reconstruction errors over subsets $K \subset \mathbb{R}^N$, where $\mathbb{R}^N$ is endowed with a norm $\|\cdot\|_X$, are defined as

$$\begin{aligned} \sigma_k(K)_X &:= \sup_{x\in K} \sigma_k(x)_X, \\ E_m(K, X) &:= \inf_{(A,\Delta)\in A_{m,N}} \sup_{x\in K} \|x - \Delta(Ax)\|_X. \end{aligned}$$

In words, $E_n(K, X)$ is the worst reconstruction error for the best pair of encoder / decoder. The goal is to find the largest k such that

$$E_m(K, X) \leq C_0 \sigma_k(K)_X.$$

Of particular interest for compressive sensing are the unit balls $K = B_p^N$ for $0 < p \leq 1$ and $X = \ell_2^N$ because the elements of B_p^N are well-approximated by sparse vectors due to Lemma 3.1. The proper estimate of $E_m(K, X)$ turns out to be linked to the geometrical concept of *Gelfand width*.

Definition 3.3. *Let K be a compact set in a normed space X. Then the* Gelfand width *of K of order m is*

$$d^m(K, X) := \inf_{\substack{Y \subset X \\ \operatorname{codim}(Y) \leq m}} \sup\{\|x\|_X : x \in K \cap Y\},$$

where the infimum is over all linear subspaces Y of X of codimension less or equal to m.

The following fundamental relationship between $E_m(K, X)$ and the Gelfand widths holds.

Proposition 3.8. *Let $K \subset \mathbb{R}^N$ be a closed compact set such that $K = -K$ and $K + K \subset C_0 K$ for some constant C_0. Let $X = (\mathbb{R}^N, \|\cdot\|_X)$ be a normed space. Then*

$$d^m(K, X) \leq E_m(K, X) \leq C_0 d^m(K, X).$$

Proof. For a matrix $A \in \mathbb{R}^{m \times N}$, the subspace $Y = \ker A$ has codimension less or equal to m. Conversely, to any subspace $Y \subset \mathbb{R}^N$ of codimension less or equal to m, a matrix $A \in \mathbb{R}^{m \times N}$ can be associated, the rows of which form a basis for $Y^\perp$. This identification yields

$$d^m(K, X) = \inf_{A \in \mathbb{R}^{m \times N}} \sup\{\|\eta\|_X : \eta \in \ker A \cap K\}.$$

Let (A, Δ) be an encoder / decoder pair in $\mathcal{A}_{m,N}$ and $z = \Delta(0)$. Denote $Y = \ker(A)$. Then with $\eta \in Y$ also $-\eta \in Y$, and either $\|\eta - z\|_X \geq \|\eta\|_X$ or $\| - \eta - z\|_X \geq \|\eta\|_X$. Indeed, if both inequalities were false then

$$\|2\eta\|_X = \|\eta - z + z + \eta\|_X \leq \|\eta - z\|_X + \| - \eta - z\|_X < 2\|\eta\|_X,$$

a contradiction. Since $K = -K$ it follows that

$$\begin{aligned} d^m(K, X) &= \inf_{A \in \mathbb{R}^{m \times N}} \sup\{\|\eta\|_X : \eta \in Y \cap K\} \leq \sup_{\eta \in Y \cap K} \|\eta - z\|_X \\ &= \sup_{\eta \in Y \cap K} \|\eta - \Delta(A\eta)\|_X \leq \sup_{x \in K} \|x - \Delta(Ax)\|_X. \end{aligned}$$

Taking the infimum over all $(A, \Delta) \in \mathcal{A}_{m,N}$ yields

$$d^m(K, X) \leq E_m(K, X).$$

To prove the converse inequality, choose an optimal Y such that

$$d^m(K, X) = \sup\{\|x\|_X : x \in Y \cap K\}.$$

(An optimal subspace Y always exists [60].) Let A be a matrix whose rows form a basis for $Y^\perp$. Denote the affine solution space $\mathcal{F}(y) := \{x : Ax = y\}$. One defines then a decoder as follows. If $\mathcal{F}(y) \cap K \neq \emptyset$ then choose some

$\bar{x}(y) \in \mathcal{F}(y)$ and set $\Delta(y) = \bar{x}(y)$. If $\mathcal{F}(y) \cap K = \emptyset$ then $\Delta(y) \in \mathcal{F}(y)$. The following chain of inequalities is then deduced

$$\begin{aligned} E_m(K,X) &\leq \sup_y \sup_{x,x' \in \mathcal{F}(y) \cap K} \|x - x'\|_X \\ &\leq \sup_{\eta \in C_0(Y \cap K)} \|\eta\|_X \leq C_0 d^m(K,X), \end{aligned}$$

which concludes the proof. ■

The assumption $K + K \subset C_0 K$ clearly holds for norm balls with $C_0 = 2$ and for quasi-norm balls with some $C_0 \geq 2$. The next theorem provides a two-sided estimate of the Gelfand widths $d^m(B_p^N, \ell_2^N)$ [44, 27, 95]. Note that the case $p = 1$ was considered much earlier in [56, 47, 44].

Theorem 3.9. *Let $0 < p \leq 1$. There exist universal constants $C_p, D_p > 0$ such that the Gelfand widths $d^m(B_p^N, \ell_2^N)$ satisfy*

$$\begin{aligned} C_p \min\left\{1, \frac{\ln(2N/m)}{m}\right\}^{1/p-1/2} &\leq d^m(B_p^N, \ell_2^N) \\ &\leq D_p \min\left\{1, \frac{\ln(2N/m)}{m}\right\}^{1/p-1/2} \end{aligned} \tag{3.20}$$

Combining Proposition 3.8 and Theorem 3.9 gives in particular, for large m,

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq \tilde{D}_1 \sqrt{\frac{\log(2N/m)}{m}}. \tag{3.21}$$

This estimate implies a lower estimate for the minimal number of required samples which allows for approximate sparse recovery using any measurement matrix and any recovery method whatsoever. The reader should compare the next statement with Theorem 3.4.

Corollary 3.10. *Suppose that $A \in \mathbb{R}^{m \times N}$ and $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ such that*

$$\|x - \Delta(Ax)\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}}$$

for all $x \in B_1^N$ and some constant $C > 0$. Then necessarily

$$m \geq C' k \log(2N/m). \tag{3.22}$$

Proof. Since $\sigma_k(x)_1 \leq \|x\|_1 \leq 1$, the assumption implies $E_m(B_1^N, \ell_2^N) \leq Ck^{-1/2}$. The lower bound in (3.21) combined with Proposition 3.8 yields

$$\tilde{C}_1 \sqrt{\frac{\log(2N/m)}{m}} \leq E_m(B_1^N, \ell_2^N) \leq Ck^{-1/2}.$$

Consequently, $m \geq C'k \log(eN/m)$ as claimed. ■

In particular, the above lemma applies to ℓ_1-minimization and consequently $\delta_k \leq 0.4$ (say) for a matrix $A \in \mathbb{R}^{m \times N}$ implies $m \geq Ck \log(N/m)$. Therefore, the recovery results for Gaussian or Bernoulli random matrices with ℓ_1-minimization stated above are optimal.

It can also be shown that a stability estimate in the ℓ_1-norm of the form $\|x - \Delta(Ax)\|_1 \leq C\sigma_k(x)_1$ for all $x \in \mathbb{R}^N$ implies (3.22) as well [44, 24].

