An Introduction to Compressed Sensing

Niklas Koep, Arash Behboodi and Rudolf Mathar

Abstract Compressed sensing and many research activities associated with it can be seen as a framework for signal processing of low-complexity structures. A cornerstone of the underlying theory is the study of inverse problems with linear or nonlinear measurements. Whether it is sparsity, low-rankness, or other familiar notions of low complexity, the theory addresses necessary and sufficient conditions behind the measurement process to guarantee signal reconstruction with efficient algorithms. This includes consideration of robustness to measurement noise and stability with respect to signal model inaccuracies. This introduction aims to provide an overall view of some of the most important results in this direction. After discussing various examples of low-complexity signal models, two approaches to linear inverse problems are introduced which, respectively, focus on the recovery of individual signals and recovery of all low-complexity signals simultaneously. In particular, we focus on the former setting, giving rise to so-called nonuniform signal recovery problems. We discuss different necessary and sufficient conditions for stable and robust signal reconstruction using convex optimization methods. Appealing to concepts from non-asymptotic random matrix theory, we outline how certain classes of random sensing matrices, which fully govern the measurement process, satisfy certain sufficient conditions for signal recovery. Finally, we review some of the most prominent algorithms for signal recovery proposed in the literature.

N. Koep · A. Behboodi (✉) · R. Mathar
RWTH Aachen Theoretische Informationstechnik, 52056 Aachen, Germany
e-mail: arash.behboodi@ti.rwth-aachen.de

N. Koep
e-mail: niklas.koep@ti.rwth-aachen.de

R. Mathar
e-mail: rudolf.mathar@ti.rwth-aachen.de

© Springer Nature Switzerland AG 2019

H. Boche et al. (eds.), *Compressed Sensing and Its Applications*,
Applied and Numerical Harmonic Analysis,
https://doi.org/10.1007/978-3-319-73074-5_1

1 Introduction

The field of compressed sensing was originally established with the publication of the seminal papers "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information" [25] by Terence Tao, Justin Romberg and Emmanuel Candès, and the aptly titled "Compressed sensing" [40] by David Donoho. The research activity by hundreds of researchers that followed over time transformed the field into a mature mathematical theory with far-reaching implications in applied mathematics and engineering alike. While deemed impossible by the celebrated Shannon–Nyquist sampling theorem, as well as fundamental facts in linear algebra, their work demonstrated that unique solutions of underdetermined systems of linear equations do in fact exist if one limits attention to signal sets exhibiting some type of low-complexity structure. In particular, Tao, Romberg, Candès, and Donoho considered so-called sparse vectors containing only a limited number of nonzero coefficients and demonstrated that solving a simple linear program minimizing the ℓ_1-norm of a vector subject to an affine constraint allowed for an efficient way to recover such signals. While examples of ℓ_1-regularized methods as a means to retrieve sparse estimates of linear inverse problems can be traced back as far as the 1970s to work in seismology, the concept was first put on a rigorous footing in a series of landmark papers [25–28, 40]. Today, compressed sensing is considered a mature field firmly positioned at the intersection of linear algebra, probability theory, convex analysis, and Banach space theory.

This chapter serves as a concise overview of the field of compressed sensing, highlighting some of the most important results in the theory, as well as some more recent developments. In light of the popularity of the field, there truly exists no shortage of excellent surveys and introductions to the topic. We want to point out the following references in particular: [14, 46, 47, 51, 52, 54], which include extended monographs focusing on a rigorous presentation of the mathematical theory, as well as works more focused on the application side, e.g., in the context of wireless communication [60] or more generally in sparse signal processing [29]. Due to the volume of excellent references, we decided on a rather opinionated selection of topics for this introduction. For instance, a notable omission of our text is a discussion on the so-called Gelfand widths, a concept in the theory of Banach spaces that is commonly used in compressed sensing to prove the optimality of bounds on the number of measurements required to establish certain properties of random matrices. Moreover, in the interest of space, we opted to omit most of the proofs in this chapter, and instead make frequent reference to the excellent material found in the literature.

Organization

Given the typical syllabus of introductions to compressed sensing, we decided to go a slightly different route than usual by motivating the underlying problem from an extended view at the problem of individual vector recovery before moving on to the so-called uniform recovery case which deals with the simultaneous recovery of all vectors in a particular signal class at once.

