Chapter 1
A Survey of Compressed Sensing

Holger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybíral

Abstract Compressed sensing was introduced some ten years ago as an effective way of acquiring signals, which possess a sparse or nearly sparse representation in a suitable basis or dictionary. Due to its solid mathematical backgrounds, it quickly attracted the attention of mathematicians from several different areas, so that the most important aspects of the theory are nowadays very well understood. In recent years, its applications started to spread out through applied mathematics, signal processing, and electrical engineering. The aim of this chapter is to provide an introduction into the basic concepts of compressed sensing. In the first part of this chapter, we present the basic mathematical concepts of compressed sensing, including the Null Space Property, Restricted Isometry Property, their connection to basis pursuit and sparse recovery, and construction of matrices with small restricted isometry constants. This presentation is easily accessible, largely self-contained, and includes proofs of the most important theorems. The second part gives an overview of the most important extensions of these ideas, including recovery of vectors with sparse representation in frames and dictionaries, discussion of (in)coherence and its implications for compressed sensing, and presentation of other algorithms of sparse recovery.

Holger Boche
Technische Universität München, Theresienstr. 90/IV, München, Germany e-mail: boche@tum.de

Robert Calderbank
Duke University, 317 Gross Hall, Durham NC, U. S. A. e-mail: robert.calderbank@duke.edu

Gitta Kutyniok
Technische Universität Berlin, Straße des 17. Juni 136, Berlin, Germany e-mail: kutyniok@math.tu-berlin.de

Jan Vybíral
Technische Universität Berlin, Straße des 17. Juni 136, Berlin, Germany e-mail: vybiral@math.tu-berlin.de

1.1 Introduction

Compressed sensing is a novel method of signal processing, which was introduced in [25] and [14] and which profited from its very beginning from fruitful interplay between mathematicians, applied mathematicians, and electrical engineers. The mathematical concepts are inspired by ideas from a number of different disciplines, including numerical analysis, stochastic, combinatorics, and functional analysis. On the other hand, the applications of compressed sensing range from image processing [29], medical imaging [51], and radar technology [5] to sampling theory [55, 68], and statistical learning.

The aim of this chapter is twofold. In Section 1.3 we collect the basic mathematical ideas from numerical analysis, stochastic, and functional analysis used in the area of compressed sensing to give an overview of basic notions, including the Null Space Property and the Restricted Isometry Property, and the relations between them. Most of the material in this section is presented with a self-contained proof, using only few simple notions from approximation theory and stochastic recalled in Section 1.2. We hope that this presentation will make the mathematical concepts of compressed sensing appealing and understandable both to applied mathematicians and electrical engineers. Although it can also be used as a basis for a lecture on compressed sensing for a wide variety of students, depending on circumstances, it would have to be complemented by other subjects of the lecturers choice to make a full one-semester course. Let us stress that the material presented in this section is by no means new or original, actually it is nowadays considered classical, or "common wisdom" throughout the community.

The second aim of this Chapter is to give (without proof) an overview of the most important extensions (Section 1.4).In this part, we refer to original research papers or to more extensive summaries of compressed sensing [23, 35, 40] for more details and further references.

1.2 Preliminaries

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

1.2.1 Norms and quasi-norms

In the most simple setting of discrete signals on finite domain, signals are modeled as (column) vectors in then n-dimensional Euclidean space, denoted by $\mathbb{R}^n$. We shall use different ways how to measure the size of such a vector. The most typical

way, however, is to consider its ℓ_p^n-norm, which is defined for $x = (x_1, \dots, x_n)^T$ and $p \in (0, \infty]$ as

$$\|x\|_p = \begin{cases} \Big(\sum_{j=1}^{n} |x_j|^p \Big)^{1/p}, & p \in (0, \infty); \\ \max\limits_{j=1,\dots,n} |x_j|, & p = \infty. \end{cases} \tag{1.1}$$

If $p < 1$, this expression does not satisfy the triangle inequality. Instead of that the following inequalities hold

$$\|x+z\|_p \le 2^{1/p-1} \big(\|x\|_p + \|z\|_p \big),$$
$$\|x+z\|_p^p \le \|x\|_p^p + \|z\|_p^p$$

for all $x \in \mathbb{R}^n$ and all $z \in \mathbb{R}^n$. If $p = 2$, ℓ_2^n is a (real) Hilbert space with the scalar product

$$\langle x, z \rangle = z^T x = \sum_{i=j}^{n} x_j z_j.$$

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

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

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

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

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

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

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

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

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

$$\sigma_k(x)_p = \inf_{z \in \Sigma_k} \|x - z\|_p = \begin{cases} \Big(\sum\limits_{j=k+1}^{n} (x_j^*)^p \Big)^{1/p}, & p \in (0, \infty); \\ x_{k+1}^*, & p = \infty. \end{cases} \tag{1.2}$$

The notions introduced so far, can be easily transferred to n-dimensional complex spaces. Especially, the scalar product of $x, y \in \mathbb{C}^n$ is defined by

$$\langle x,y\rangle = \sum_{j=1}^{n} x_j \overline{y_j},$$

where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$.

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

1.2.2 Random Variables

As several important constructions from the field of compressed sensing rely on randomness, we recall the basic notions from probability theory.

We denote by $(\Omega,\Sigma,\mathbb{P})$ a probability space. Here stands Ω for the sample space, Σ for a σ-algebra of subsets of Ω and $\mathbb{P}$ is a probability measure on (Ω,Σ). The sets $B \in \Sigma$ are called events, and their probability is denoted by

$$\mathbb{P}(B) = \int_B d\mathbb{P}(\omega).$$

A random variable X is a measurable function $X : \Omega \to \mathbb{R}$ and we denote by

$$\mu = \mathbb{E}X = \int_\Omega X(\omega) d\mathbb{P}(\omega)$$

its expected value, or mean, and by $\sigma^2 = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - (\mathbb{E}X)^2$ its variance. We recall Markov's inequality, which states

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \quad \text{for all } t > 0. \tag{1.3}$$

A random variable X is called *normal* (or *Gaussian*), if it has a density function

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(t-\mu)^2}{2\sigma^2}\Big), \quad t \in \mathbb{R}$$

for some real μ and positive σ^2, i.e. if $\mathbb{P}(a < X \leq b) = \int_a^b f(t)dt$ for all real $a < b$. In that case, the expected value of X is equal to μ and its variance to σ^2 and we often write $X \sim \mathcal{N}(\mu,\sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$, the normal variable is called *standard* and its density function is

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{t^2}{2}\Big), \quad t \in \mathbb{R}.$$

A random variable X is called *Rademacher* if

$$\mathbb{P}(X=1)=\mathbb{P}(X=-1)=1/2. \tag{1.4}$$

Random variables $X_1,\dots,X_N$ are called *independent*, if for every real $t_1,\dots,t_N$ the following formula holds

$$\mathbb{P}(X_1\le t_1,\dots,X_N\le t_N)=\prod_{j=1}^{N}\mathbb{P}(X_j\le t_j).$$

In that case,

$$\mathbb{E}\Big[\prod_{j=1}^{N}X_j\Big]=\prod_{j=1}^{N}\mathbb{E}(X_j). \tag{1.5}$$

If the random variables $X_1,\dots,X_N$ are independent and have the same distribution, we call them *independent identically distributed*, which is usually abbreviated as *i.i.d.*

1.3 Basic ideas of compressed sensing

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

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

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

1.3.1 Basis pursuit

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

We shall start our tour through compressed sensing by showing that even every other algorithm solving (P_0) suffers from this drawback. This will be formulated in the language of complexity theory as the statement, that the (P_0) problem is NP-hard. Before we come to that, we introduce the basic terms used in the sequel. We refer for example to [2] for an introduction to computational complexity.

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

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

Let us reformulate (P_0) as a decision problem. Namely, if the natural numbers k,m,n, $m \times n$ matrix A and $y \in \mathbb{R}^m$ are given, decide if there is a k-sparse solution x of the equation $Ax = y$. It is easy to see that this version of (P_0) is in the NP-class. Indeed, if the answer to the problem is "yes" and a certificate $x \in \mathbb{R}^n$ is given, then it can be verified in polynomial time if x is k-sparse and $Ax = y$.

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

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

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

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

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

Theorem 1.1. *The ℓ_0-minimization problem is NP-hard.*

Proof. It is sufficient to show that any algorithm solving the ℓ_0-minimization problem can be transferred in polynomial time into an algorithm solving the exact cover

problem. Let therefore $\{T_j : j = 1,\dots,n\}$ be a system of subsets of $\{1,\dots,m\}$ with $|T_j| = 3$ for all $j = 1,\dots,n$. Then we construct a matrix $A \in \mathbb{R}^{m\times n}$ by putting

$$a_{ij} := \begin{cases} 1 & \text{if } i \in T_j, \\ 0 & \text{if } i \notin T_j, \end{cases}$$

i.e. the jth column of A is the indicator function of T_j (denoted by $\chi_{T_j} \in \{0,1\}^m$) and

$$Ax = \sum_{j=1}^{n} x_j \chi_{T_j}. \tag{1.6}$$

The construction of A can of course be done in polynomial time.

Let now x be the solution to the ℓ_0-minimization problem with the matrix A and the vector $y = (1,\dots,1)^T$. It follows by (1.6), that $m = \|y\|_0 = \|Ax\|_0 \le 3\|x\|_0$, i.e. that $\|x\|_0 \ge m/3$. We will show that the exact cover problem has a positive solution if, and only if, $\|x\|_0 = m/3$.

Indeed, if the exact cover problem has a positive solution, then there is a set $J \subset \{1,\dots,n\}$ with $|J| = m/3$ and

$$\chi_{\{1,\dots,m\}} = \sum_{j\in J} \chi_{T_j}.$$

Hence $y = Ax$ for $x = \chi_J$ and $\|x\|_0 = |J| = m/3$. If, on the other hand, $y = Ax$ and $\|x\|_0 = m/3$, then $\{T_j : j \in \operatorname{supp}(x)\}$ solves the exact cover problem. □

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

We shall discuss later on, under which conditions the solution to (P_0) coincides with the solution of the following convex optimization problem called *basis pursuit*

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

which was introduced in [19]. But before we come to that, let us show, that in the real case this problem may be reformulated as a linear optimization problem, i.e. as the search for the minimizer of a linear function over a set given by linear constraints, whose number depends polynomially on the dimension. We refer to [42] for an introduction to linear programming.