3.10 Applications

Compressive sensing can be potentially used in all applications where the task is the reconstruction of a signal or an image from linear measurements, while taking many of those measurements – in particular, a complete set of measurements – is a costly, lengthy, difficult, dangerous, impossible or otherwise undesired procedure. Additionally, there should be reasons to believe that the signal is sparse in a suitable basis (or frame). Empirically, the latter applies to most types of signals.

In computerized tomography, for instance, one would like to obtain an image of the inside of a human body by taking X-ray images from different angles. Taking an almost complete set of images would expose the patient to a large and dangerous dose of radiation, so the amount of measurements should be as small as possible, and nevertheless guarantee a good enough image quality. Such images are usually nearly piecewise constant and therefore nearly sparse in the gradient, so there is a good reason to believe that compressive sensing is well applicable. And indeed, it is precisely this application that started the investigations on compressive sensing in the seminal paper [13].

Also radar imaging seems to be a very promising application of compressive sensing techniques [38, 83]. One is usually monitoring only a small number of targets, so that sparsity is a very realistic assumption. Standard methods for radar imaging actually also use the sparsity assumption, but only at the very end of the signal processing procedure in order to clean up the noise in the resulting image. Using sparsity systematically from the very

beginning by exploiting compressive sensing methods is therefore a natural approach. First numerical experiments in [38, 83] are very promising.

Further potential applications include wireless communication [86], astronomical signal and image processing [8], analog to digital conversion [93], camera design [35] and imaging [77].

4 Numerical Methods

The previous sections showed that ℓ_1-minimization performs very well in recovering sparse or approximately sparse vectors from undersampled measurements. In applications it is important to have fast methods for actually solving ℓ_1-minimization problems. Two such methods – the homotopy (LARS) method introduced in [68, 36] and iteratively reweighted least squares (IRLS) [23] – will be explained in more detail below.

As a first remark, the ℓ_1-minimization problem

$$\min \|x\|_1 \quad \text{subject to } Ax = y \tag{4.1}$$

is in the real case equivalent to the linear program

$$\min \sum_{j=1}^{2N} v_j \quad \text{subject to} \quad v \geq 0,\ (A|-A)v = y. \tag{4.2}$$

The solution x^* to (4.1) is obtained from the solution v^* of (4.2) via $x^* = (\mathrm{Id}\,|\,-\mathrm{Id})v^*$. Any linear programming method may therefore be used for solving (4.1). The simplex method as well as interior point methods apply in particular [65], and standard software may be used. (In the complex case, (4.1) is equivalent to a second order cone program (SOCP) and can also be solved with interior point methods.) However, such methods and software are of general purpose and one may expect that methods specialized to (4.1) outperform such existing standard methods. Moreover, standard software often has the drawback that one has to provide the full matrix rather than fast routines for matrix-vector multiplication which are available for instance in the case of partial Fourier matrices. In order to obtain the full performance of such methods one would therefore need to reimplement them, which is a daunting task because interior point methods usually require much fine tuning. On the contrary the two specialized methods described below are rather simple to implement and very efficient. Many more methods are available nowadays, including greedy methods, such as orthogonal matching pursuit [91], CoSaMP [90], and iterative hard thresholding [7, 39], which

may offer better complexity than standard interior point methods. Due to space limitations, however, only the two methods below are explained in detail.

4.1 The Homotopy Method

The homotopy method – or modified LARS – [68, 67, 36, 33] solves (4.1) in the real-valued case. One considers the ℓ_1-regularized least squares functionals

$$F_\lambda(x) = \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1, \quad x \in \mathbb{R}^N,\ \lambda > 0, \tag{4.3}$$

and its minimizer x_λ. When $\lambda = \hat{\lambda}$ is large enough then $x_{\hat{\lambda}} = 0$, and furthermore, $\lim_{\lambda\to 0} x_\lambda = x^*$, where x^* is the solution to (4.1). The idea of the homotopy method is to trace the solution x_λ from $x_{\hat{\lambda}} = 0$ to x^*. The crucial observation is that the solution path $\lambda \mapsto x_\lambda$ is piecewise linear, and it is enough to trace the endpoints of the linear pieces.

The minimizer of (4.3) can be characterized using the subdifferential, which is defined for a general convex function $F : \mathbb{R}^N \to \mathbb{R}$ at a point $x \in \mathbb{R}^N$ by

$$\partial F(x) = \{v \in \mathbb{R}^N, F(y) - F(x) \geq \langle v, y - x\rangle \ \text{ for all } y \in \mathbb{R}^N\}.$$

Clearly, x is a minimizer of F if and only if $0 \in \partial F(x)$. The subdifferential of F_λ is given by

$$\partial F_\lambda(x) = A^*(Ax - y) + \lambda\partial\|x\|_1$$

where the subdifferential of the ℓ_1-norm is given by

$$\partial\|x\|_1 = \{v \in \mathbb{R}^N : v_\ell \in \partial|x_\ell|, \ell = 1, \ldots, N\}$$

with the subdifferential of the absolute value being

$$\partial|z| = \begin{cases} \{\operatorname{sgn}(z)\}, & \text{if } z \neq 0, \\ [-1, 1] & \text{if } z = 0. \end{cases}$$

The inclusion $0 \in \partial F_\lambda(x)$ is equivalent to

$$(A^*(Ax - y))_\ell = \lambda \operatorname{sgn}(x_\ell) \qquad \text{if } x_\ell \neq 0, \tag{4.4}$$

$$|(A^*(Ax - y)_\ell| \leq \lambda \qquad \text{if } x_\ell = 0, \tag{4.5}$$

for all $\ell = 1, \ldots, N$.

As already mentioned above the homotopy method starts with $x^{(0)} = x_\lambda = 0$. By conditions (4.4) and (4.5) the corresponding λ can be chosen

as $\lambda = \lambda^{(0)} = \|A^*y\|_\infty$. In the further steps $j = 1, 2, \ldots$, the algorithm computes minimizers $x^{(1)}, x^{(2)}, \ldots$, and maintains an active (support) set T_j. Denote by

$$c^{(j)} = A^*(Ax^{(j-1)} - y)$$

the current residual vector.

Step 1: Let

$$\ell^{(1)} := \arg\max_{\ell=1,\ldots,N} |(A^*y)_\ell| = \arg\max_{\ell=1,\ldots,N} |c_\ell^{(1)}|.$$

One assumes here and also in the further steps that the maximum is attained at only one index ℓ. The case that the maximum is attained simultaneously at two or more indexes ℓ (which almost never happens) requires more complications that will not be covered here. The reader is referred to [36] for such details.

Now set $T_1 = \{\ell^{(1)}\}$. The vector $d \in \mathbb{R}^N$ describing the direction of the solution (homotopy) path has components

$$d_{\ell^{(1)}}^{(1)} = \|a_{\ell^{(1)}}\|_2^{-2} \operatorname{sgn}((Ay)_{\ell^{(1)}}) \text{ and } d_\ell^{(1)} = 0, \quad \ell \neq \ell^{(1)}.$$

The first linear piece of the solution path then takes the form

$$x = x(\gamma) = x^{(0)} + \gamma d^{(1)} = \gamma d^{(1)}, \quad \gamma \in [0, \gamma^{(1)}].$$

One verifies with the definition of $d^{(1)}$ that (4.4) is always satisfied for $x = x(\gamma)$ and $\lambda = \lambda(\gamma) = \lambda^{(0)} - \gamma$, $\gamma \in [0, \lambda^{(0)}]$. The next breakpoint is found by determining the maximal $\gamma = \gamma^{(1)} > 0$ for which (4.5) is still satisfied, which is

$$\gamma^{(1)} = \min_{\ell \neq \ell^{(1)}} \left\{ \frac{\lambda^{(0)} - c_\ell^{(1)}}{1 - (A^*A\, d^{(1)})_\ell}, \frac{\lambda^{(0)} + c_\ell^{(1)}}{1 + (A^*A\, d^{(1)})_\ell} \right\}. \tag{4.6}$$

Here, the minimum is taken only over positive arguments. Then $x^{(1)} = x(\gamma^{(1)}) = \gamma^{(1)} d^{(1)}$ is the next minimizer of F_λ for $\lambda = \lambda^{(1)} := \lambda^{(0)} - \gamma^{(1)}$. This $\lambda^{(1)}$ satisfies $\lambda^{(1)} = \|c^{(1)}\|_\infty$. Let $\ell^{(2)}$ be the index where the minimum in (4.6) is attained (where we again assume that the minimum is attained only at one index) and put $T_2 = \{\ell^{(1)}, \ell^{(2)}\}$.