In Sect. 2, we briefly recall a few basic definitions of norms and random variables. We also define some basic notions about so-called *subgaussian* random variables as they play a particularly important role in modern treatments of compressed sensing.

In Sect. 3, we introduce a variety of signal models for different applications and contexts. To that end, we adopt the notion of *simple sets* generated by so-called *atomic sets*, and the associated concept of *atomic norms* which provide a convenient abstraction for the formulation of nonuniform recovery problems in a multitude of different domains. In the context of sparse recovery, we also discuss the important class of so-called *compressible vectors* as a practical alternative to exactly sparse vectors to model real-world signals such as natural images, audio signals, and the like.

Equipped with the concept of the atomic norm which gives rise to a tractable recovery program of central importance in the context of linear inverse problems, we discuss in Sect. 4 conditions for perfect or robust recovery of low-complexity signals. We also comment on a rather recent development in the theory which connects the problem of sparse recovery with the field of conic integral geometry.

Starting with Sect. 5, we finally turn our attention to the important case of uniform recovery of sparse or compressible vectors where we are interested in establishing guarantees which—given a particular measurement matrix—hold uniformly over the entire signal class. Such results stand in stark contrast to the problems we discuss in Sect. 4 where recovery conditions are allowed to locally depend on the choice of the particular vector one aims to recover.

In Sect. 6, we introduce a variety of properties of sensing matrices such as the null space property and the restricted isometry property which are commonly used to assert that recovery conditions as teased in Sect. 5 hold for a particular matrix. While the deterministic construction of matrices with provably optimal number of measurements remains a yet unsolved problem, random matrices—including a broad class of structured random matrices—which satisfy said properties can be shown to exist in abundance. We therefore complement our discussion with an overview of some of the most important classes of random matrices considered in compressed sensing in Sect. 7.

We conclude our introduction to the field of compressed sensing with a short survey of some of the most important sparse recovery algorithms in Sect. 8.

Motivation

At the heart of compressed sensing (CS) lies a very simple question. Given a d-dimensional vector $\mathring{\mathbf{x}}$, and a set of m measurements of the form $y_i = \langle \mathbf{a}_i, \mathring{\mathbf{x}} \rangle$, under what conditions are we able to infer $\mathring{\mathbf{x}}$ from knowledge of

$$\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)^\top \quad \text{and} \quad \mathbf{y} = (y_1, \dots, y_m)^\top$$

alone? Historically, the answer to this question was “as soon as $m \geq d$” or more precisely, as soon as $\operatorname{rank}(\mathbf{A}) = d$. In other words, the number of independent observations of $\mathring{\mathbf{x}}$ has to exceed the number of unknowns in $\mathring{\mathbf{x}}$, namely, the dimension of the vector space V containing it. The beautiful insight of compressed sensing is that this statement is actually too pessimistic if the information content in $\mathring{\mathbf{x}}$ is less than d. The only exception to this rule that was known prior to the inception of the field of compressed sensing was when $\mathring{\mathbf{x}}$ was known to live in a lower dimensional linear subspace $W \subset V$ with $\dim(W) \leq d$. A highly oversimplified summary of the contribution of compressed sensing therefore says that the field extended the previous observation from single subspaces to unions of subspaces. This interpretation of the set of sparse vectors is therefore also known as the *union-of-subspaces* model. While sparsity is certainly firmly positioned at the forefront of CS research, the concept of low-complexity models encompasses many other interesting structures such as block- or group-sparsity, as well as low-rankness of matrices to name a few.

We will comment on such signal models in Sect. 3. As hinted at before, the recovery of these signal classes can be treated in a unified way using the atomic norm formalism (cf. Sect. 4) as long as we are only interested in nonuniform recovery results. Establishing similar results which hold uniformly over entire signal classes, however, usually requires more specialized analyses. In the later parts of this introduction, we therefore limit our discussions to sparse vectors. Note that while more restrictive low-complexity structures such as block- or group-sparsity overlap with the class of sparse vectors, the recovery guarantees obtained by merely modeling such signals as sparse are generally suboptimal as they do not exploit all latent structure inherent to their respective class.