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

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

is the unique solution of

$$\min_{u,v\in\mathbb{R}^n} \sum_{j=1}^{n} (u_j + v_j) \text{ s.t. } Au - Av = y \text{ and } u_j \geq 0 \text{ and } v_j \geq 0 \text{ for all } j = 1,\dots,n. \quad (1.7)$$

If namely (u',v') is another pair of vectors admissible in (1.7), then $x' = u' - v'$ satisfies $Ax' = y$ and x' is therefore admissible in (P_1). As x is the solution of (P_1), we get

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

If, on the other hand, the pair (u,v) is the unique solution of (1.7), then $x = u - v$ is the unique solution of (P_1). If namely z is another admissible vector in (P_1), then $u' = z^+$ and $v' = z^-$ are admissible in (1.7) and we obtain

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

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

1.3.2 Null Space Property

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

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

We start the discussion of the properties of basis pursuit by introducing the notion of Null Space Property, which first appeared in [20].

Definition 1.1. Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then A is said to have the *Null Space Property* (NSP) of order k if

$$\|v_T\|_1 < \|v_{T^c}\|_1 \quad \text{for all } v \in \ker A \setminus \{0\} \text{ and all } T \subset \{1,\dots,n\} \text{ with } |T| \leq k. \quad (1.8)$$

Remark 1.1. (i) The condition (1.8) states that vectors from the kernel of A are well spread, i.e. not supported on a set of small size. Indeed, if $v \in \mathbb{R}^n \setminus \{0\}$ is k-sparse and $T = \text{supp}\,(v)$, then (1.8) shows immediately, that v can not lie in the kernel of A.
(ii) If we add $\|v_{T^c}\|_1$ to both sides of (1.8), we obtain $\|v\|_1 < 2\|v_{T^c}\|_1$. If then T are the indices of the k largest coordinates of v taken in the absolute value, this inequality becomes $\|v\|_1 < 2\sigma_k(v)_1$.

Theorem 1.2. *Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$ if, and only if, A has the NSP of order k.*

Proof. Let us assume that every k-sparse vector x is the unique solution of (P_1) with $y = Ax$. Let $v \in \ker A \setminus \{0\}$ and let $T \subset \{1,\dots,n\}$ with $|T| \le k$ be arbitrary. Then v_T is k-sparse, and is therefore the unique solution of

$$\min_z \|z\|_1, \quad \text{s.t.} \quad Az = Av_T. \tag{1.9}$$

As $A(-v_{T^c}) = A(v - v_{T^c}) = A(v_T)$, this gives especially $\|v_T\|_1 < \|v_{T^c}\|_1$ and A has the NSP of order k.

Let us on the other hand assume that A has the NSP of order k. Let $x \in \mathbb{R}^n$ be a k-sparse vector and let $T = \operatorname{supp}(x)$. We have to show that $\|x\|_1 < \|z\|_1$ for every $z \in \mathbb{R}^n$ different from x with $Az = Ax$. But this follows easily by using (1.8) for the vector $(x-z) \in \ker A \setminus \{0\}$

$$\begin{aligned}\|x\|_1 &\le \|x - z_T\|_1 + \|z_T\|_1 = \|(x-z)_T\|_1 + \|z_T\|_1 < \|(x-z)_{T^c}\|_1 + \|z_T\|_1 \\ &= \|z_{T^c}\|_1 + \|z_T\|_1 = \|z\|_1.\end{aligned}$$

□

Remark 1.2. Theorem 1.2 states that the solutions of (P_0) may be found by (P_1), if A has the NSP of order k and if $y \in \mathbb{R}^m$ is such that, there exists a k-sparse solution x of the equation $Ax = y$. Indeed, if in such a case, $\hat{x}$ is a solution of (P_0), then $\|\hat{x}\|_0 \le \|x\|_0 \le k$. Finally, it follows by Theorem 1.2, that $\hat{x}$ is also a solution of (P_1) and that $x = \hat{x}$.

In the language of complexity theory, if we restrict the inputs of the ℓ_0-minimization problem to matrices with the NSP of order k and to vectors y, for which there is a k-sparse solution of the equation $Ax = y$, the problem belongs to the P-class and the solving algorithm with polynomial running time is any standard algorithm solving (P_1), or the corresponding linear problem (1.7).

1.3.3 Restricted Isometry Property

Although the Null Space Property is equivalent to the recovery of sparse solutions of underdetermined linear systems by basis pursuit in the sense just described, it is somehow difficult to construct matrices satisfying this property. We shall therefore present a sufficient condition called Restricted Isometry Property, which was first introduced in [14], and which ensures that the Null Space Property is satisfied.

Definition 1.2. Let $A \in \mathbb{R}^{m\times n}$ and let $k \in \{1,\dots,n\}$. Then the *restricted isometry constant* $\delta_k = \delta_k(A)$ of A of order k is the smallest $\delta \ge 0$, such that

$$(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2 \quad \text{for all} \quad x \in \Sigma_k. \tag{1.10}$$

Furthermore, we say that A satisfies the *Restricted Isometry Property* (RIP) of order k with the constant δ_k if $\delta_k > 0$.

Remark 1.3. The condition (1.10) states that A acts nearly isometrically when restricted to vectors from Σ_k. Of course, the smaller the constant $\delta_k(A)$ is, the closer is the matrix A to isometry on Σ_k. We will be therefore later interested in constructing matrices with small RIP constants. Finally, the inequality $\delta_1(A) \leq \delta_2(A) \leq \cdots \leq \delta_k(A)$ follows trivially.

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

Theorem 1.3. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then A has the NSP of order k.*

Proof. Let $v \in \ker A$ and let $T \subset \{1,\dots,n\}$ with $|T| \leq k$. We shall show, that

$$\|v_T\|_2 \leq \frac{\delta_{2k}}{1-\delta_k} \cdot \frac{\|v\|_1}{\sqrt{k}}. \tag{1.11}$$

If $\delta_k \leq \delta_{2k} < 1/3$, then Hölder's inequality gives immediately $\|v_T\|_1 \leq \sqrt{k}\|v_T\|_2 < \|v\|_1/2$ and the NSP of A of order k follows.

Before we come to the proof of (1.11), let us make the following observation. If $x,z \in \Sigma_k$ are two vectors with disjoint supports and $\|x\|_2 = \|z\|_2 = 1$, then $x \pm z \in \Sigma_{2k}$ and $\|x \pm z\|_2^2 = 2$. If we now combine the RIP of A

$$2(1-\delta_{2k}) \leq \|A(x \pm z)\|_2^2 \leq 2(1+\delta_{2k})$$

with the polarization identity, we get

$$|\langle Ax, Az\rangle| = \frac{1}{4}\left|\|Ax+Az\|_2^2 - \|Ax-Az\|_2^2\right| \leq \delta_{2k}.$$

Hence if A has the RIP of order $2k$ and $x,z \in \Sigma_k$ have disjoint supports, then

$$|\langle Ax, Az\rangle| \leq \delta_{2k}\|x\|_2\|z\|_2. \tag{1.12}$$

To show (1.11), let us assume that $v \in \ker A$ is fixed. It is enough to consider $T = T_0$ the set of the k largest entries of v taken in the absolute value. Furthermore, we denote by T_1 the set of k largest entries of $v_{T_0^c}$ in the absolute value, by T_2 the set of k largest entries of $v_{(T_0\cup T_1)^c}$ in the absolute value, etc. Using $0 = Av = A(v_{T_0} + v_{T_1} + v_{T_2} + \dots)$ and (1.12), we arrive at

$$\begin{aligned}\|v_{T_0}\|_2^2 &\leq \frac{1}{1-\delta_k}\|Av_{T_0}\|_2^2 = \frac{1}{1-\delta_k}\langle Av_{T_0}, A(-v_{T_1}) + A(-v_{T_2}) + \dots\rangle \\ &\leq \frac{1}{1-\delta_k}\sum_{j\geq 1}|\langle Av_{T_0}, Av_{T_j}\rangle| \leq \frac{\delta_{2k}}{1-\delta_k}\sum_{j\geq 1}\|v_{T_0}\|_2 \cdot \|v_{T_j}\|_2.\end{aligned}$$

We divide this inequality by $\|v_{T_0}\|_2 \neq 0$ and obtain

$$\|v_{T_0}\|_2 \leq \frac{\delta_{2k}}{1-\delta_k} \sum_{j\geq 1} \|v_{T_j}\|_2.$$

The proof is then completed by the following simple chain of inequalities, which involve only the definition of the sets $T_j, j \geq 0$.

$$\begin{aligned} \sum_{j\geq 1} \|v_{T_j}\|_2 &= \sum_{j\geq 1} \Big(\sum_{l\in T_j} |v_l|^2 \Big)^{1/2} \leq \sum_{j\geq 1} \Big(k \max_{l\in T_j} |v_l|^2 \Big)^{1/2} \\ &= \sum_{j\geq 1} \sqrt{k} \max_{l\in T_j} |v_l| \leq \sum_{j\geq 1} \sqrt{k} \min_{l\in T_{j-1}} |v_l| \leq \sum_{j\geq 1} \sqrt{k} \cdot \frac{\sum_{l\in T_{j-1}} |v_l|}{k} \\ &= \sum_{j\geq 1} \frac{\|v_{T_{j-1}}\|_1}{\sqrt{k}} = \frac{\|v\|_1}{\sqrt{k}}. \end{aligned} \tag{1.13}$$

□

Combining Theorems 1.2 and 1.3, we obtain immediately the following corollary.

Corollary 1.1. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then every k-sparse vector x is the unique solution of (P_1) with $y = Ax$.*

1.3.4 RIP for random matrices

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

We denote until the end of this section

$$A = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \cdots & \omega_{mn} \end{pmatrix}, \tag{1.14}$$

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

1.3.4.1 Concentration inequalities

Before we come to the main result of this chapter, we need some properties of independent standard normal variables.

Lemma 1.1. *(i) Let ω be a standard normal variable. Then $\mathbb{E}(e^{\lambda\omega^2}) = 1/\sqrt{1-2\lambda}$ for $-\infty < \lambda < 1/2$.*

(ii) (2-stability of the normal distribution) Let $m \in \mathbb{N}$, let $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ and let $\omega_1, \dots, \omega_m$ be i.i.d. standard normal variables. Then $\lambda_1\omega_1 + \dots + \lambda_m\omega_m \sim (\sum_{i=1}^m \lambda_i^2)^{1/2} \cdot \mathcal{N}(0,1)$, i.e. it is equidistributed with a multiple of a standard normal variable.

Proof. The proof of (i) follows from the substitution $s := \sqrt{1-2\lambda}\cdot t$ in the following way.

$$\begin{aligned}\mathbb{E}(e^{\lambda\omega^2}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda t^2}\cdot e^{-t^2/2}dt &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{(\lambda-1/2)t^2}dt \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-s^2/2}\cdot\frac{ds}{\sqrt{1-2\lambda}} = \frac{1}{\sqrt{1-2\lambda}}.\end{aligned}$$

Although the property (ii) is very well known (and there are several different ways to prove it), we provide a simple geometric proof for the sake of completeness. It is enough to consider the case $m = 2$. The general case then follows by induction.