Step j: Determine the new direction $d^{(j)}$ of the homotopy path by solving

$$A_{T_j}^* A_{T_j} d_{T_j}^{(j)} = \operatorname{sgn}(c_{T_j}^{(j)}), \tag{4.7}$$

which is a linear system of equations of size $|T_j| \times |T_j|$, $|T_j| \leq j$. Outside the components in T_j one sets $d_\ell^{(j)} = 0$, $\ell \notin T_j$. The next piece of the path is then given by

$$x(\gamma) = x^{(j-1)} + \gamma d^{(j)}, \quad \gamma \in [0, \gamma^{(j)}].$$

The maximal γ such that $x(\gamma)$ satisfies (4.5) is

$$\gamma_+^{(j)} = \min_{\ell \notin T_j} \left\{ \frac{\lambda^{(j-1)} - c_\ell^{(j)}}{1 - (A^* A d^{(j)})_\ell}, \frac{\lambda^{(j-1)} + c_\ell^{(j)}}{1 + (A^* A d^{(j)})_\ell} \right\}. \tag{4.8}$$

The maximal γ such that $x(\gamma)$ satisfies (4.4) is determined as

$$\gamma_-^{(j)} = \min_{\ell \in T_j} \{-x_\ell^{(j-1)} / d_\ell^{(j)}\}. \tag{4.9}$$

Both in (4.8) and (4.9) the minimum is taken only over positive arguments. The next breakpoint is given by $x^{(j+1)} = x(\gamma^{(j)})$ with $\gamma^{(j)} = \min\{\gamma_+^{(j)}, \gamma_-^{(j)}\}$. If $\gamma_+^{(j)}$ determines the minimum then the index $\ell_+^{(j)} \notin T_j$ providing the minimum in (4.8) is added to the active set, $T_{j+1} = T_j \cup \{\ell_+^{(j)}\}$. If $\gamma^{(j)} = \gamma_-^{(j)}$ then the index $\ell_-^{(j)} \in T_j$ is removed from the active set, $T_{j+1} = T_j \setminus \{\ell_-^{(j)}\}$. Further, one updates $\lambda^{(j)} = \lambda^{(j-1)} - \gamma^{(j)}$. By construction $\lambda^{(j)} = \|c^{(j)}\|_\infty$.

The algorithm stops when $\lambda^{(j)} = \|c^{(j)}\|_\infty = 0$, i.e., when the residual vanishes, and outputs $x^* = x^{(j)}$. Indeed, this happens after a finite number of steps. In [36] the following result was shown.

Theorem 4.1. *If in each step the minimum in* (4.8) *and* (4.9) *is attained in only one index ℓ, then the homotopy algorithm as described yields the minimizer of the ℓ_1-minimization problem* (4.1).

If the algorithm is stopped earlier at some iteration j then obviously it yields the minimizer of $F_\lambda = F_{\lambda^{(j)}}$. In particular, obvious stopping rules may also be used to solve the problems

$$\min \|x\|_1 \quad \text{subject to } \|Ax - y\|_2 \leq \epsilon \tag{4.10}$$

$$\text{or} \quad \min \|Ax - y\|_2 \quad \text{subject to } \|x\|_1 \leq \delta. \tag{4.11}$$

The first of these appears in (3.14), and the second is called the *lasso* (least absolute shrinkage and selection operator) [88].

The LARS (least angle regression) algorithm is a simple modification of the homotopy method, which only adds elements to the active set in each

step. So $\gamma_-^{(j)}$ in (4.9) is not considered. (Sometimes the homotopy method is therefore also called modified LARS.) Clearly, LARS is not guaranteed any more to yield the solution of (4.1). However, it is observed empirically — and can be proven rigorously in certain cases [33] — that often in sparse recovery problems, the homotopy method does never remove elements from the active set, so that in this case LARS and homotopy perform the same steps. It is a crucial point that if the solution of (4.1) is k-sparse and the homotopy method never removes elements then the solution is obtained after precisely k-steps. Furthermore, the most demanding computational part at step j is then the solution of the $j \times j$ linear system of equations (4.7). In conclusion, the homotopy and LARS methods are very efficient for sparse recovery problems.

4.2 Iteratively Reweighted Least Squares

This section is concerned with an iterative algorithm which, under the condition that A satisfies the NSP (see Definition 3.1), is guaranteed to reconstruct vectors with the same error estimate (3.6) as ℓ_1-minimization. Again we restrict the following discussion to the real case. This algorithm has a guaranteed linear rate of convergence which can even be improved to a superlinear rate with a small modification. First a brief introduction aims at shedding light on the basic principles of this algorithm and their interplay with sparse recovery and ℓ_1-minimization.

Denote $\mathcal{F}(y) = \{x : Ax = y\}$ and $\mathcal{N} = \ker A$. The starting point is the trivial observation that $|t| = \frac{t^2}{|t|}$ for $t \neq 0$. Hence, an ℓ_1-minimization can be recasted into a weighted ℓ_2-minimization, with the hope that

$$\arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^{N} |x_j| \approx \arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^{N} x_j^2 |x_j^*|^{-1},$$

as soon as x^* is the desired ℓ_1-norm minimizer. The advantage of the reformulation consists in the fact that minimizing the smooth quadratic function t^2 is an easier task than the minimization of the nonsmooth function $|t|$. However, the obvious drawbacks are that neither one disposes of x^* a priori (this is the vector one is interested to compute!) nor one can expect that $x_j^* \neq 0$ for all $j = 1, \ldots, N$, since one hopes for k-sparse solutions.

Suppose one has a good approximation w_j^n of $|(x_j^*)^2 + \epsilon_n^2|^{-1/2} \approx |x_j^*|^{-1}$, for some $\epsilon_n > 0$. One computes

$$x^{n+1} = \arg\min_{x \in \mathcal{F}(y)} \sum_{j=1}^{N} x_j^2 w_j^n, \tag{4.12}$$

and then updates $\epsilon_{n+1} \leq \epsilon_n$ by some rule to be specified later. Further, one sets

$$w_j^{n+1} = |(x_j^{n+1})^2 + \epsilon_{n+1}^2|^{-1/2}, \tag{4.13}$$

and iterates the process. The hope is that a proper choice of $\epsilon_n \to 0$ allows the iterative computation of an ℓ_1-minimizer. The next sections investigate convergence of this algorithm and properties of the limit.