Before moving on to a more detailed discussion of the most common signal models, we briefly want to comment on a particular line of research that deals with low-complexity signal recovery from nonlinear observations. Consider an arbitrary univariate, scalar-valued function f acting element-wise on vectors:

$$\mathbf{y} = f(\mathbf{A}\mathbf{x}). \tag{1}$$

An interesting instance of Eq. (1) is when f models the effects of an analog-to-digital converter (ADC), mapping the infinite-precision observations $\mathbf{A}\mathbf{x}$ on a finite quantization alphabet. Since this extension of the linear observation model gives rise to its very own set of problems which require specialized tools beyond what is needed in the basic theory of compressed sensing, we will not discuss this particular measurement paradigm in this introduction. A good introduction to the general topic of nonlinear signal recovery can be found in [100]. For a detailed survey focusing on the comparatively young field of quantized compressed sensing, we refer interested readers to [16].

2 Preliminaries

Compressed sensing builds on various mathematical tools from linear algebra, optimization theory, probability theory, and geometric functional analysis. In this section, we review some of the mathematical notions used throughout this chapter. We start with a few remarks regarding notation.

Notation

We use lower- and uppercase boldface letters to denote vectors and matrices, respectively. The all ones vector of appropriate dimension is denoted by $\mathbf{1}$, the zero vector is $\mathbf{0}$, and the identity matrix is Id. Given a natural number $n \in \mathbb{N}$, we denote by $[n]$ the set of integers from 1 to n, i.e., $[n] := \{1, \ldots, n\} = \mathbb{N} \cap [1, n]$. The complement of a subset $A \subset B$ is denoted by $\overline{A} = B \setminus A$. For a vector $\mathbf{x} \in \mathbb{C}^d$ and an index set $S \subset [d]$ with $|S| = k$, the meaning of $\mathbf{x}_S$ may change slightly depending on context. In particular, it might denote the vector $\mathbf{x}_S \in \mathbb{C}^d$ which agrees with $\mathbf{x}$ only on the index set S, and vanishes identically otherwise. On the other hand, it might represent the k-dimensional vector restricted to the coordinates indexed by S. The particular meaning should be apparent from context. Finally, for $a, b > 0$, the notation $a \lesssim b$ hides an absolute constant $C > 0$, which does not depend on either a or b, such that $a \leq Cb$ holds.

2.1 Norms and Quasinorms

The vectors we consider in this chapter are generally assumed to belong to a finite- or infinite-dimensional Hilbert space $\mathcal{H}$, i.e., a vector space endowed with a bilinear form $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ known as *inner product*, which induces a norm on the underlying vector space by[1]

$$\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

The d-dimensional Euclidean space $\mathbb{R}^d$ is an example of a vector space with the inner product between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i.$$

The norm induced by this inner product corresponds to the so-called ℓ_2-norm. In general, the family of ℓ_p-norms on $\mathbb{R}^d$ is defined as

$$\|\mathbf{x}\|_p := \begin{cases} \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, & p \in [1, \infty) \\ \max_{i \in [d]} |x_i|, & p = \infty. \end{cases}$$

Note that the ℓ_2-norm is the only ℓ_p-norm on $\mathbb{R}^d$ that is induced by an inner product since it satisfies the parallelogram identity. One can extend the definition of ℓ_p-norms to the case $p \in (0, 1)$. However, the resulting "ℓ_p-norm" ceases to be a norm as it no longer satisfies the triangle inequality. Instead, the collection of ℓ_p-norms for $p \in (0, 1)$ defines a family of quasinorms which satisfy the weaker condition

[1] Technically, a Hilbert space is an inner product space in which every Cauchy sequence converges to a point in the same space.