Let therefore $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2, \lambda \neq 0$, be fixed and let ω_1 and ω_2 be i.i.d. standard normal random variables. We put $S := \lambda_1\omega_1 + \lambda_2\omega_2$. Let $t \geq 0$ be an arbitrary non-negative real number. We calculate

$$\begin{aligned}\mathbb{P}(S \leq t) &= \frac{1}{2\pi}\int_{(u,v):\lambda_1 u+\lambda_2 v\leq t} e^{-(u^2+v^2)/2}dudv = \frac{1}{2\pi}\int_{u\leq c;v\in\mathbb{R}} e^{-(u^2+v^2)/2}dudv \\ &= \frac{1}{\sqrt{2\pi}}\int_{u\leq c} e^{-u^2/2}du.\end{aligned}$$

We have used the rotational invariance of the function $(u,v) \to e^{-(u^2+v^2)/2}$. The value of c is given by the distance of the origin from the line $\{(u,v) : \lambda_1 u + \lambda_2 v = t\}$. It follows by elementary geometry and Pythagorean theorem that

$$c = \left\| \left(\frac{\lambda_1 t}{\lambda_1^2+\lambda_2^2}, \frac{\lambda_2 t}{\lambda_1^2+\lambda_2^2} \right) \right\|_2 = \frac{t}{\sqrt{\lambda_1^2+\lambda_2^2}}.$$

We therefore get

$$\mathbb{P}(S \leq t) = \frac{1}{\sqrt{2\pi}}\int_{\sqrt{\lambda_1^2+\lambda_2^2}\cdot u\leq t} e^{-u^2/2}du = \mathbb{P}\left(\sqrt{\lambda_1^2+\lambda_2^2}\cdot\omega \leq t\right).$$

The same estimate holds for negative t's by symmetry and the proof is finished. □

If $\omega_1, \dots, \omega_m$ are (possibly dependent) standard normal random variables, then $\mathbb{E}(\omega_1^2 + \cdots + \omega_m^2) = m$. If $\omega_1, \dots, \omega_m$ are even independent, then the value of $\omega_1^2 + \cdots + \omega_m^2$ concentrates very strongly around m. This effect is known as *concentration of measure*, cf. [48, 49, 54].

Lemma 1.2. *Let $m \in \mathbb{N}$ and let $\omega_1, \dots, \omega_m$ be i.i.d. standard normal variables. Let $0 < \varepsilon < 1$. Then*

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

and

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

Proof. We prove only the first inequality. The second one follows in exactly the same manner. Let us put $\beta := 1 + \varepsilon > 1$ and calculate

$$\begin{aligned}
\mathbb{P}(\omega_1^2 + \cdots + \omega_m^2 \geq \beta m) &= \mathbb{P}(\omega_1^2 + \cdots + \omega_m^2 - \beta m \geq 0) \\
&= \mathbb{P}(\lambda(\omega_1^2 + \cdots + \omega_m^2 - \beta m) \geq 0) \\
&= \mathbb{P}(\exp(\lambda(\omega_1^2 + \cdots + \omega_m^2 - \beta m)) \geq 1) \\
&\leq \mathbb{E}\exp(\lambda(\omega_1^2 + \cdots + \omega_m^2 - \beta m)),
\end{aligned}$$

where $\lambda > 0$ is a positive real number, which shall be chosen later on. We have used the Markov's inequality (1.3) in the last step. Further we use the elementary properties of exponential function and (1.5) for the independent variables $\omega_1, \dots, \omega_m$. This leads to

$$\mathbb{E}\exp(\lambda(\omega_1^2 + \cdots + \omega_m^2 - \beta m)) = e^{-\lambda\beta m} \cdot \mathbb{E}e^{\lambda\omega_1^2} \cdots e^{\lambda\omega_m^2} = e^{-\lambda\beta m} \cdot (\mathbb{E}e^{\lambda\omega_1^2})^m$$

and with the help of Lemma 1.1 we get finally (for $0 < \lambda < 1/2$)

$$\mathbb{E}\exp(\lambda(\omega_1^2 + \cdots + \omega_m^2 - \beta m)) = e^{-\lambda\beta m} \cdot (1-2\lambda)^{-m/2}.$$

We now look for the value of $0 < \lambda < 1/2$, which would minimize the last expression. Therefore, we take the derivative of $e^{-\lambda\beta m} \cdot (1-2\lambda)^{-m/2}$ and put it equal to zero. After a straightforward calculation, we get

$$\lambda = \frac{1 - 1/\beta}{2},$$

which obviously satisfies also $0 < \lambda < 1/2$. Using this value of λ we obtain

$$\begin{aligned}
\mathbb{P}(\omega_1^2 + \cdots + \omega_m^2 \geq \beta m) &\leq e^{-\frac{1-1/\beta}{2} \cdot \beta m} \cdot (1 - (1 - 1/\beta))^{-m/2} = e^{-\frac{\beta-1}{2}m} \cdot \beta^{m/2} \\
&= e^{-\frac{\varepsilon m}{2}} \cdot e^{\frac{m}{2}\ln(1+\varepsilon)}.
\end{aligned}$$

The result then follows from the inequality

$$\ln(1+t) \leq t - \frac{t^2}{2} + \frac{t^3}{3}, \quad -1 < t < 1.$$

□

Using 2-stability of the normal distribution, Lemma 1.2 shows immediately that A defined as in (1.14) acts with high probability as isometry on one fixed $x \in \mathbb{R}^n$.

Theorem 1.4. *Let $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ and let A be as in* (1.14)*. Then*

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - 1\Big| \geq t\Big) \leq 2e^{-\frac{m}{2}[t^2/2 - t^3/3]} \leq 2e^{-Cmt^2} \tag{1.15}$$

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

Proof. Let $x = (x_1, x_2, \ldots, x_n)^T$. Then we get by the 2-stability of normal distribution and Lemma 1.2

$$\begin{aligned}
&\mathbb{P}\Big(\Big|\|Ax\|_2^2 - 1\Big| \geq t\Big) \\
&= \mathbb{P}\Big(\big|(\omega_{1,1}x_1 + \cdots + \omega_{1n}x_n)^2 + \cdots + (\omega_{m1}x_1 + \cdots + \omega_{mn}x_n)^2 - m\big| \geq mt\Big) \\
&= \mathbb{P}\Big(\big|\omega_1^2 + \cdots + \omega_m^2 - m\big| \geq mt\Big) \\
&= \mathbb{P}\Big(\omega_1^2 + \cdots + \omega_m^2 \geq m(1+t)\Big) + \mathbb{P}\Big(\omega_1^2 + \cdots + \omega_m^2 \leq m(1-t)\Big) \\
&\leq 2e^{-\frac{m}{2}[t^2/2 - t^3/3]}.
\end{aligned}$$

This gives the first inequality in (1.15). The second one follows by simple algebraic manipulations (for $C = 1/12$). □

Remark 1.4. (i) Observe, that (1.15) may be easily rescaled to

$$\mathbb{P}\Big(\Big|\|Ax\|_2^2 - \|x\|_2^2\Big| \geq t\|x\|_2^2\Big) \leq 2e^{-Cmt^2}, \tag{1.16}$$

which is true for every $x \in \mathbb{R}^n$.
(ii) A slightly different proof of (1.15) is based on the rotational invariance of the distribution underlying the random structure of matrices defined by (1.14). Therefore, it is enough to prove (1.15) only for one fixed element $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. Taking $x = e_1 = (1, 0, \ldots, 0)^T$ to be the first canonical unit vector allows us to use Lemma 1.2 without the necessity of applying the 2-stability of normal distribution.

1.3.4.2 RIP for random Gaussian matrices

The proof of restricted isometry property of random matrices generated as in (1.14) is based on two main ingredients. The first is the concentration of measure phenomenon described in its most simple form in Lemma 1.2, and reformulated in

Theorem 1.4. The second is the following entropy argument, which allows to extend Theorem 1.4 and (1.15) from one fixed $x \in \mathbb{R}^n$ to the set Σ_k of all k-sparse vectors.

Lemma 1.3. *Let $t > 0$. Then there is a set $\mathcal{N} \subset \mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ with*
(i) $|\mathcal{N}| \leq (1+2/t)^n$ and
(ii) for every $z \in \mathbb{S}^{n-1}$, there is a $x \in \mathcal{N}$ with $\|x-z\|_2 \leq t$.

Proof. Choose any $x^1 \in \mathbb{S}^{n-1}$. If $x^1,\dots,x^j \in \mathbb{S}^{n-1}$ were already chosen, take $x^{j+1} \in \mathbb{S}^{n-1}$ arbitrarily with $\|x^{j+1}-x^l\|_2 > t$ for all $l = 1,\dots,j$. This process is then repeated as long as possible, i.e. until we obtain a set $\mathcal{N} = \{x^1,\dots,x^N\} \subset \mathbb{S}^{n-1}$, such that for every $z \in \mathbb{S}^{n-1}$ there is a $j \in \{1,\dots,N\}$ with $\|x^j - z\|_2 \leq t$. This gives the property (ii).

We will use volume arguments to prove (i). It follows by construction, that $\|x^i - x^j\|_2 > t$ for every $i, j \in \{1,\dots,N\}$ with $i \neq j$. By triangle inequality, the balls $B(x^j, t/2)$ are all disjoint and are all included in the ball with the center in the origin and radius $1+t/2$. By comparing the volumes we get

$$N \cdot (t/2)^n \cdot V \leq (1+t/2)^n \cdot V,$$

where V is the volume of the unit ball in $\mathbb{R}^n$. Hence, we get $N = |\mathcal{N}| \leq (1+2/t)^n$. □

With all these tools at hand, we can now state the main theorem of this section, whose proof follows closely the arguments of [4].

Theorem 1.5. *Let $n \geq m \geq k \geq 1$ be natural numbers and let $0 < \varepsilon < 1$ and $0 < \delta < 1$ be real numbers with*

$$m \geq C\delta^{-2}\Big(k\ln(en/k) + \ln(2/\varepsilon)\Big), \tag{1.17}$$

where $C > 0$ is an absolute constant. Let A be again defined by (1.14)*. Then*

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

Proof. The proof follows by the concentration inequality of Theorem 1.4 and the entropy argument described in Lemma 1.3. By this lemma, there is a set

$$\mathcal{N} \subset Z := \{z \in \mathbb{R}^n : \operatorname{supp}(z) \subset \{1,\dots,k\}, \|z\|_2 = 1\},$$

such that
(i) $|\mathcal{N}| \leq 9^k$ and
(ii) $\min_{x\in\mathcal{N}} \|z-x\|_2 \leq 1/4$ for every $z \in Z$.

We show that if $\big|\|Ax\|_2^2 - 1\big| \leq \delta/2$ for all $x \in \mathcal{N}$, then $\big|\|Az\|_2^2 - 1\big| \leq \delta$ for all $z \in Z$.