4.2.1 Weighted ℓ_2-minimization

Suppose that the weight w is *strictly positive* which means that $w_j > 0$ for all $j \in \{1, \ldots, N\}$. Then $\ell_2(w)$ is a Hilbert space with the inner product

$$\langle u, v \rangle_w := \sum_{j=1}^{N} w_j u_j v_j. \tag{4.14}$$

Define

$$x^w := \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2,w}, \tag{4.15}$$

where $\|z\|_{2,w} = \langle z, z \rangle_w^{1/2}$. Because the $\|\cdot\|_{2,w}$-norm is strictly convex, the minimizer x^w is necessarily unique; it is characterized by the orthogonality conditions

$$\langle x^w, \eta \rangle_w = 0, \qquad \text{for all } \eta \in \mathcal{N}. \tag{4.16}$$

4.2.2 An iteratively re-weighted least squares algorithm (IRLS)

An IRLS algorithm appears for the first time in the Ph.D. thesis of Lawson in 1961 [57], in the form of an algorithm for solving uniform approximation problems. This iterative algorithm is now well-known in classical approximation theory as Lawson's algorithm. In [20] it is proved that it obeys a linear convergence rate. In the 1970s, extensions of Lawson's algorithm for ℓ_p-minimization, and in particular ℓ_1-minimization, were introduced. In signal analysis, IRLS was proposed as a technique to build algorithms for sparse signal reconstruction in [52]. The interplay of the NSP, ℓ_1-minimization, and a reweighted least square algorithm has been clarified only recently in the work [23].

The analysis of the algorithm (4.12) and (4.13) starts from the observation that

$$|t| = \min_{w>0} \frac{1}{2} \left(wt^2 + w^{-1} \right),$$

the minimum being attained for $w = \frac{1}{|t|}$. Inspired by this simple relationship, given a real number $\epsilon > 0$ and a weight vector $w \in \mathbb{R}^N$, with $w_j > 0$, $j = 1, \ldots, N$, one introduces the functional

$$\mathcal{J}(z, w, \epsilon) := \frac{1}{2} \sum_{j=1}^{N} \left(z_j^2 w_j + \epsilon^2 w_j + w_j^{-1} \right), \quad z \in \mathbb{R}^N. \tag{4.17}$$

The algorithm roughly described in (4.12) and (4.13) can be recast as an alternating method for choosing minimizers and weights based on the functional $\mathcal{J}$. To describe this more rigorously, recall that $r(z)$ denotes the nonincreasing rearrangement of a vector $z \in \mathbb{R}^N$.

Algorithm IRLS. Initialize by taking $w^0 := (1, \ldots, 1)$. Set $\epsilon_0 := 1$. Then recursively define, for $n = 0, 1, \ldots,$

$$x^{n+1} := \arg \min_{z \in \mathcal{F}(y)} \mathcal{J}(z, w^n, \epsilon_n) = \arg \min_{z \in \mathcal{F}(y)} \|z\|_{2,w^n} \tag{4.18}$$

and

$$\epsilon_{n+1} := \min \left\{ \epsilon_n, \frac{r_{K+1}(x^{n+1})}{N} \right\}, \tag{4.19}$$

where K is a fixed integer that will be specified later. Set

$$w^{n+1} := \arg \min_{w>0} \mathcal{J}(x^{n+1}, w, \epsilon_{n+1}). \tag{4.20}$$

The algorithm stops if $\epsilon_n = 0$; in this case, define $x^j := x^n$ for $j > n$. In general, the algorithm generates an infinite sequence $(x^n)_{n \in \mathbb{N}}$ of vectors.

Each step of the algorithm requires the solution of a weighted least squares problem. In matrix form

$$x^{n+1} = D_n^{-1} A^* (A D_n^{-1} A^*)^{-1} y, \tag{4.21}$$

where D_n is the $N \times N$ diagonal matrix the j-th diagonal entry of which is w_j^n. Once x^{n+1} is found, the weight w^{n+1} is given by

$$w_j^{n+1} = [(x_j^{n+1})^2 + \epsilon_{n+1}^2]^{-1/2}, \quad j = 1, \ldots, N. \tag{4.22}$$

4.2.3 Convergence properties

Lemma 4.2. *Set $L := \mathcal{J}(x^1, w^0, \epsilon_0)$. Then*

$$\|x^n - x^{n+1}\|_2^2 \leq 2L \left[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1}) \right].$$

Hence $(J(x^n, w^n, \epsilon_n))_{n\in\mathbb{N}}$ *is a monotonically decreasing sequence and*

$$\lim_{n\to\infty} \|x^n - x^{n+1}\|_2^2 = 0.$$

Proof. Note that $\mathcal{J}(x^n, w^n, \epsilon_n) \geq \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})$ for each $n = 1, 2, \ldots,$ and

$$L = \mathcal{J}(x^1, w^0, \epsilon_0) \geq \mathcal{J}(x^n, w^n, \epsilon_n) \geq (w_j^n)^{-1}, \quad j = 1, \ldots, N.$$

Hence, for each $n = 1, 2, \ldots,$ the following estimates hold,

$$\begin{aligned}
&2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^{n+1}, \epsilon_{n+1})] \\
&\geq 2[\mathcal{J}(x^n, w^n, \epsilon_n) - \mathcal{J}(x^{n+1}, w^n, \epsilon_n)] = \langle x^n, x^n\rangle_{w^n} - \langle x^{n+1}, x^{n+1}\rangle_{w^n} \\
&= \langle x^n + x^{n+1}, x^n - x^{n+1}\rangle_{w^n} = \langle x^n - x^{n+1}, x^n - x^{n+1}\rangle_{w^n} \\
&= \sum_{j=1}^{N} w_j^n (x_j^n - x_j^{n+1})^2 \geq L^{-1}\|x^n - x^{n+1}\|_2^2,
\end{aligned}$$

In the third line it is used that $\langle x^{n+1}, x^n - x^{n+1}\rangle_{w^n} = 0$ due to (4.16) since $x^n - x^{n+1}$ is contained in $\mathcal{N}$. ■

Moreover, if one assumes that $x^n \to \bar{x}$ and $\epsilon_n \to 0$, then, formally,

$$\mathcal{J}(x^n, w^n, \epsilon_n) \to \|\bar{x}\|_1.$$

Hence, one expects that this algorithm performs similar to ℓ_1-minimization. Indeed, the following convergence result holds.

Theorem 4.3. *Suppose* $A \in \mathbb{R}^{m\times N}$ *satisfies the NSP of order* K *with constant* $\gamma < 1$*. Use* K *in the update rule* (4.19)*. Then, for each* $y \in \mathbb{R}^m$*, the sequence* x^n *produced by the algorithm converges to a vector* $\bar{x}$*, with* $r_{K+1}(\bar{x}) = N \lim_{n\to\infty} \epsilon_n$ *and the following holds:*
(i) *If* $\epsilon = \lim_{n\to\infty} \epsilon_n = 0$*, then* $\bar{x}$ *is* K*-sparse; in this case there is therefore a unique* ℓ_1*-minimizer* x^**, and* $\bar{x} = x^*$*; moreover, we have, for* $k \leq K$*, and any* $z \in \mathcal{F}(y)$*,*

$$\|z - \bar{x}\|_1 \leq \frac{2(1+\gamma)}{1-\gamma}\sigma_k(z)_1; \tag{4.23}$$

(ii) *If* $\epsilon = \lim_{n\to\infty} \epsilon_n > 0$*, then* $\bar{x} = x^\epsilon := \arg\min_{z\in\mathcal{F}(y)} \sum_{j=1}^{N} \left(z_j^2 + \epsilon^2\right)^{1/2}$*;*
(iii) *In this last case, if* γ *satisfies the stricter bound* $\gamma < 1 - \frac{2}{K+2}$ *(or,*

equivalently, if $\frac{2\gamma}{1-\gamma} < K$*), then we have, for all* $z \in \mathcal{F}(y)$ *and any* $k < K - \frac{2\gamma}{1-\gamma}$*, that*

$$\|z - \bar{x}\|_1 \leq \tilde{c}\sigma_k(z)_1, \qquad \textit{with } \tilde{c} := \frac{2(1+\gamma)}{1-\gamma}\left[\frac{K-k+\frac{3}{2}}{K-k-\frac{2\gamma}{1-\gamma}}\right] \tag{4.24}$$