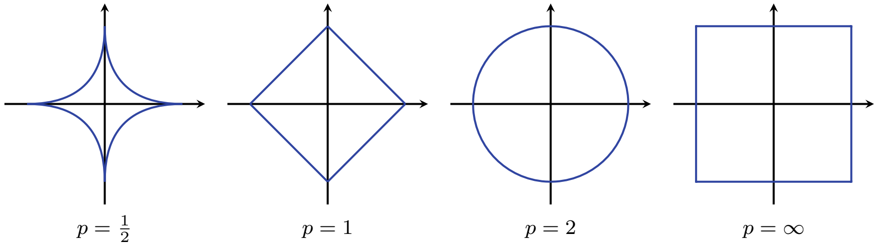


Fig. 1 The ℓ_p-unit spheres in $\mathbb{R}^2$ for different values of p. The interiors (including their respective boundaries) correspond to the ℓ_p-balls $\mathbb{B}_p^d$

$$\|\mathbf{x} + \mathbf{y}\|_p \leq 2^{1/p-1}(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p).$$

Additionally, we will make frequent use of the egregiously termed ℓ_0-norm of $\mathbf{x}$ which is defined as the number of nonzero coefficients,

$$\|\mathbf{x}\|_0 := \lim_{p \to 0} \|\mathbf{x}\|_p^p = |\text{supp}(\mathbf{x})| \, .$$

Note that the ℓ_0-norm, as a measure of *sparsity* of a vector, is neither a norm nor a quasinorm (or even a seminorm) as it is not positively homogeneous, i.e., for $t > 0$ we have $\|t\mathbf{x}\|_0 = \|\mathbf{x}\|_0 \neq t \, \|\mathbf{x}\|_0$. As we will see later, both the ℓ_1-norm, and the ℓ_p-quasinorms are of particular interest in the theory of compressed sensing. The ℓ_p-unit ball, defined as

$$\mathbb{B}_p^d := \left\{ \mathbf{x} \in \mathbb{C}^d : \|\mathbf{x}\|_p \leq 1 \right\},$$

forms a convex body for $p \geq 1$ and a nonconvex one for $p \in (0, 1)$. The boundaries $\partial\mathbb{B}_p^d = \{\mathbf{x} : \|\mathbf{x}\|_p = 1\}$ correspond to the ℓ_p-unit spheres. For $p = 2$, the boundary $\partial\mathbb{B}_2^d$ of the ℓ_2-ball corresponds to the unit Euclidean sphere denoted $\mathbb{S}^{d-1}$. Some examples of the ℓ_p-unit spheres are given in Fig. 1.

Another commonly used space in compressed sensing is the space of linear transformations from $\mathbb{R}^d$ to $\mathbb{R}^m$. This particular function space is isomorphic to the collection of $\mathbb{R}^{m \times d}$ matrices and forms a vector space on which we can define an inner product via

$$\langle \mathbf{A}, \mathbf{B} \rangle := \text{tr}(\mathbf{A}^\top \mathbf{B}).$$

The norm induced by this inner product is called the Frobenius norm and is given by

$$\|\mathbf{A}\|_\text{F} := \sqrt{\text{tr}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\sum_{i \in [d]} \sum_{j \in [m]} a_{ij}^2}.$$

In this context, the inner product above is also known as the so-called *Frobenius* inner product. Another commonly used norm defined on the space of linear transformations is the operator norm

$$\|\mathbf{A}\|_{p\to q} := \sup_{\|\mathbf{x}\|_p \leq 1} \|\mathbf{A}\mathbf{x}\|_q .$$

In particular, the operator norm $\|\mathbf{A}\|_{2\to 2}$ between two normed spaces equipped with their respective ℓ_2-norm is given by the maximum singular value of $\mathbf{A}$ denoted by $\sigma_{\max}(\mathbf{A})$.

2.2 Random Variables, Vectors, and Matrices

Let $(\boldsymbol{\Omega}, \Sigma, \mathbb{P})$ be a probability space consisting of the sample space $\boldsymbol{\Omega}$, the Borel measurable event space Σ, and a probability measure $\mathbb{P}\colon \Sigma \to [0, 1]$. The space of matrix-valued, Borel measurable functions from $\boldsymbol{\Omega}$ to $\mathbb{R}^{m\times d}$ are called *random matrices*. This space inherits a probability measure as the pushforward of the measure $\mathbb{P}$. For $d = 1$, we obtain the set of random vectors; the space of random variables corresponds to the choice $m = d = 1$. Given a scalar random variable X, the *expected value* of X is defined as

$$\mathbb{E}X := \int X \mathrm{d}\mathbb{P} = \int_{\boldsymbol{\Omega}} X(\omega)\mathrm{d}\mathbb{P}(\omega)$$