We proceed by the following bootstrap argument. Let $\gamma > 0$ be the smallest number, such that $\big|\|Az\|_2^2 - 1\big| \leq \gamma$ for all $z \in Z$. Then $\big|\|Au\|_2^2 - \|u\|_2^2\big| \leq \gamma\|u\|_2^2$ for all $u \in \mathbb{R}^n$ with $\operatorname{supp}(u) \subset \{1,\dots,k\}$ and, by polarization identity,

$$|\langle Au, Av\rangle - \langle u, v\rangle| \leq \gamma \|u\|_2 \|v\|_2 \tag{1.18}$$

for all $u, v \in \mathbb{R}^n$ with supp $(u) \cup$ supp $(v) \subset \{1, \dots, k\}$.

Let now again $z \in Z$. Then there is an $x \in \mathcal{N}$, such that $\|z - x\|_2 \leq 1/4$. We obtain by triangle inequality and (1.18)

$$\begin{aligned} \big| \|Az\|_2^2 - 1 \big| &= \big| \|Ax\|_2^2 - 1 + \langle A(z+x), A(z-x)\rangle - \langle z+x, z-x\rangle \big| \\ &\leq \delta/2 + \gamma \|z+x\|_2 \|z-x\|_2 \leq \delta/2 + \gamma/2. \end{aligned}$$

As the supremum of the left-hand side over all admissible z's is equal to γ, we obtain that $\gamma \leq \delta$ and the statement follows.

Equipped with this tool, the rest of the proof follows by a simple union bound.

$$\begin{aligned} \mathbb{P}(\delta_k(A) > \delta) &\leq \sum_{\substack{T \subset \{1,\dots,n\} \\ |T| \leq k}} \mathbb{P}\Big(\exists z \in \mathbb{R}^n : \text{supp } (z) \subset T, \|z\|_2 = 1 \text{ and } \big| \|Az\|_2^2 - 1 \big| > \delta \Big) \\ &= \binom{n}{k} \mathbb{P}\Big(\exists z \in Z \text{ with } \big| \|Az\|_2^2 - 1 \big| > \delta \Big) \\ &\leq \binom{n}{k} \mathbb{P}\Big(\exists x \in \mathcal{N} : \big| \|Ax\|_2^2 - 1 \big| > \delta/2 \Big). \end{aligned}$$

By Theorem 1.4, the last probability may be estimated from above by $2e^{-C'm\delta^2}$. Hence we obtain

$$\mathbb{P}(\delta_k(A) > \delta) \leq 9^k \binom{n}{k} \cdot 2e^{-C'm\delta^2}$$

Hence it is enough to show that the last quantity is at most ε if (1.17) is satisfied. But this follows by straightforward algebraic manipulations and the well-known estimate

$$\binom{n}{k} \leq \frac{n^k}{k!} \leq \Big(\frac{en}{k} \Big)^k.$$

□

1.3.4.3 Lemma of Johnson and Lindenstrauss

Concentration inequalities similar to (1.15) play an important role in several areas of mathematics. We shall present their connection to the famous result from functional analysis called Johnson-Lindenstrauss lemma, cf. [45]. The lemma states that a set of points in a high-dimensional space can be embedded into a space of much lower dimension in such a way that the mutual distances between the points are nearly preserved. The connection between this classical result and compressed sensing was first highlighted in [4], cf. also [46].

Lemma 1.4. *Let* $0 < \varepsilon < 1$ *and let* m, N *and* n *be natural numbers with*

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

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

$$(1-\varepsilon)\|x^i - x^j\|_2^2 \leq \|f(x^i) - f(x^j)\|_2^2 \leq (1+\varepsilon)\|x^i - x^j\|_2^2, \qquad i,j \in \{1,\ldots,N\}. \tag{1.19}$$

Proof. We put $f(x) = Ax$, where again

$$Ax = \frac{1}{\sqrt{m}} \begin{pmatrix} \omega_{1,1} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{m1} & \cdots & \omega_{mn} \end{pmatrix} x,$$

and $\omega_{ij}, i = 1,\ldots,m, j = 1,\ldots,n$ are i.i.d. standard normal variables. We show that with this choice f satisfies (1.19) with positive probability. This proves the existence of such a mapping.

Let $i,j \in \{1,\ldots,N\}$ arbitrary with $x^i \neq x^j$. Then we put $z = \frac{x^i - x^j}{\|x^i - x^j\|_2}$ and evaluate the probability that the right hand side inequality in (1.19) does not hold. Theorem 1.4 then implies

$$\begin{aligned} \mathbb{P}\left(\left| \|f(x^i) - f(x^j)\|_2^2 - \|x^i - x^j\|_2^2 \right| > \varepsilon \|x^i - x^j\|_2^2 \right) &= \mathbb{P}\left(\left| \|Az\|^2 - 1 \right| > \varepsilon \right) \\ &\leq e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]}. \end{aligned}$$

The same estimate is also true for all $\binom{N}{2}$ pairs $\{i,j\} \subset \{1,\ldots,N\}$ with $i \neq j$. The probability, that one of the inequalities in (1.19) is not satisfied is therefore at most

$$2 \cdot \binom{N}{2} \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} < N^2 \cdot e^{-\frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3]} = \exp\left(2\ln N - \frac{m}{2}[\varepsilon^2/2 - \varepsilon^3/3] \right) \leq e^0 = 1$$

for $m \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln N$. Therefore, the probability that (1.19) holds for all $i,j \in \{1,\ldots,N\}$ is positive and the result follows. □

1.3.5 Stability and Robustness

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

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

k-term approximation of $x \in \mathbb{R}^n$ is, the better we should be able to approximate x.

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

We shall show that the methods of compressed sensing can be extended also to this kind of scenario. There is a number of different estimates in the literature, which show that the technique of compressed sensing is stable and robust. We will present only one of them (with more to come in Section 1.4.3). Its proof is a modification of the proof of Theorem 1.3, and follows closely [11].

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

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

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

Theorem 1.6. *Let $\delta_{2k} < \sqrt{2} - 1$ and $\|e\|_2 \leq \eta$. Then the solution $\hat{x}$ of $(P_{1,\eta})$ satisfies*

$$\|x - \hat{x}\|_2 \leq \frac{C\sigma_k(x)_1}{\sqrt{k}} + D\eta, \tag{1.20}$$

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

Proof. First, let us recall that if A has RIP of order $2k$ and $u, v \in \Sigma_k$ are two vectors with disjoint supports, then we have by (1.12)

$$|\langle Au, Av \rangle| \leq \delta_{2k} \|u\|_2 \|v\|_2. \tag{1.21}$$

Let us put $h = \hat{x} - x$ and let us define the index set $T_0 \subset \{1, \dots, n\}$ as the locations of k largest entries of x taken in the absolute value. Furthermore, we define $T_1 \subset T_0^c$ to be the indices of k largest absolute entries of $h_{T_0^c}$, T_2 the indices of k largest absolute entries of $h_{(T_0 \cup T_1)^c}$, etc. As $\hat{x}$ is an admissible point in $(P_{1,\eta})$, the triangle inequality gives

$$\|Ah\|_2 = \|A(x - \hat{x})\|_2 \leq \|Ax - y\|_2 + \|y - A\hat{x}\|_2 \leq 2\eta. \tag{1.22}$$

As $\hat{x}$ is the minimizer of $(P_{1,\eta})$, we get $\|\hat{x}\|_1 = \|x + h\|_1 \leq \|x\|_1$, which we use to show that h must be small outside of T_0. Indeed, we obtain

$$\begin{aligned}\|h_{T_0^c}\|_1 &= \|(x+h)_{T_0^c} - x_{T_0^c}\|_1 + \|(x+h)_{T_0} - h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &\le \|(x+h)_{T_0^c}\|_1 + \|x_{T_0^c}\|_1 + \|(x+h)_{T_0}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &= \|x+h\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &\le \|x\|_1 + \|x_{T_0^c}\|_1 + \|h_{T_0}\|_1 - \|x_{T_0}\|_1 \\ &= \|h_{T_0}\|_1 + 2\|x_{T_0^c}\| \le k^{1/2}\|h_{T_0}\|_2 + 2\sigma_k(x)_1.\end{aligned}$$

Using this together with the approach applied already in (1.13), we derive

$$\sum_{j\ge 2} \|h_{T_j}\|_2 \le k^{-1/2}\|h_{T_0^c}\|_1 \le \|h_{T_0}\|_2 + 2k^{-1/2}\sigma_k(x)_1. \tag{1.23}$$

We use the RIP property of A, (1.21), (1.22), (1.23) and the simple inequality $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \le \sqrt{2}\|h_{T_0\cup T_1}\|_2$ and get

$$\begin{aligned}(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2^2 &\le \|Ah_{T_0\cup T_1}\|_2^2 = \langle Ah_{T_0\cup T_1}, Ah\rangle - \langle Ah_{T_0\cup T_1}, \sum_{j\ge 2} Ah_{T_j}\rangle \\ &\le \|Ah_{T_0\cup T_1}\|_2\|Ah\|_2 + \sum_{j\ge 2} |\langle Ah_{T_0}, Ah_{T_j}\rangle| + \sum_{j\ge 2} |\langle Ah_{T_1}, Ah_{T_j}\rangle| \\ &\le 2\eta\sqrt{1+\delta_{2k}}\|h_{T_0\cup T_1}\|_2 + \delta_{2k}(\|h_{T_0}\|_2 + \|h_{T_1}\|_2)\sum_{j\ge 2}\|h_{T_j}\|_2 \\ &\le \|h_{T_0\cup T_1}\|_2\Big(2\eta\sqrt{1+\delta_{2k}} + \sqrt{2}\delta_{2k}\|h_{T_0}\|_2 + 2\sqrt{2}\delta_{2k}k^{-1/2}\sigma_k(x)_1\Big).\end{aligned}$$

We divide this inequality with $(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2$, replace $\|h_{T_0}\|_2$ with the larger quantity $\|h_{T_0\cup T_1}\|_2$ and subtract $\sqrt{2}\delta_{2k}/(1-\delta_{2k})\|h_{T_0\cup T_1}\|_2$ to arrive at

$$\|h_{T_0\cup T_1}\|_2 \le (1-\rho)^{-1}(\alpha\eta + 2\rho k^{-1/2}\sigma_k(x)_1), \tag{1.24}$$

where

$$\alpha = \frac{2\sqrt{1+\delta_{2k}}}{1-\delta_{2k}} \quad \text{and} \quad \rho = \frac{\sqrt{2}\delta_{2k}}{1-\delta_{2k}}. \tag{1.25}$$

We conclude the proof by using this estimate and (1.23)

$$\begin{aligned}\|h\|_2 &\le \|h_{(T_0\cup T_1)^c}\|_2 + \|h_{T_0\cup T_1}\|_2 \le \sum_{j\ge 2}\|h_{T_j}\|_2 + \|h_{T_0\cup T_1}\|_2 \\ &\le 2\|h_{T_0\cup T_1}\|_2 + 2k^{-1/2}\sigma_k(x)_1 \le C\frac{\sigma_k(x)_1}{\sqrt{k}} + D\eta\end{aligned}$$

with $C = 2(1-\rho)^{-1}\alpha$ and $D = 2(1+\rho)(1-\rho)^{-1}$.