As a consequence, this case is excluded if $\mathcal{F}(y)$ *contains a vector of sparsity* $k < K - \frac{2\gamma}{1-\gamma}$*.*

Note that the approximation properties (4.23) and (4.24) are exactly of the same order as the one (3.6) provided by ℓ_1-minimization. However, in general, $\bar{x}$ is not necessarily an ℓ_1-minimizer, unless it coincides with a sparse solution.
The proof of this result is not included and the interested reader is referred to [23, 39] for the details.

4.2.4 Local linear rate of convergence

It is instructive to show a further result concerning the local rate of convergence of this algorithm, which again uses the NSP as well as the optimality conditions we introduced above. One assumes here that $\mathcal{F}(y)$ contains the k-sparse vector x^*. The algorithm produces a sequence x^n, which converges to x^*, as established above. One denotes the (unknown) support of the k-sparse vector x^* by T.

For now, one introduces an auxiliary sequence of error vectors $\eta^n \in \mathcal{N}$ via $\eta^n := x^n - x^*$ and

$$E_n := \|\eta^n\|_1 = \|x^* - x^n\|_1.$$

Theorem 4.3 guarantees that $E_n \to 0$ for $n \to \infty$. A useful technical result is reported next.

Lemma 4.4. *For any* $z, z' \in \mathbb{R}^N$*, and for any* j*,*

$$|\sigma_j(z)_1 - \sigma_j(z')_1| \leq \|z - z'\|_1, \tag{4.25}$$

while for any $J > j$*,*

$$(J-j)r_J(z) \leq \|z - z'\|_1 + \sigma_j(z')_1. \tag{4.26}$$

Proof. To prove (4.25), approximate z by a best j-term approximation $z'_{[j]} \in \Sigma_j$ of z' in ℓ_1. Then

$$\sigma_j(z)_1 \le \|z - z'_{[j]}\|_1 \le \|z - z'\|_1 + \sigma_j(z')_1,$$

and the result follows from symmetry. To prove (4.26), it suffices to note that $(J-j)\, r_J(z) \le \sigma_j(z)_1$. ■

The following theorem gives a bound on the rate of convergence of E_n to zero.

Theorem 4.5. *Assume A satisfies the NSP of order K with constant γ. Suppose that $k < K - \frac{2\gamma}{1-\gamma}$, $0 < \rho < 1$, and $0 < \gamma < 1 - \frac{2}{K+2}$ are such that*

$$\mu := \frac{\gamma(1+\gamma)}{1-\rho}\left(1 + \frac{1}{K+1-k}\right) < 1.$$

Assume that $\mathcal{F}(y)$ contains a k-sparse vector x^* and let $T = \operatorname{supp}(x^*)$. Let n_0 be such that

$$E_{n_0} \le R^* := \rho \min_{i \in T} |x_i^*|. \tag{4.27}$$

Then, for all $n \ge n_0$, we have

$$E_{n+1} \le \mu\, E_n. \tag{4.28}$$

Consequently, x^n converges to x^* exponentially.

Proof. The relation (4.16) with $w = w^n$, $x^w = x^{n+1} = x^* + \eta^{n+1}$, and $\eta = x^{n+1} - x^* = \eta^{n+1}$, gives

$$\sum_{i=1}^{N} (x_i^* + \eta_i^{n+1})\eta_i^{n+1} w_i^n = 0.$$

Rearranging the terms and using the fact that x^* is supported on T, one obtains

$$\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n = -\sum_{i \in T} x_i^* \eta_i^{n+1} w_i^n = -\sum_{i \in T} \frac{x_i^*}{[(x_i^n)^2 + \epsilon_n^2]^{1/2}} \eta_i^{n+1}. \tag{4.29}$$

The proof of the theorem is by induction. Assume that $E_n \le R^*$ has already been established. Then, for all $i \in T$,

$$|\eta_i^n| \le \|\eta^n\|_1 = E_n \le \rho |x_i^*|,$$

so that

$$\frac{|x_i^*|}{[(x_i^n)^2+\epsilon_n^2]^{1/2}} \le \frac{|x_i^*|}{|x_i^n|} = \frac{|x_i^*|}{|x_i^*+\eta_i^n|} \le \frac{1}{1-\rho}, \tag{4.30}$$

and hence (4.29) combined with (4.30) and the NSP gives

$$\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n \le \frac{1}{1-\rho}\|\eta_T^{n+1}\|_1 \le \frac{\gamma}{1-\rho}\|\eta_{T^c}^{n+1}\|_1$$

The Cauchy–Schwarz inequality combined with the above estimate yields

$$\begin{aligned} \|\eta_{T^c}^{n+1}\|_1^2 &\le \left(\sum_{i\in T^c} |\eta_i^{n+1}|^2 w_i^n\right)\left(\sum_{i\in T^c}[(x_i^n)^2+\epsilon_n^2]^{1/2}\right) \\ &= \left(\sum_{i=1}^{N} |\eta_i^{n+1}|^2 w_i^n\right)\left(\sum_{i\in T^c}[(\eta_i^n)^2+\epsilon_n^2]^{1/2}\right) \\ &\le \frac{\gamma}{1-\rho}\|\eta_{T^c}^{n+1}\|_1 \left(\|\eta^n\|_1 + N\epsilon_n\right). \end{aligned} \tag{4.31}$$

If $\eta_{T^c}^{n+1} = 0$, then $x_{T^c}^{n+1} = 0$. In this case x^{n+1} is k-sparse and the algorithm has stopped by definition; since $x^{n+1} - x^*$ is in the null space $\mathcal{N}$, which contains no k-sparse elements other than 0, one has already obtained the solution $x^{n+1} = x^*$. If $\eta_{T^c}^{n+1} \neq 0$, then cancelling the factor $\|\eta_{T^c}^{n+1}\|_1$ in (4.31) yields

$$\|\eta_{T^c}^{n+1}\|_1 \le \frac{\gamma}{1-\rho}\left(\|\eta^n\|_1 + N\epsilon_n\right),$$

and thus

$$\|\eta^{n+1}\|_1 = \|\eta_T^{n+1}\|_1 + \|\eta_{T^c}^{n+1}\|_1 \le (1+\gamma)\|\eta_{T^c}^{n+1}\|_1 \le \frac{\gamma(1+\gamma)}{1-\rho}\left(\|\eta^n\|_1 + N\epsilon_n\right). \tag{4.32}$$

Now, by (4.19) and (4.26) it follows

$$N\epsilon_n \le r_{K+1}(x^n) \le \frac{1}{K+1-k}(\|x^n - x^*\|_1 + \sigma_k(x^*)_1) = \frac{\|\eta^n\|_1}{K+1-k}, \tag{4.33}$$

since by assumption $\sigma_k(x^*)_1 = 0$. Together with (4.32) this yields the desired bound,

$$E_{n+1} = \|\eta^{n+1}\|_1 \le \frac{\gamma(1+\gamma)}{1-\rho}\left(1 + \frac{1}{K+1-k}\right)\|\eta^n\|_1 = \mu E_n.$$