if the integral exists. Moreover, if $\mathbb{E}e^{tX}$ exists for all $|t| < h$ for some $h \in \mathbb{R}$, then the map

$$M_X\colon \mathbb{R} \to \mathbb{R}\colon t \mapsto M_X(t) = \mathbb{E}e^{tX} = \int e^{tX}\mathrm{d}\mathbb{P},$$

known as the moment generating function (MGF), fully determines the distribution of X. The pth absolute moment of a random variable X is defined as

$$\mathbb{E}|X|^p = \int_{\boldsymbol{\Omega}} |X(\omega)|^p \mathrm{d}\mathbb{P}(\omega).$$

This leads to the notion of the so-called L^p norm

$$\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p},$$

which turns the space of random variables equipped with $\|\cdot\|_{L^p}$ into a normed vector space. A particular class of random variables which finds widespread use in compressed sensing is the so-called subgaussian random variables whose L^p norm increases at most as $\sqrt{p}$. The name *subgaussian* is owed to the fact that subgaussian random variables have tail probabilities which decay at least as fast as the tails of the Gaussian distribution [99]. This leads to the following definition.

Definition 1 (*Subgaussian random variables*) A random variable X is called subgaussian if it satisfies one of the following equivalent properties:

1. The tails of X satisfy
$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2) \quad t \geq 0.$$
2. The absolute moments of X satisfy
$$(\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \forall p \geq 1.$$
3. The super-exponential moment of X satisfies
$$\mathbb{E}\exp(X^2/K_3^2) \leq 2.$$
4. If $\mathbb{E}X = 0$, then the MGF of S satisfies
$$\mathbb{E}\exp(tX) \leq \exp(K_4^2t^2) \quad \forall t \in \mathbb{R}.$$

The constants $K_1, \ldots, K_4$ are universal.

Note that the constants $K_i > 0$ for $i = 1, 2, 3, 4$ differ from each other by at most a constant factor, which, in turn, deviate only by a constant factor from the so-called subgaussian norm $\|\cdot\|_{\psi_2}$.

Definition 2 (*Subgaussian norm*) Given a random variable X, we define the subgaussian norm of X as
$$\|X\|_{\psi_2} := \inf\{s > 0 : \mathbb{E}\psi_2(X/s) \leq 1\},$$
where $\psi_2(t) := \exp(t^2) - 1$ is called an *Orlicz function*.

The set of subgaussian random variables defined on a common probability space equipped with the norm $\|\cdot\|_{\psi_2}$ therefore forms a normed space known as *Orlicz space*. Note that some authors instead define the subgaussian norm as
$$\|X\|_{\psi_2} := \sup_{p\geq 1} \frac{1}{\sqrt{p}}(\mathbb{E}|X|^p)^{1/p}. \tag{2}$$

In light of Definition 1, these definitions are equivalent up to a multiplicative constant. As a consequence of Eq. (2) and Definition 2 above, a random variable is subgaussian if its subgaussian norm is finite. For instance, the subgaussian norm of a Gaussian random variable $X \sim \mathsf{N}(0, \sigma^2)$ is—up to a constant—multiplicatively bounded from above by σ. The subgaussian norm of a Rademacher random variable is given by $\|X\|_{\psi_2} = 1/\sqrt{\log 2}$. Gaussian and Bernoulli random variables are therefore typical instances of subgaussian random variables. Other examples include random variables following the Steinhaus[2] distribution, as well as any bounded random variables in general.

A convenient property of subgaussian random variables is that their tail probabilities can be expressed in terms of their subgaussian norm:

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\frac{ct^2}{\|X\|_{\psi_2}^2}\right) \quad \forall t > 0.$$

If $X_i \sim \mathsf{N}(0, \sigma_i^2)$ are independent Gaussian random variables, then due to the rotation invariance of the normal distribution, the linear combination $X = \sum_i X_i$ is still a zero-mean Gaussian random variable with variance $\sum_i \sigma_i^2$. This property also extends to subgaussians barring a dependence a multiplicative constant, i.e., if $(X_i)_i$ is a sequence of centered subgaussian random variables, then

$$\|\sum_i X_i\|_{\psi_2}^2 \leq C \sum_i \|X_i\|_{\psi_2}^2.$$

This can easily be shown with the help of the moment generating function of $X = \sum_i X_i$. The rotational invariance along with the tail property of subgaussian distributions makes it possible to generalize many familiar tools such as Hoeffding-type inequalities to subgaussian distributions, e.g.,

$$\mathbb{P}\left(\left|\sum_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_i \|X_i\|_{\psi_2}^2}\right) \quad \forall t > 0.$$