We shall give more details on stability and robustness of compressed sensing in Section 1.4.3.

1.3.6 Optimality of bounds

When recovering k-sparse vectors one obviously needs at least $m \geq k$ linear measurements. Even when the support of the unknown vector would be known, this number of measurements would be necessary to identify the value of the non-zero coordinates. Therefore, the dependence of the bound (1.17) on k can possibly only be improved in the logarithmic factor. We shall show that even that is not possible and that this dependence is already optimal as soon as a stable recovery of k-sparse vectors is requested. The approach presented here is essentially taken over from [40].

The proof is based on the following combinatorial lemma.

Lemma 1.5. *Let $k \leq n$ be two natural numbers. Then there are N subsets $T_1, \dots, T_N$ of $\{1, \dots, n\}$, such that*

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

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

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

Proof. We may assume that $k \leq n/4$, otherwise one can take $N = 1$ and the statement becomes trivial. The main idea of the proof is straightforward (and similar to the proof of Lemma 1.3). We choose the sets $T_1, T_2, \dots$ inductively one after another as long as possible, satisfying (ii) and (iii) on the way, and then we show that this process will run for at least N steps with N fulfilling (i).

Let $T_1 \subset \{1, \dots, n\}$ be any set with k elements. The number of subsets of $\{1, \dots, n\}$ with exactly k elements, whose intersection with T_1 has at least $k/2$ elements is bounded by the product of 2^k (i.e. the number of all subsets of T_1) and $\binom{n-k}{\lfloor k/2 \rfloor}$, which is the number of all subsets of T_1^c with at most $k/2$ elements. Therefore there are at least

$$\binom{n}{k} - 2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

sets $T \subset \{1, \dots, n\}$ with k elements and $|T \cap T_1| < k/2$. We select T_2 to be any of them. After the jth step, we have selected sets $T_1, \dots, T_j$ with (ii) and (iii) and there are still

$$\binom{n}{k} - j2^k \binom{n-k}{\lfloor k/2 \rfloor}$$

to choose from. The process stops if this quantity is not positive any more, i.e. after at least

$$\begin{aligned} N &\geq \frac{\binom{n}{k}}{2^k \binom{n-k}{\lfloor k/2 \rfloor}} \geq 2^{-k} \frac{\binom{n}{k}}{\binom{n-\lceil k/2 \rceil}{\lfloor k/2 \rfloor}} = 2^{-k} \frac{n!}{(n-k)!k!} \cdot \frac{(\lfloor k/2 \rfloor)!(n-k)!}{(n-\lceil k/2 \rceil)!} \\ &= 2^{-k} \frac{n(n-1)\dots(n-\lceil k/2 \rceil+1)}{k(k-1)\dots(k-\lceil k/2 \rceil+1)} \geq 2^{-k} \left(\frac{n}{k}\right)^{\lceil k/2 \rceil} \geq \left(\frac{n}{4k}\right)^{k/2} \end{aligned}$$

steps.

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

Theorem 1.7. *Let $k \le m \le n$ be natural numbers, let $A \in \mathbb{R}^{m\times n}$ be a measurement matrix, and let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ be an arbitrary recovery map such that for some constant $C > 0$*

$$\|x - \Delta(Ax)\|_2 \le C\frac{\sigma_k(x)_1}{\sqrt{k}} \quad \textit{for all} \quad x \in \mathbb{R}^n. \tag{1.26}$$

Then

$$m \ge C'k\ln(en/k) \tag{1.27}$$

with some other constant C' depending only on C.

Proof. We may assume that $C \ge 1$. Furthermore, if k is proportional to n (say $k \ge n/8$), then (1.27) becomes trivial. Hence we may also assume that $k \le n/8$.

By Lemma 1.5, there exists index sets $T_1, \dots, T_N$ with $N \ge (n/4k)^{k/2}$, $|T_i| = k$ and $|T_i \cap T_j| < k/2$ if $i \ne j$. We put $x_i = \chi_{T_i}/\sqrt{k}$. Then $\|x_i\|_2 = 1$, $\|x_i\|_1 = \sqrt{k}$ and $\|x_i - x_j\|_2 > 1$ for $i \ne j$.

Let

$$\mathscr{B} = \left\{ z \in \mathbb{R}^n : \|z\|_1 \le \frac{\sqrt{k}}{4C} \quad \text{and} \quad \|z\|_2 \le 1/4 \right\}.$$

Then $x_i \in 4C \cdot \mathscr{B}$ for all $i = 1, \dots, N$.

We claim that the sets $A(x_i + \mathscr{B})$ are mutually disjoint. Indeed, let us assume that this is not the case. Then there is a pair of indices $i, j \in \{1, \dots, n\}$ and $z, z' \in \mathscr{B}$ with $i \ne j$ and $A(x_i + z) = A(x_j + z')$. It follows that $\Delta(A(x_i + z)) = \Delta(A(x_j + z'))$ and we get a contradiction by

$$\begin{aligned}
1 &< \|x_i - x_j\|_2 = \|(x_i + z - \Delta(A(x_i + z)) - (x_j + z' - \Delta(A(x_j + z')) - z + z')\|_2 \\
&\le \|(x_i + z - \Delta(A(x_i + z))\|_2 + \|x_j + z' - \Delta(A(x_j + z'))\|_2 + \|z\|_2 + \|z'\|_2 \\
&\le C\frac{\sigma_k(x_i + z)_1}{\sqrt{k}} + C\frac{\sigma_k(x_j + z')_1}{\sqrt{k}} + \|z\|_2 + \|z'\|_2 \\
&\le C\frac{\|z\|_1}{\sqrt{k}} + C\frac{\|z'\|_1}{\sqrt{k}} + \|z\|_2 + \|z'\|_2 \le 1.
\end{aligned}$$

Furthermore,

$$A(x_i + \mathscr{B}) \subset A((4C + 1)\mathscr{B}), \quad i = 1, \dots, N$$

Let $d \le m$ be the dimension of the range of A. We denote by $V \ne 0$ the d-dimensional volume of $A(\mathscr{B})$ and compare the volumes

$$\sum_{j=1}^{N} \operatorname{vol}\big(A(x_j + \mathscr{B})\big) \le \operatorname{vol}\big(A((4C + 1)\mathscr{B})\big).$$

Using linearity of A, we obtain

$$\left(\frac{n}{4k}\right)^{k/2} V \leq N \cdot V \leq (4C+1)^d V \leq (4C+1)^m V.$$

We divide by V and take the logarithm to arrive at

$$\frac{k}{2}\ln\left(\frac{n}{4k}\right) \leq m\ln(4C+1). \tag{1.28}$$

If $k \leq n/8$, then it is easy to check that there is a constant $c' > 0$, such that

$$\ln\left(\frac{n}{4k}\right) \geq c' \ln\left(\frac{en}{k}\right).$$

Putting this into (1.28) finishes the proof. □

1.4 Extensions

Section 1.3 gives a detailed overview of the most important features of compressed sensing. On the other hand, inspired by many questions coming from application driven research, various additional aspects of the theory were studied in the literature. We present here few selected extensions of the ideas of compressed sensing, which turned out to be the most useful in practice. To keep the presentation reasonable short, we do not give any proofs, and only refer to relevant sources.

1.4.1 Frames and Dictionaries

We have considered in Section 1.3 vectors $x \in \mathbb{R}^n$, which are sparse with respect to the natural canonical basis $\{e_j\}_{j=1}^n$ of $\mathbb{R}^n$. In practice, however, the signal has a sparse representation with respect to a basis (or, more general, with respect to a frame or dictionary). Let us first recall some terminology.

A set of vectors $\{\phi_j\}_{j=1}^n$ in $\mathbb{R}^n$, which is linearly independent and which spans the whole space $\mathbb{R}^n$ is called a basis. It follows easily that such a set necessarily has n elements. Furthermore, every $x \in \mathbb{R}^n$ can be expressed uniquely as a linear combination of the basis vectors, i.e. there is a unique $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$, such that

$$x = \sum_{j=1}^n c_j \phi_j. \tag{1.29}$$

A basis is called orthonormal, if it satisfies the orthogonality relations

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1.30}$$

If $\{\phi\}_{j=1}^n$ is an orthonormal basis and $x \in \mathbb{R}^n$, then the decomposition coefficients c_j in (1.29) are given by $c_j = \langle x, \phi_j \rangle$. Furthermore, the relation

$$\|x\|_2^2 = \sum_{j=1}^n |c_j|^2 \tag{1.31}$$

holds true.

Equations (1.29)–(1.30) can be written also in matrix notation. If Φ is an $n \times n$ matrix with j-th column equal to ϕ_j, then (1.29) becomes $x = \Phi c$ and (1.30) reads $\Phi^T \Phi = I$, where I denoted the $n \times n$ identity matrix. As a consequence, $c = \Phi^T x$. We shall say that x has sparse or compressible representation with respect to the basis $\{\phi_j\}_{j=1}^n$ if the vector $c \in \mathbb{R}^n$ is sparse or compressible, respectively.

To allow for more flexibility in representation of signals, it is often useful to drop the condition of linear independence of the set $\{\phi_j\}_{j=1}^N \subset \mathbb{R}^n$. As before, we represent such a system of vectors by a $n \times N$ matrix Φ. We say that $\{\phi_j\}_{j=1}^N$ is a frame, if there are two positive finite constants $0 < A \leq B$, such that

$$A\|x\|_2^2 \leq \sum_{j=1}^N |\langle x, \phi_j \rangle|^2 \leq B\|x\|_2^2. \tag{1.32}$$

From $A > 0$, it follows that the span of the frame vectors is the whole $\mathbb{R}^n$ and, therefore, that $N \geq n$. If one can choose $A = B$ in (1.32), then the frame is called tight. Dual frame of Φ is any other frame $\tilde{\Phi}$ with

$$\Phi \tilde{\Phi}^T = \tilde{\Phi} \Phi^T = I. \tag{1.33}$$

In general, for a given signal $x \in \mathbb{R}^n$ we can find infinitely many coefficients c, such that $x = \Phi c$. Actually, if $\tilde{\Phi}$ is a dual frame to Φ, one can take $c = \tilde{\Phi}^T x$. One is often interested in finding a vector of coefficients c with $x = \Phi c$, which is optimal in some sense. Especially, we shall say that x has a sparse or compressible representation with respect to the frame $\{\phi_j\}_{j=1}^N$ if c can be chosen sparse or compressible, cf. [33].