In particular, since $\mu < 1$, one has $E_{n+1} \le R^*$, which completes the induction step. It follows that $E_{n+1} \le \mu E_n$ for all $n \ge n_0$. ■

4.2.5 Superlinear convergence promoting ℓ_τ-minimization for $\tau < 1$

The linear rate (4.28) can be improved significantly, by a very simple modification of the rule of updating the weight:

$$w_j^{n+1} = \left((x_j^{n+1})^2 + \epsilon_{n+1}^2\right)^{-\frac{2-\tau}{2}}, \quad j = 1, \ldots, N, \text{ for any } 0 < \tau < 1.$$

This corresponds to the substitution of the function $\mathcal{J}$ with

$$\mathcal{J}_\tau(z, w, \epsilon) := \frac{\tau}{2} \sum_{j=1}^{N} \left(z_j^2 w_j + \epsilon^2 w_j + \frac{2-\tau}{\tau} \frac{1}{w_j^{\frac{\tau}{2-\tau}}} \right),$$

where $z \in \mathbb{R}^N, w \in \mathbb{R}_+^N, \epsilon \in \mathbb{R}_+$. With this new up-date rule for the weight, which depends on $0 < \tau < 1$, we have formally, for $x^n \to \bar{x}$ and $\epsilon_n \to 0$,

$$\mathcal{J}_\tau(x^n, w^n, \epsilon_n) \to \|\bar{x}\|_\tau^\tau.$$

Hence such an iterative optimization tends to promote the ℓ_τ-quasi-norm minimization.

Surprisingly the rate of local convergence of this modified algorithm is superlinear; the rate is larger for smaller τ, and approaches a quadratic rate as $\tau \to 0$. More precisely, the local error $E_n := \|x^n - x^*\|_\tau^\tau$ satisfies

$$E_{n+1} \leq \mu(\gamma, \tau) E_n^{2-\tau}, \tag{4.34}$$

where $\mu(\gamma, \tau) < 1$ for $\gamma > 0$ sufficiently small. The validity of (4.34) is restricted to x^n in a (small) ball centered at x^*. In particular, if x^0 is close enough to x^* then (4.34) ensures the convergence of the algorithm to the k-sparse solution x^*, see Figure 4.

4.3 Numerical Experiments

Figure 5 shows a typical *phase transition* diagram related to the (experimentally determined) probability of successful recovery of sparse vectors by means of the iteratively re-weighted least squares algorithm. For each point of this diagram with coordinates $(m/N, k/m) \in [0, 1]^2$, we indicate the empirical success probability of recovery of a k-sparse vector $x \in \mathbb{R}^N$ from m measurements $y = Ax$. The brightness level corresponds to the probability. As measurement matrix a real random Fourier type matrix A was used, with entries given by

$$A_{k,j} = \cos(2\pi j \xi_k), \quad j = 1, \ldots, N,$$

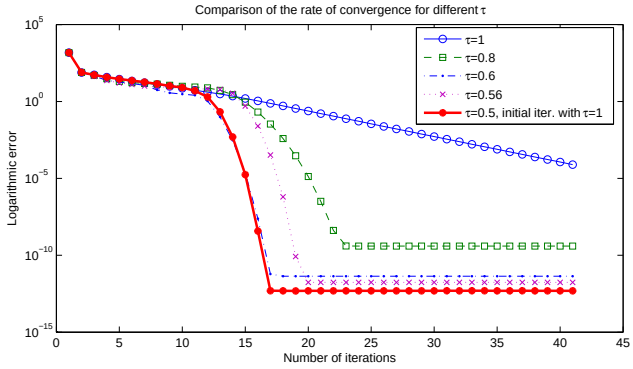

Figure 4: The decay of logarithmic error is shown, as a function of the number of iterations of IRLS for different values of τ (1, 0.8, 0.6, 0.56). We show also the results of an experiment in which the initial 10 iterations are performed with $\tau = 1$ and the remaining iterations with $\tau = 0.5$.

and the ξ_k, $k = 1, ..., m$, are sampled independently and uniformly at random from $[0, 1]$. (Theorem 3.7 does not apply directly to real random Fourier matrices, but an analogous result concerning the RIP for such matrices can be found in [75].)

Figure 6 shows a section of a phase transition diagram related to the (experimentally determined) probability of successful recovery of sparse vectors from linear measurements $y = Ax$, where the matrix A has i.i.d. Gaussian entries. Here both m and N are fixed and only k is variable. This diagram establishes the transition from a situation of exact reconstruction for sparse vectors with high probability to very unlikely recovery for vectors with many nonzero entries. These numerical experiments used the iteratively reweighted least squares algorithm with different parameters $0 < \tau \leq 1$. It is of interest to emphasize the enhanced success rate when using the algorithm for $\tau < 1$. Similarly, many other algorithms are tested by showing the corresponding phase transition diagrams and comparing them, see [6] for a detailed account of phase transitions for greedy algorithms and [28, 32] for ℓ_1-minimization.

This section is concluded by showing applications of ℓ_1-minimization methods to a real-life image recolorization problem [41, 42] in Figure 7. The image is known completely only on very few colored portions, while on the remaining areas only gray levels are provided. With this partial information, the use of ℓ_1-minimization with respect to wavelet or curvelets coefficients allows for high fidelity recolorization of the whole images.

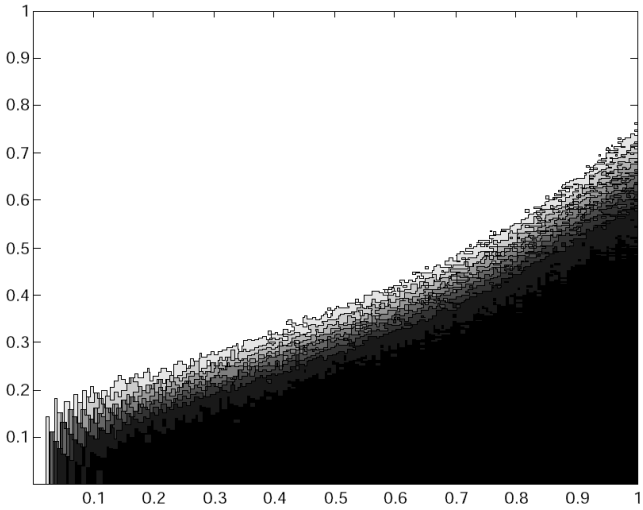

Figure 5: Empirical success probability of recovery of k-sparse vectors $x \in \mathbb{R}^N$ from measurements $y = Ax$, where $A \in \mathbb{R}^{m \times N}$ is a real random Fourier matrix. The dimension $N = 300$ of the vectors is fixed. Each point of this diagram with coordinates $(m/N, k/m) \in [0,1]^2$ indicates the empirical success probability of recovery, which is computed by running 100 experiments with randomly generated k-sparse vectors x and randomly generated matrix. The algorithm used for the recovery is the iteratively re-weighted least squares method tuned to promote ℓ_1-minimization.

5 Open Questions

The field of compressed sensing is rather young so there remain many directions to be explored and it is questionable whether one can assign certain problems in the field already at this point the status of an "open problem". Anyhow, below we list two problems that remained unsolved until the time of writing of this article.