Oftentimes, it is convenient to extend the notion of subgaussianity from random variables to random vectors. In particular, we say that a random vector $\mathbf{X} \in \mathbb{R}^m$ is subgaussian if the random variable $X = \langle \mathbf{X}, \mathbf{y} \rangle$ is subgaussian for all $\mathbf{y} \in \mathbb{R}^m$. Taking the supremum of the subgaussian norm of X over all unit directions then leads to the definition of the subgaussian norm for random vectors.

Definition 3 (*Subgaussian vector norm*) The subgaussian norm of an m-dimensional random vector $\mathbf{X}$ is

$$\|\mathbf{X}\|_{\psi_2} := \sup_{\mathbf{y} \in \mathbb{S}^{m-1}} \| \langle \mathbf{X}, \mathbf{y} \rangle \|_{\psi_2}.$$

Finally, a random vector $\mathbf{X}$ is called isotropic if $\mathbb{E}| \langle \mathbf{X}, \mathbf{y} \rangle |^2 = \|\mathbf{y}\|_2^2$ for all $\mathbf{y} \in \mathbb{R}^m$.

[2] A Steinhaus random variable is a complex random variable distributed uniformly on the complex unit circle.

3 Signal Models

As a basic framework for the types of signals discussed in this introduction, we decided to adopt the notion of so-called *atomic sets* as coined by Chandrasekaran et al. [31]. This serves two purposes. First, it elegantly emphasizes the notion of *low complexity* of the signals one aims to recover or estimate in practice. Second, the associated notion of *atomic norm* (cf. Definition 5) provides a convenient way to motivate certain geometric ideas in the recovery of low-complexity models. Let us emphasize that this viewpoint is not necessarily required when discussing so-called uniform recovery results where one is interested in conditions allowing for the recovery of entire signal classes given a fixed draw of a measurement matrix (cf. Sect. 7). However, the concept provides a suitable level of abstraction to discuss recovery conditions for individual vectors of a variety of different interesting signal models in a unified manner which were previously studied in isolation by researchers in their respective fields.

As alluded to in the motivation, one of the most common examples of a “low-complexity” structure of a signal $\mathring{\mathbf{x}} \in \mathbb{C}^d$ is the assumption that it belongs to a lower dimensional subspace of dimension k. Given a matrix $\mathbf{U} \in \mathbb{C}^{d \times k}$ whose columns $\mathbf{u}_i$ span said subspace, and the linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, we may simply solve the least-squares problem

$$\underset{\mathbf{c} \in \mathbb{C}^k}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{A}\mathbf{U}\mathbf{c}\|_2 \tag{3}$$

to recover $\mathring{\mathbf{x}} = \mathbf{U}\mathbf{c}^\star$ where the solution $\mathbf{c}^\star$ of Problem (3) admits a closed-form expression in terms of the Moore–Penrose pseudoinverse of $\mathbf{AU}$. Once again, this strategy succeeds if $m \geq \dim \operatorname{span}(\{\mathbf{u}_i\}_{i=1}^k)$, i.e., if we obtain at least as many measurements as the subspace dimension. As a canonical example, assume that $\mathbf{U}$ corresponds to the identity matrix Id restricted to the columns indexed by a set $S \subset [d]$ of cardinality $|S| = k$, i.e., $\mathbf{U} = \mathrm{Id}_S$. The columns of this matrix form a basis for a k-dimensional coordinate subspace of $\mathbb{C}^d$. If we lift the restriction that $\mathring{\mathbf{x}}$ lives in this particular subspace, and rather assume instead that $\mathring{\mathbf{x}}$ belongs to any of the $\binom{d}{k}$ coordinate subspaces of dimension k, we arrive at a special case of the so-called *union-of-subspaces* model. In particular, we have

$$\mathring{\mathbf{x}} \in \bigcup_{\substack{S \subset [d], \\ |S| = k}} W_S =: \Sigma_k,$$

where W_S denotes the coordinate subspace of $\mathbb{C}^d$ with basis matrix Id_S. The set Σ_k therefore corresponds to the set of sparse vectors supported on an index set S of cardinality at most k. This signal class represents a central object of study in the field of compressed sensing.