It can be shown that the smallest coefficient sequence in the ℓ_2^N sense is obtained by the choice $c = \Phi^\dagger x$, where $\Phi^\dagger$ is the Penrose pseudoinverse. In this context, $\Phi^\dagger$ is also called the canonical dual frame. Finally, let us note that (1.33) implies that

$$\sum_{j=1}^N \langle x, \phi_j \rangle \tilde{\phi}_j = \sum_{j=1}^N \langle x, \tilde{\phi}_j \rangle \phi_j = x$$

for every $x \in \mathbb{R}^n$.

The theory of compressed sensing was extended to the setting of sparse representations with respect to frames and dictionaries in [59]. The measurements now take the form $y = Ax = A\Phi c$, where c is sparse. Essentially, it turns out that if A satisfies the concentration inequalities from Section 1.3.4 and the dictionary Φ has

small coherence, then the matrix $A\Phi$ has small RIP constants, and the methods of compressed sensing can be applied.

1.4.2 Coherence

We have provided in Section 1.3.4 a simple recipe how to construct matrices with small RIP constants - namely to choose each entry independently at random with respect to a correctly normalized standard distribution. On the other hand, if the matrix A is given beforehand, it is quite difficult to check if this matrix really satisfies the RIP, or to calculate its RIP constants. Another property of A, which is easily verifiable and which also ensures good recovery guarantees, is the coherence of A.

Definition 1.3. Let A be a $m \times n$ matrix and let $a_1, \dots, a_n \in \mathbb{R}^m$ be its columns. Then the coherence of A is the number $\mu(A)$ defined as

$$\mu(A) = \max_{1 \le i < j \le n} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2}. \tag{1.34}$$

Due to Cauchy-Schwartz inequality, $\mu(A) \le 1$ is always true. If $m \le n$, then there is a lower bound (known as the Welch bound [70]) on the coherence given by $\mu(A) \ge \sqrt{\frac{n-m}{m(n-1)}}$. Let us observe that if $n \gg m$, then this bound reduces to approximately $\mu(A) \ge 1/\sqrt{m}$. There is a lot of possible ways how to construct matrices with small coherence. Not surprisingly, one possible option is to consider random matrices A with each entry generated independently at random, cf. [57, Chapter 11]. Nevertheless the construction of matrices achieving the Welch bound exactly is still an active area of research, making use of ideas from algebra and number theory. On the other hand, it is easy to show that the Welch bound can not be achieved if n is much larger then m. It can be done only if $n \le m(m+1)/2$ in the real case, and if $n \le m^2$ in the complex case.

The connection of coherence to RIP is given by the following Lemma.

Lemma 1.6. *If A has unit-norm columns and coherence $\mu(A)$, then it satisfies the RIP of order k with $\delta_k(A) \le (k-1)\mu(A)$ for all $k < 1/\mu(A)$.*

Combining this with Theorem 1.5, it gives recovery guarantees for the number of measurements m growing quadratically in the sparsity k.

1.4.3 Stability and Robustness

Basic discussion of stability and robustness of the methods of compressed sensing was given already in Section 1.3.5 with Theorem 1.6 being the most important representative of the variety of noise-aware estimates in the area. Its proof follows

closely the presentation of [11]. The proof can be easily transformed to the spirit of Section 1.3.2 and 1.3.3 using the following modification of the Null Space Property.

Definition 1.4. We say that $A \in \mathbb{R}^{m\times n}$ satisfies the ℓ_2*-Robust Null Space Property* of order k with constants $0 < \rho < 1$ and $\tau > 0$ if

$$\|v_T\|_2 \leq \frac{\rho\|v_{T^c}\|_1}{\sqrt{k}} + \tau\|Av\|_2 \tag{1.35}$$

for all $v \in \mathbb{R}^n$ and all sets $T \subset \{1,\dots,n\}$ with $|T| \leq k$.

The following theorem (which goes essentially back to [15]), is then the noise-aware replacement of Theorem 1.2.

Theorem 1.8. *Let $A \in \mathbb{R}^{m\times n}$ with ℓ_2-Robust Null Space Property of order k with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $x \in \mathbb{R}^n$ the solution $\hat{x}$ of $(P_{1,\eta})$ with $y = Ax + e$ and $\|e\|_2 \leq \eta$ satisfies*

$$\|x - \hat{x}\|_2 \leq \frac{C}{\sqrt{k}}\sigma_k(x)_1 + D\eta \tag{1.36}$$

with constants $C, D > 0$ depending only on ρ and τ.

Finally, it turns out that the Restricted Isometry Property is also sufficient to guarantee the ℓ_2-Robust Null Space Property and Theorem 1.3 can be extended to

Theorem 1.9. *Let $A \in \mathbb{R}^{m\times n}$ and let k be a natural number with $k \leq n/2$. If $\delta_{2k}(A) < 1/3$, then A satisfies the ℓ_2-Robust Null Space Property of order k with constants $0 < \rho < 1$ and $\tau > 0$ depending only on $\delta_{2k}(A)$.*

Let us only point out, that the constant $1/3$ is by no means optimal, and that the same result (with more technical analysis) holds also if $\delta_{2k}(A) < 4/\sqrt{41}$, cf. [9, 10, 38, 39].

Theorems 1.6 and 1.8 are sufficient to analyze the situation, when the noise is bounded in the ℓ_2-norm, no matter what the structure of the noise is. If we assume, that the noise is Gaussian, i.e. that $e = (e_1,\dots,e_m)$, where e_i's are independent normal variables, then the estimate (1.36) suffers from the following drawback. If we increase the number of measurements m, then also the expected value of $\|e\|_2$ increases and, therefore, the estimate (1.36) actually becomes *worse*.

To deal with this issue, the following recovery algorithm, called *Dantzig selector*

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|A^*(Az - y)\|_\infty \leq \tau, \tag{1.37}$$

was proposed and analyzed in [17]. It deals with the case, when $\|A^T e\|_2$ is small.

Theorem 1.10. *Let A be a $m \times n$ matrix with RIP of order $2k$ and $\delta_{2k} < \sqrt{2} - 1$. Let the measurements y take the form $y = Ax + e$, where $\|A^T e\|_\infty \leq \tau$. Then the solution $\hat{x}$ of* (1.37) *satisfies*

$$\|\hat{x}-x\|_2 \leq \frac{C}{\sqrt{k}}\sigma_k(x)_1 + D\sqrt{k}\tau, \tag{1.38}$$

where $C, D > 0$ *depend only on* $\delta_{2k}(A)$.

To see, how this is related to measurements corrupted with Gaussian noise, let us assume that the components of $e \in \mathbb{R}^m$ are i.i.d. normal variables with variance σ^2. If A has also unit-norm columns, then the 2-stability of normal variables gives that each coordinate of $A^T e$ is also a normal variable with mean zero and variance σ^2. Then we obtain

$$\mathbb{P}(|[A^T e]_i| \geq t\sigma) \leq 2\exp(-t^2/2)$$

and using the union bound this becomes

$$\mathbb{P}\big(\|A^T e\|_\infty \geq 2\sqrt{\ln n}\sigma\big) \leq 2n\exp(-2\ln n) = \frac{2}{n}. \tag{1.39}$$

Combining this with Theorem 1.10, we obtain for the case of exactly sparse vectors the following theorem.

Theorem 1.11. *Let A be a $m \times n$ matrix with unit-norm columns and with RIP of order $2k$ and $\delta_{2k} < \sqrt{2}-1$. Let the measurements y take the form $y = Ax + e$, where the entries of e are i.i.d. normal variables with variance σ^2. Then the solution $\hat{x}$ of* (1.37) *with $\tau = 2\sqrt{\ln n}\sigma$ satisfies*

$$\|\hat{x}-x\|_2 \leq C\sqrt{k\ln n}\sigma \tag{1.40}$$

with probability at least $1-2/n$.

Observe that (1.40) depends only on the sparsity level of x and not on m any more.

1.4.4 Recovery algorithms

Although we concentrated on ℓ_1-minimization in the first part of this chapter, there is a number of different algorithms solving the problem of sparse signal recovery. Similarly to ℓ_1-minimization, which was used successfully in machine learning much before the advent of compressed sensing, many of these algorithms also predate the field of compressed sensing. We give an overview of some of these algorithms and refer to [40] for more extensive treatment.

1.4.4.1 ℓ_1-minimization

The ℓ_1-minimization problems (P_1) or $(P_{1,\eta})$ presented before form a backbone of the theory of compressed sensing. Their geometrical background allows for theoretical recovery guarantees, including corresponding stability and robustness extensions. They are formulated as convex optimization problems, which can be solved

effectively by any general purpose numerical solver. Furthermore, several implementations dealing with the specific setting of compressed sensing are available nowadays.

Sometimes, it is more convenient to work with some of the equivalent reformulations of $(P_{1,\eta})$. Let us discuss two most important of them. Let $\eta \geq 0$ be given and let $\hat{x}$ be a solution of the optimization problem $(P_{1,\eta})$

$$\hat{x} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{s.t.} \quad \|Az - y\|_2 \leq \eta. \tag{$P_{1,\eta}$}$$

Then there is a $\lambda \geq 0$, such that $\hat{x}$ is also a solution of the non-constrained convex problem

$$\hat{x} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \frac{1}{2}\|Az - y\|_2^2 + \lambda \|z\|_1. \tag{1.41}$$

This version of ℓ_1-minimization is probably the mostly studied one, see, for example, [34, 41, 50, 72]. On the other hand, if $\lambda > 0$ is given and $\hat{x}$ is a solution to (1.41), then there is an $\eta > 0$, such that $\hat{x}$ is also a solution of $(P_{1,\eta})$. In the same sense, $(P_{1,\eta})$ and (1.41) is also equivalent to *Lasso* (least absolute shrinkage and selection operator, cf. [63])

$$\hat{x} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \|Az - y\|_2^2 \quad \text{s.t.} \quad \|z\|_1 \leq \tau. \tag{1.42}$$

Unfortunately, the values of λ and $\tau > 0$ making these problems equivalent are a-priori unknown.

The last prominent example of an optimization problem, which takes a form of ℓ_1-minimization is the Dantzig selector (1.37). Let us also point out, that [7] provides solvers for a variety of ℓ_1-minimization problems.

1.4.4.2 Greedy algorithms

Another approach to sparse recovery is based on iterative identification/approximation of the support of the unknown vector x and of its components. For example, one adds in each step of the algorithm one index to the support to minimize the mismatch to the measured data as much as possible. Therefore, such algorithms are usually referred to as greedy algorithms. For many of them, remarkable theoretical guarantees are available in the literature, sometimes even optimal in the sense of the lower bounds discussed above. Nevertheless, the techniques necessary to achieve these results are usually completely different from those needed to analyze ℓ_1-minimization. We will discuss three of these algorithms, *Orthogonal Matching Pursuit*, *Compressive Sampling Matching Pursuit* and *Iterative Hard Thresholding*.