5.1 Deterministic compressed sensing matrices

So far only several types of random matrices $A \in \mathbb{C}^{m \times N}$ are known to satisfy the RIP $\delta_s \leq \delta \leq 0.4$ (say) for

$$m = C_\delta s \log^\alpha(N) \tag{5.1}$$

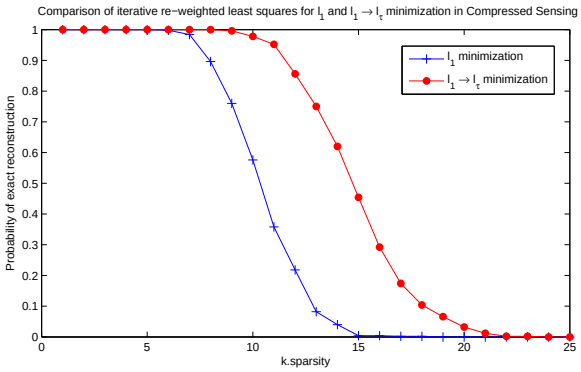

Figure 6: Empirical success probability of recovery of a k-sparse vector $x \in \mathbb{R}^{250}$ from measurements $y = Ax$, where $A \in \mathbb{R}^{50\times 250}$ is Gaussian. The matrix is generated once; then, for each sparsity value k shown in the plot, 500 attempts were made, for randomly generated k-sparse vectors x. Two different IRLS algorithms were compared: one with weights inspired by ℓ_1-minimization, and the IRLS with weights that gradually moved during the iterations from an ℓ_1- to an ℓ_τ-minimization goal, with final $\tau = 0.5$.

for some constant C_δ and some exponent α (with high probability). This is a strong form of existence statement. It is open, however, to provide deterministic and explicit $m \times N$ matrices that satisfy the RIP $\delta_s \leq \delta \leq 0.4$ (say) in the desired range (5.1).

In order to show RIP estimates in the regime (5.1) one has to take into account cancellations of positive and negative (or more generally complex) entries in the matrix, see also Section 3.6. This is done "automatically" with probabilistic methods but seems to be much more difficult to exploit when the given matrix is deterministic. It may be conjectured that certain equiangular tight frames or the "Alltop matrix" in [82, 70] do satisfy the RIP under (5.1). This is supported by numerical experiments in [70]. It is expected, however, that a proof is very hard and requires a good amount of analytic number theory.

The best deterministic construction of CS matrices known so far uses deterministic expander graphs [5]. Instead of the usual RIP, one shows that the adjacency matrix of such an expander graph has the 1-RIP, where the ℓ_2-norm is replaced by the ℓ_1-norm at each occurence in (3.8). The 1-RIP also implies recovery by ℓ_1-minimization. The best known deterministic expanders [17] yield sparse recovery under the condition $m \geq Cs(\log N)^{c\log^2(N)}$. Although the scaling in s is linear as desired, the term

Figure 7: Iterations of the recolorization methods proposed in [41, 42] via ℓ_1 and total variation minimization, for the virtual restoration of the frescoes of A. Mantegna (1452), which were destroyed by a bombing during World War II. Only a few colored fragments of the images were saved from the disaster, together with good quality gray level pictures dated to 1920.

$(\log N)^{c \log^2(N)}$ grows faster than any polynomial in $\log N$. Another drawback is that the deterministic expander graph is the output of a polynomial time algorithm, and it is questionable whether the resulting matrix can be regarded as *explicit*.

5.2 Removing log-factors in the Fourier-RIP estimate

It is known [16, 73, 78, 75] that a random partial Fourier matrix $A \in \mathbb{C}^{m \times N}$ satisfies the RIP with high probability provided

$$\frac{m}{\log(m)} \geq C_\delta s \log^2(s) \log(N).$$

(The condition stated in (3.19) implies this one.) It is conjectured that one can remove some of the log-factors. It must be hard, however, to improve this to a better estimate than $m \geq C_{\delta,\epsilon} s \log(N) \log(\log N)$. Indeed, this would imply an open conjecture of Talagrand [85] concerning the equivalence of the ℓ_1 and ℓ_2 norm of a linear combination of a subset of characters (complex exponentials).

6 Conclusions

Compressive sensing established itself by now as a new sampling theory which exhibits fundamental and intriguing connections with several mathematical fields, such as probability, geometry of Banach spaces, harmonic

analysis, theory of computability and information-based complexity. The link to convex optimization and the development of very efficient and robust numerical methods make compressive sensing a concept useful for a broad spectrum of natural science and engineering applications, in particular, in signal and image processing and acquisition. It can be expected that compressive sensing will enter various branches of science and technology to notable effect.

Recent developments, for instance the work [14, 76] on low rank matrix recovery via nuclear norm minimization, suggest new possible extensions of compressive sensing to more complex structures. Moreover, new challenges are now emerging in numerical analysis and simulation where high-dimensional problems (e.g., stochastic partial differential equations in finance and electron structure calculations in chemistry and biochemistry) became the frontier. In this context, besides other forms of efficient approximation, such as sparse grid and tensor product methods [10], compressive sensing is a promising concept which is likely to cope with the "curse of dimensionality". In particular, further systematic developments of adaptivity in the presence of different scales, randomized algorithms, an increasing role for combinatorial aspects of the underlying algorithms, are examples of possible future developments, which are inspired by the successful history of compressive sensing [84].

7 Cross-References

Compressive sensing has connections with the following chapters of the book: Wavelets, Fourier Analysis, Compression, Astronomy, CT, Variational Methods for Image Analysis, Numerical Methods for Variational Approach in Image Analysis, Duality and Convex Minimization, Mumford Shah, Phase Field Models, Iterative Solution Methods, Learning, Classification, Data Mining, Radar, Sampling Methods, Linear Inverse Problems, Nonlinear Inverse Problems, Regularization Methods for Ill-Posed Problems, Seismic.

8 Recommended Reading

The initial papers on the subject are [13, 16, 26]. An introduction to compressive sensing is contained in the monograph [45] by Rauhut and Foucart under current preparation. Another introductory source are the lecture notes [39, 75] of the summer school "Theoretical Foundations and Numerical Methods for Sparse Recovery ", held at RICAM in September

2009. The overview papers [12, 3, 11, 77] introduce to various theoretical and applied aspects of compressive sensing. A large collection of the vastly growing research literature on the subject is available on the webpage `http://www.compressedsensing.com`.

References

[1] D. Achlioptas. Database-friendly random projections. In *Proc. 20th Annual ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, pages 274–281, 2001.

[2] F. Affentranger and R. Schneider. Random projections of regular simplices. *Discrete Comput. Geom.*, 7(3):219–226, 1992.

[3] R. Baraniuk. Compressive sensing. *IEEE Signal Process. Magazine*, 24(4):118–121, 2007.

[4] 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.

[5] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. *preprint*, 2008.

[6] J. D. Blanchard, C. Cartis, J. Tanner, and A. Thompson. Phase transitions for greedy sparse approximation algorithms. *preprint*, 2009.

[7] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.

[8] J. Bobin, J.-L. Starck, and R. Ottensamer. Compressed sensing in astronomy. *IEEE J. Sel. Topics Signal Process.*, 2(5):718–726, 2008.

[9] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2004.

[10] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[11] E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Magazine*, 25(2):21–30, 2008.

[12] E. J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, 2006.

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

[14] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.,*, 9:717–772, 2009.

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

[16] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[17] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson. Randomness conductors and constant-degree lossless expanders. In *Proceedings of the Thirty-Fourth Annual ACM*, pages 659–668 (electronic). ACM, 2002.

[18] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[19] O. Christensen. *An Introduction to Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2003.

[20] A. K. Cline. Rate of convergence of Lawson's algorithm. *Math. Comp.*, 26:167–176, 1972.

[21] A. Cohen, W. Dahmen, and R. A. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[22] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. In *CISS*, 2006.

[23] I. Daubechies, R. DeVore, M. Fornasier, and C. Güntürk. Iteratively re-weighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63(1):1–38, 2010.

[24] B. Do, P. Indyk, E. Price, and D. Woodruff. Lower bounds for sparse recovery. In *Proc. SODA*, 2010.

[25] D. Donoho and B. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(2):577–591, 1992.

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

[27] 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.

[28] D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.*, 35(4):617–652, 2006.

[29] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ell^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202, 2003.

[30] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[31] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452–9457, 2005.

[32] D. L. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 22(1):1–53, 2009.

[33] D. L. Donoho and Y. Tsaig. Fast solution of l1-norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54(11):4789–4812, 2008.

[34] R. Dorfman. The detection of defective members of large populations. *Ann. Statist.*, 14:436–440, 1943.

[35] M. Duarte, M. Davenport, D. Takhar, J. Laska, S. Ting, K. Kelly, and R. Baraniuk. Single-Pixel Imaging via Compressive Sampling. *Signal Processing Magazine, IEEE,*, 25(2):83–91, March , 2008.

[36] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

[37] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48(9):2558–2567, 2002.

[38] A. Fannjiang, P. Yan, and T. Strohmer. Compressed Remote Sensing of Sparse Objects. *preprint*, 2009.

[39] M. Fornasier. Numerical methods for sparse recovery. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Series Comp. Appl. Math. deGruyter, in preparation.

[40] M. Fornasier, A. Langer, and C. B. Schönlieb. A convergent overlapping domain decomposition method for total variation minimization. *preprint*, 2009.

[41] M. Fornasier and R. March. Restoration of color images by vector valued BV functions and variational calculus. *SIAM J. Appl. Math.*, 68(2):437–460, 2007.

[42] M. Fornasier, R. Ramlau, and G. Teschke. The application of joint sparsity and total variation minimization algorithms to a real-life art restoration problem. *Adv. Comput. Math.*, 31(1-3):157–184, 2009.

[43] S. Foucart. A note on guaranteed sparse recovery via ℓ_1-minimization. *Appl. Comput. Harmon. Anal.*, to appear.

[44] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of ℓ_p-balls for $0 < p \leq 1$. *preprint*, 2010.

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

[46] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6):1341–1344, 2004.

[47] A. Garnaev and E. Gluskin. On widths of the Euclidean ball. *Sov. Math., Dokl.*, 30:200–204, 1984.

[48] A. C. Gilbert, S. Muthukrishnan, S. Guha, P. Indyk, and M. Strauss. Near-Optimal Sparse Fourier Representations via Sampling. In *Proc. STOC'02*, pages 152 – 161. others, Association for Computing Machinery, 2002.

[49] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, USA, January 12-14, 2003*, pages 243–252. SIAM and Association for Computing Machinery, 2003.

[50] A. C. Gilbert, M. Strauss, J. A. Tropp, and R. Vershynin. One sketch for all: Fast algorithms for compressed sensing. *preprint*, 2006.

[51] E. Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. USSR-Sb.*, 48:173–182, 1984.

[52] I. Gorodnitsky and B. Rao. Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. *IEEE Trans. Signal Process.*, 45(3):600–616, 1997.

[53] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, 2003.

[54] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[55] W. B. Johnson and J. Lindenstrauss, editors. *Handbook of the Geometry of Banach Spaces Vol I.* North-Holland Publishing Co., 2001.

[56] B. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Math. USSR, Izv.*, 11:317–333, 1977.

[57] C. Lawson. *Contributions to the Theory of Linear Least Maximum Approximation*. PhD thesis, University of California, Los Angeles, 1961.

[58] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[59] B. Logan. *Properties of High-Pass Signals*. PhD thesis, Columbia University, 1965.

[60] G. G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*. Springer, Berlin, 1996.

[61] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[62] S. Marple. *Digital Spectral Analysis with Applications*. Prentice - Hall, 1987.

[63] S. Mendelson, A. Pajor, and N. Tomczak Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2009.

[64] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995.

[65] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.

[66] E. Novak. Optimal recovery and n-widths for convex classes of functions. *J. Approx. Theory*, 80(3):390–408, 1995.

[67] M. Osborne, B. Presnell, and B. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.

[68] M. Osborne, B. Presnell, and B. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000.

[69] G. E. Pfander and H. Rauhut. Sparsity in time-frequency representations. *J. Fourier Anal. Appl.*, 16(2):233–260, 2010.

[70] G. E. Pfander, H. Rauhut, and J. Tanner. Identification of matrices having a sparse representation. *IEEE Trans. Signal Process.*, 56(11):5376–5388, 2008.

[71] R. Prony. Essai expérimental et analytique sur les lois de la Dilatabilité des uides élastique et sur celles de la Force expansive de la vapeur de leau et de la vapeur de lalkool, à différentes températures. *J. École Polytechnique*, 1:24–76, 1795.

[72] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.*, 22(1):16–42, 2007.

[73] H. Rauhut. Stability results for random sampling of sparse trigonometric polynomials. *IEEE Trans. Information Theory*, 54(12):5661–5670, 2008.

[74] H. Rauhut. Circulant and Toeplitz matrices in compressed sensing. In *Proc. SPARS'09*, 2009.

[75] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Series Comp. Appl. Math. deGruyter, to appear.

[76] B. Recht, M. Fazel, and P. Parillo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Rev.*, to appear.

[77] J. Romberg. Imaging via Compressive Sampling. *IEEE Signal Process. Magazine*, 25(2):14–20, March, 2008.

[78] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[79] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992.

[80] F. Santosa and W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330, 1986.

[81] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Trans. Signal Process.*, 56(5):1994–2002, 2008.

[82] T. Strohmer and R. W. j. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.

[83] T. Strohmer and M. Hermann. Compressed Sensing Radar. *IEEE Proc. Int. Conf. Acoustic, Speech, and Signal Processing, 2008*, pages 1509–1512, 2008.

[84] E. Tadmor. Numerical methods for nonlinear partial differential equations. In *Encyclopedia of Complexity and Systems Science*. Springer, 2009.

[85] M. Talagrand. Selecting a proportion of characters. *Israel J. Math.*, 108:173–191, 1998.

[86] G. Tauböck, F. Hlawatsch, and H. Rauhut. Compressive Estimation of Doubly Selective Channels: Exploiting Channel Sparsity to Improve Spectral Efficiency in Multicarrier Transmissions. 2009.

[87] H. Taylor, S. Banks, and J. McCoy. Deconvolution with the ℓ_1-norm. *Geophys. J. Internat.*, 44(1):39–52, 1979.

[88] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[89] J. Traub, G. Wasilkowski, and H. Wo'zniakowski. *Information-based complexity*. Computer Science and Scientific Computing. Academic Press Inc., 1988.

[90] J. Tropp and D. Needell. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, page 30, 2008.

[91] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[92] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 51(3):1030–1051, 2006.

[93] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Trans. Inform. Theory*, 56(1):520 –544, 2010.

[94] M. Unser. Sampling—50 Years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

[95] J. Vybiral. Widths of embeddings in function spaces. *J. Complexity*, 24(4):545–570, 2008.

[96] G. Wagner, P. Schmieder, A. Stern, and J. Hoch. Application of nonlinear sampling schemes to cosy-type spectra. *J. Biomolecular NMR*, 3(5):569, 1993.