Orthogonal Matching Pursuit(OMP)

Orthogonal Matching Pursuit [52, 64, 66] adds in each iteration exactly one entry into the support of $\hat{x}$. After k iterations, it therefore outputs a k-sparse vector $\hat{x}$.

The algorithm finds in each step the column of A most correlated with the residual of the measurements. Its index is then added to the support. Finally, it updates the target vector $\hat{x}_i$ as the vector supported on T_i that best fits the measurements, i.e. which minimizes $\|y - Az\|_2$ among all $z \in \mathbb{R}^n$ with supp $(z) \subset T_i$. It is well known, that this vector is given as the the product of the Penrose pseudoinverse $A^\dagger$ of A and y.

The formal transcription of this algorithm is given as follows.

Orthogonal Matching Pursuit (OMP)

Input: Compressed sensing matrix A, measurement vector y
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until stopping criterion is met

$i := i + 1$	
$T_i \leftarrow T_{i-1} \cup \text{supp } H_1(A^T r)$	add largest residual entry to the support
$\hat{x}_i\|_{T_i} \leftarrow A_{T_i}^\dagger y$	update the estimate of the signal
$r \leftarrow y - A\hat{x}_i$	update the residual of the measurements

Output: $\hat{x}_i$

It makes use of the hard thresholding operator $H_k(x)$. If $x \in \mathbb{R}^n$ and $k \in \{0, 1, \ldots, n\}$, then $H_k : x \to H_k(x)$ associates to x a vector $H_k(x) \in \mathbb{R}^n$, which is equal to x on the k entries of x with largest magnitude and zero otherwise. The stopping criteria can either limit the overall number of iteration (limiting also the size of the support of the output vector $\hat{x}$), or ensure, that the distance between y and $A\hat{x}$ is small in some norm.

The simplicity of OMP is unfortunately connected with one of its weak points. If an incorrect index is added to the support in some step (which can happen in general and depends on the properties of the input parameters), it can not be removed any more, and stays there until the end of OMP. We refer also to [26] for another variant of OMP.

Compressive Sampling Matching Pursuit (CoSaMP)

One attempt to overcome this drawback is presented in the following algorithm called *Compressive Sampling Matching Pursuit* [56]. It assumes, that an additional input is given - namely the expected sparsity of the output. At each step it again enlarges the support, but in contrast to OMP, it will add at least k new entries. Afterwards, it again uses the Penrose pseudo-inverse to find the minimizer of $\|Az - y\|_2$ among all $z \in \mathbb{R}^n$ with supp $(z) \subset T_i$, but this time only the k largest of coordinates of this minimizer are stored.

The formal description is given by the following scheme.

Compressive Sampling Matching Pursuit (CoSaMP)

Input: Compressed sensing matrix A, measurement vector y, sparsity level k
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until stopping criterion is met

$i := i+1$	
$T_i \leftarrow \text{supp}\,(\hat{x}_{i-1}) \cup \text{supp}\, H_{2k}(A^T r)$	update the support
$\hat{x}_i\|_{T_i} \leftarrow H_k(A_{T_i}^{\dagger} y)$	update the estimate of the signal
$r \leftarrow y - A\hat{x}_i$	update the residual

Output: $\hat{x}_i$

Iterative Hard Thresholding (IHT)

The last algorithm [8] we shall discuss is also making use of the hard thresholding operator H_k. The equation $Az = y$ is transformed into $A^T Az = A^T y$, which again can be interpreted as looking for the fixed point of the mapping $z \to (I - A^T A)z + A^T y$. Classical approach is then to iterate this mapping and to put $\hat{x}_i = (I - A^T A)\hat{x}_{i-1} + A^T y = \hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1})$. Iterative Hard Thresholding algorithm is doing exactly this, only combined with the hard thresholding operator H_k.

Iterative Hard Thresholding (IHT)

Input: Compressed sensing matrix A, measurement vector y, sparsity level k
Initial values: $\hat{x}_0 = 0, i = 0$
Iteration step: Repeat until stopping criterion is met

$i := i+1$	
$\hat{x}_i = H_k(\hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1}))$	update the estimate of the signal

Output: $\hat{x}_i$

1.4.4.3 Combinatorial algorithms

The last class of algorithms for sparse recovery we shall review, were developed mainly in the context of theoretical computer science and they are based on classical ideas from this field, which usually pre-date the area of compressed sensing. Nevertheless, they were successfully adapted to the setting of compressed sensing.

Let us present the basic idea on the example of Group Testing, which was introduced by Robert Dorfman [27] in 1943. One task of United States Public Health Service during the Second World War was to identify all syphilitic soldiers. However, syphilis test in that time was expensive and the naive approach of testing every soldier independently would have been very costly.

If the portion of infected soldiers would be large (say above 50 percent) then the method of individual testing would be reasonable (and nearly optimal). A realistic assumption however is that only a tiny fraction of all the soldiers is infected, say one in thousand, or one in ten thousand. The main idea of the area of Group Testing in this setting is that we can combine blood samples and test a combined sample

to check if at least one soldier in the group has syphilis. Another example of this technique is the false coin problem from recreational mathematics, in which one is supposed to identify in a group of n coins a false coin weighting less than a real coin. We refer to [28] to an overview of the methods of Group Testing.

To relate this problem to compressed sensing, let us consider a vector $x = (x_1, \dots, x_n) \in \{0,1\}^n$, where n is the number of soldiers, with $x_i = 0$ if the ith soldier is healthy, or $x_i = 1$ if he has syphilis. The grouping is then represented by a $m \times n$ matrix $A = (a_{ij})$, where $a_{ij} = 1$, if the blood sample of jth soldier was added to ith combined sample. The methods of Group Testing then allow to design efficient matrices A, such that the recovery of x can be done in a surprisingly small number of steps - even linear in the length of the sparse representation of x, i.e. in its sparsity k, cf. [43, 44].

1.4.5 Structured sparsity

In many applications, one has much more prior knowledge about the signal x, than just assuming that it possesses a sparse representation with respect to certain basis, frame, or dictionary.

For example, the image coder JPEG2000 exploits not only the fact that natural images have compressible representation in the wavelet basis (i.e. that most of their wavelet coefficients are small) but it also uses the fact that the values and locations of the large coefficients have a special structure. It turns out that they tend to cluster into a connected subtree inside the wavelet parent-child tree. Using this additional information can of course help to improve the properties of the coder and provide better compression rates [30, 31, 47].

Another model appearing frequently in practice is the model of block-sparse (or joint-sparse) signals. Assume, that we want to recover N correlated signals $x^1, \dots, x^N \in \mathbb{R}^n$ with (nearly) the same locations of their most significant elements. A simple example of such a situation are the three color channels of a natural RGB image, where we intuitively expect the important wavelet coefficients in all three channels to be on nearly the same locations. Furthermore, the same model often appears in the study of DNA microarrays, magnetoencephalography, sensor networks and MIMO communication [6, 32, 62, 69]. It is usually convenient to represent the signals as columns of a $n \times N$ matrix $X = [x^1 \dots x^N]$. The recovery algorithms are then based on mixed matrix norms, which are defined for such an X as

$$\|X\|_{(p,q)} = \Big(\sum_{i=1}^{n} \|\tilde{x}^i\|_p^q \Big)^{1/q},$$

where $p, q \geq 1$ are real numbers and $\tilde{x}^i$, $i = 1, \dots, n$, are the rows of the matrix X. If A is again the sensing matrix and $Y = AX$ are the measurements, then the analogue of (P_1) in this setting is then

$$\hat{X} = \underset{Z \in \mathbb{R}^{n \times N}}{\operatorname{argmin}} \|Z\|_{(p,q)} \quad \text{s. t.} \quad Y = AZ$$

for a suitable choice of p and q, typically $(p,q) = (2,1)$. We refer for example to [36, 65, 67] for further results.

Finally, let us point out that *model-based compressive sensing* [3] provides a general framework for many different kinds of structured sparsity.

1.4.6 Compressed Learning

In this last part, we will discuss applications of compressed sensing to a classical task of approximation theory, namely to learning of an unknown function f from a limited number of its samples $f(x^1), \ldots, f(x^m)$. In its most simple form, treated already in [13] and elaborated in [58], one assumes that the function f is known to be a sparse combination of trigonometric polynomials of maximal order q in dimension d, i.e. that

$$f(x) = \sum_{l \in \{-q, -q+1, \ldots, q-1, q\}^d} c_l e^{il \cdot x}$$

and $\|c\|_0 \leq k$, where $k \in \mathbb{N}$ is the level of sparsity. Theorem 2.1 of [58] then shows that, with probability at least $1 - \varepsilon$, f can be exactly recovered from samples $f(x^1), \ldots, f(x^m)$, where $m \geq Ck \ln((2q+1)^d/\varepsilon)$ and $x_1, \ldots, x_m$ are uniformly and independently distributed in $[0, 2\pi]^d$. The recovery algorithm is given by

$$\underset{c}{\operatorname{argmin}} \|c\|_1 \quad \text{s. t.} \quad \sum_l c_l e^{il \cdot x^j} = f(x^j), \ j = 1, \ldots, m.$$

We refer to [12, 60] for further results and to [40, Chapter 12] for an overview on random sampling of functions with sparse representation in a bounded orthonormal system.

In another line of study, compressed sensing was used to approximate functions $f : [0,1]^d \to \mathbb{R}$, which depend only on $k \ll d$ (unknown) *active variables* $i_1, \ldots, i_k$, i.e.

$$f(x) = f(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_k}), \quad x \in [0,1]^d.$$

In [24] and [71], the authors presented sophisticated combinatorial (adaptive and non-adaptive) constructions of sets of sampling points, which allowed for recovery of f to a precision of $1/L$ using only $C(k)(L+1)^k \ln d$ points. Observe, that $(L+1)^k$ points would be necessary even if the location of the active coordinates would be known. The use of compressed sensing in this setting was then discussed in [61]. The algorithm developed there was based on approximation of directional derivatives of f at random points $\{x^1, \ldots, x^{m_X}\}$ and random directions $\{\varphi^1, \ldots, \varphi^{m_\Phi}\}$. Denoting the $m_\Phi \times m_X$ matrix of first order differences as Y and the $m_\Phi \times d$ matrix of random directions by Φ, it was possible to use direct estimates of probability concentrations

to ensure, that the k largest rows of $\Phi^T Y$ correspond to the k active coordinates of f with high probability. Again, only an additional $\ln d$ factor is paid for identifying the unknown active coordinates.

Finally, the paper [21] initiated a study of approximation of ridge functions of the type

$$f(x) = g(\langle a, x\rangle), \quad x \in [0,1]^d, \tag{1.43}$$

where both the direction $a \in \mathbb{R}^d \setminus \{0\}$ and the univariate function g are unknown. Due to the assumption $a_j \geq 0$ for all $j = 1, \dots, d$, posed in [21], it was first possible to approximate g by sampling on grid points along the diagonal $\{\frac{i}{L}(1, \dots, 1)^T, i = 0, \dots, L\}$. Afterwards, the methods of compressed sensing were used in connection with the first order differences to identify the vector a. The importance of derivatives of f in connection with the assumption (1.43) is best seen from the simple formula

$$\nabla f(x) = g'(\langle a, x\rangle) \cdot a. \tag{1.44}$$

Hence, approximating the gradient of f at a point x gives actually also a scalar multiple of a.

Another algorithm to approximate the ridge functions was proposed in [37]. Similarly to [61], it was based on (1.44) and on approximation of the first order derivatives by first order differences. In contrary to [21], first the ridge direction a was recovered, and only afterwards the ridge profile g was approximated by any standard one-dimensional sampling scheme. Furthermore, no assumptions on signs of a was needed and it was possible to generalize the approach also for recovery of k-ridge functions of the type $f(x) = g(Ax)$, where $A \in \mathbb{R}^{k \times d}$ and g is a function of k variables. We refer also to [18] for further results.

References

1. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J. Comput. System Sci. **66**, 671–687 (2003)
2. Arora, S., Barak, B.: Computational Complexity: A Modern Approach. Cambridge Univ. Press, Cambridge (2009)
3. Baraniuk, R., Cevher, V., Duarte, M.F., Hegde, C.: Model-based compressive sensing, IEEE Trans. Inform. Theory **56** 1982–2001 (2010)
4. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**, 253–263 (2008)
5. Baraniuk, R., Steeghs, P.: Compressive radar imaging. In Proc. IEEE Radar Conf., Boston, 128–133 (2007)
6. Baron, D., Duarte, M.F., Sarvotham, S., Wakin, M.B., Baraniuk, R., Distributed compressed sensing of jointly sparse signals. In Proc. Asilomar Conf. Signals, Systems, and Computers, Pacic Grove, CA (2005)
7. Becker, S., Candés, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Math. Prog. Comp. **3**, 165–218 (2010)
8. Blumensath, T., Davies, M.: Iterative hard thresholding for compressive sensing. Appl. Comput. Harmon. Anal. **27**, 265–274 (2009)

9. Cai, T., Wang, L., Xu, G.: New bounds for restricted isometry constants. IEEE Trans. Inform. Theor. **56**, 4388–4394 (2010)
10. Cai, T., Wang, L., Xu, G.: Shifting inequality and recovery of sparse vectors. IEEE Trans. Signal Process. **58**, 1300–1308 (2010)
11. Candés, E.J.: The restricted isometry property and its implications for compressed sensing. Compte Rendus de l'Academie des Sciences, Paris, Serie I, **346**, 589–592 (2008)
12. Candés, E.J., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. IEEE Trans. Inform. Theory **57**, 7235–7254 (2011)
13. Candés, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52**, 489–509 (2006)
14. Candés, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inform. Theory **51**, 4203–4215 (2005)
15. Candés, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. **59**, 1207–1223 (2006)
16. Candés, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory **52**, 5406–5425 (2006)
17. Candés, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. **35**, 2313–2351 (2007)
18. Cevher, V., Tyagi, H.: Active learning of multi-index function models. In: Proc. NIPS (The Neural Information Processing Systems), Lake Tahoe, Reno, Nevada, (2012)
19. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
20. Cohen, A., Dahmen, W., DeVore, R.: Compressed sensing and best k-term approximation. J. Amer. Math. Soc. **22**, 211–231 (2009)
21. Cohen, A., Daubechies, I., DeVore, R., Kerkyacharian, G., Picard, D.: Capturing ridge functions in high dimensions from point queries. Constr. Approx. **35**, 225–243 (2012)
22. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures Algorithms **22**, 60–65 (2003)
23. Davenport, M.A., Duarte, M.F., Eldar, Y.C., Kutyniok, G.: Introduction to compressed sensing. Compressed sensing, 1–64, Cambridge Univ. Press, Cambridge, (2012)
24. DeVore, R., Petrova, G., Wojtaszczyk, P.: Approximation of functions of few variables in high dimensions. Constr. Approx. **33**, 125–143 (2011)
25. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory **52**, 1289–1306 (2006)
26. Donoho, D.L., Tsaig, Y., Drori, I., Starck, J.-L.: Sparse solution of underdetermined systems of linear equations by stagewise Orthogonal Matching Pursuit. IEEE Trans. Inform. Theory **58**, 1094–1121 (2012)
27. Dorfman, R.: The detection of defective members of large populations. Annals Math. Stat. **14**, 436–440 (1943)
28. Du, D., Hwang, F.: Combinatorial group testing and its applications. World Scientic, Singapore (2000)
29. Duarte, M., Davenport, M., Takhar, D., Laska, J., Ting, S., Kelly, K., Baraniuk R.: Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. **25**, 83–91 (2008)
30. Duarte, M., Wakin, M., Baraniuk, R.: Fast reconstruction of piecewise smooth signals from random projections. In Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS), Rennes, France (2005)
31. Duarte, M., Wakin, M., Baraniuk, R.: Wavelet-domain compressive signal reconstruction using a hidden Markov tree model. In Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), Las Vegas, NV (2008)
32. Eldar, Y., Mishali, M.: Robust recovery of signals from a structured union of subspaces. IEEE Trans. Inform. Theory **55**, 5302–5316 (2009)
33. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, New York, NY (2010)

34. Figueiredo, M., Nowak, R., Wright, S.: Gradient projections for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE J. Select. Top. Signal Processing **1**, 586–597 (2007)
35. Fornasier, M., Rauhut, H., Compressive Sensing. In: Scherzer, Otmar (Ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, Heidelberg (2011)
36. Fornasier, M., Rauhut, H.: Recovery algorithms for vector valued data with joint sparsity constraints. SIAM J. Numer. Anal. **46**, 577–613 (2008)
37. Fornasier, M., Schnass, K., Vybíral, J.: Learning functions of few arbitrary linear parameters in high dimensions. Found. Comput. Math. **12**, 229262 (2012)
38. Foucart, S.: A note on guaranteed sparse recovery via 1-minimization. Appl. Comput. Harmon. Anal. **29**, 97–103 (2010)
39. Foucart, S., Lai M.: Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. Appl. Comput. Harmon. Anal. **26**, 395–407 (2009)
40. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Birkhäuser/Springer, New York (2013)
41. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stats. Software **33**, 1–22 (2010)
42. Gärtner, B., Matoušek, J.: Understanding and Using Linear Programming, Springer, Berlin (2006)
43. Gilbert, A., Li, Y., Porat, E., and Strauss, M.: Approximate sparse recovery: Optimizaing time and measurements. In Proc. ACM Symp. Theory of Comput., Cambridge, MA (2010)
44. Gilbert, A., Strauss, M., Tropp, J., Vershynin, R.: One sketch for all: Fast algorithms for compressed sensing. In Proc. ACM Symp. Theory of Comput., San Diego, CA (2007)
45. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conf. in Modern Analysis and Probability, pp. 189–206, (1984)
46. Krahmer, F., Ward, R.: New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal. **43**, 1269–1281 (2011)
47. La, C., Do, M.N.: Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In IEEE Int. Conf. Image Processing (ICIP), Atlanta, GA (2006)
48. Ledoux, M.: The concentration of measure phenomenon. American Mathematical Society, Providence, (2001)
49. Ledoux, M., Talagrand, M.: Probability in Banach spaces. Isoperimetry and processes. Springer, Berlin, (1991)
50. Loris, I.: On the performance of algorithms for the minimization of ℓ_1-penalized functions. Inverse Problems **25** 035008 (2009)
51. Lustig, M., Donoho, D., Pauly, J.M.: Sparse MRI: The application of compressed sensing for rapid MR imaging. Magn. Reson. Med. **58**, 1182–1195 (2007)
52. Mallat, S., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Processing **41**, 3397–3415 (1993)
53. Matoušek, J.: On variants of the Johnson-Lindenstrauss lemma. Random Structures Algorithms **33** 142–156 (2008)
54. Milman, V.D., Schechtman, G.: Asymptotic theory of finite-dimensional normed spaces. Springer, Berlin (1986)
55. Mishali, M., Eldar, Y.: From theory to practice: Sub-nyquist sampling of sparse wideband analog signals. IEEE J. Sel. Top. Signal Process. **4**, 375–391 (2010)
56. Needell, D., Tropp, J.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**, 301–321 (2009)
57. Pietsch, A.: Operator ideals. North-Holland Publishing Co., Amsterdam-New York (1980)
58. Rauhut, H.: Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**, 16–42 (2007)
59. Rauhut, H., Schnass, K., Vandergheynst, P.: Compressed sensing and redundant dictionaries. IEEE Trans. Inform. Theor. **54**, 2210–2219 (2008)
60. Rauhut, H., Ward, R.: Sparse Legendre expansions via ℓ_1-minimization. J. Approx. Theory **164**, 517–533 (2012)

61. Schnass, K., Vybíral, J.: Compressed learning of high-dimensional sparse functions. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 3924–3927 (2011)
62. Stojnic, M., Parvaresh, F., Hassibi, B.: On the reconstruction of block-sparse signals with an optimal number of measurements. IEEE Trans. Signal Processing **57**, 3075–3085 (2009)
63. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Royal Statist. Soc B **58**, 267–288 (1996)
64. Tropp, J.: Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inform. Theor. **50**, 2231–2242 (2004)
65. Tropp, J.: Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. Signal Processing, **86**, 589–602 (2006)
66. Tropp, J., Gilbert, A.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inform. Theory **53** 4655–4666 (2007)
67. Tropp, J., Gilbert, A., Strauss, M.: Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Processing, **86**, 572–588 (2006)
68. Tropp, J., Laska, J., Duarte, M., Romberg, J., Baraniuk, R., Beyond Nyquist: Efficient sampling of sparse bandlimited signals. IEEE Trans. Inform. Theor. **56**, 520–544 (2010)
69. Wakin, M.B., Sarvotham, S., Duarte, M.F., Baron, D., Baraniuk, R.: Recovery of jointly sparse signals from few random projections. In Proc. Workshop on Neural Info. Proc. Sys. (NIPS), Vancouver, (2005)
70. Welch, L.: Lower bounds on the maximum cross correlation of signals. IEEE Trans. Inform. Theory **20**, 397–399 (1974)
71. Wojtaszczyk, P.: Complexity of approximation of functions of few variables in high dimensions. J. Complexity **27**, 141–150 (2011)
72. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imag. Sci. **1**, 143–168 (2008)