Equipped with the knowledge that $\mathring{\mathbf{x}}$ lives in one of the k-dimensional coordinate subspaces, one could attempt to recover $\mathring{\mathbf{x}}$ by solving Problem (3) for each W_S

independently. However, even though the true solution $\mathring{\mathbf{x}}$ must be among these least-squares solutions, there is no way for us to identify the correct one. Moreover, even for moderately sized problems, the number $\binom{d}{k}$ of least-squares projections one needs to solve becomes unreasonably high. On the other hand, ignoring the information that $\mathring{\mathbf{x}}$ lives in k-dimensional subspace, and instead solving the least-squares minimization problem

$$\underset{\mathbf{x}\in\mathbb{C}^d}{\text{minimize}} \quad \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2$$

will not help either since the ℓ_2-norm we are minimizing tends to spread the signal energy over the entire support of the minimizer $\mathbf{x}^\star$ (see, e.g., the discussion in [18, Sect. 6.1.2]). We will discuss in Sect. 5 that all these issues can be resolved by imposing certain structural constraints on the measurement matrix $\mathbf{A}$, and replacing the optimization problem (3) with one that explicitly promotes the structure inherent in $\mathring{\mathbf{x}}$.

We will come back to the sparse signal model shortly. First, however, let us introduce a more flexible notion of low-complexity structures which will allow us to talk about recovery problems of more general signal models in a unified framework. As outlined above, if $\mathcal{K}$ denotes a k-dimensional subspace, then every vector in $\mathcal{K}$ can be represented as a sum of k basis vectors. To capture a similar notion of dimensionality for more general sets which do not necessarily form a subspace, we may assume that every vector in $\mathcal{K}$ can at least be represented as a linear combination of a limited number of elements in a more general generating set. While a finite-dimensional subspace is always fully determined by a finite collection of basis vectors, we now lift this finiteness requirement. The signal models generated in this fashion are simply referred to as *simple sets*.

Definition 4 (*Simple set*) Let $\mathcal{A} \subset \mathbb{C}^d$ be an origin-symmetric set whose convex hull forms a convex body.[3] Let $k \in \mathbb{N}$. Then the set

$$\mathcal{K} = \operatorname{cone}_k(\mathcal{A}) := \left\{ \mathbf{x} = \sum_{i=1}^{k} c_i \mathbf{a}_i \in \mathbb{C}^d : c_i \geq 0, \mathbf{a}_i \in \mathcal{A} \right\} \tag{4}$$

is called a *simple set*. Since $\mathcal{K}$ is generated by the set $\mathcal{A}$, we call $\mathcal{A}$ an *atomic set*.

We will discuss how this notion of simplicity leads to many familiar models in the literature on linear inverse problems. As a canonical example, however, consider the case $\mathcal{A} = \{\pm\mathbf{e}_i\} \subset \mathbb{R}^d$. The simple set $\mathcal{K}$ generated by $\operatorname{cone}_k(\mathcal{A})$ then corresponds to the set $\Sigma_k(\mathbb{R}^d)$ of k-sparse vectors.

Given an atomic set $\mathcal{A}$, we associate with it the following object.

Definition 5 (*Atomic norm*) The function

[3] A convex body is a compact convex set with non-empty interior.

$$\|\mathbf{x}\|_{\mathcal{A}} := \inf \left\{ \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} : \mathbf{x} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a}, \, c_{\mathbf{a}} \geq 0 \; \forall \mathbf{a} \in \mathcal{A} \right\}$$

associated with an atomic set $\mathcal{A} \subset \mathbb{C}^d$ is called the *atomic norm* of $\mathcal{A}$ at $\mathbf{x}$.

This definition corresponds to the so-called *Minkowski functional* or *gauge* of the set $\operatorname{conv}(\mathcal{A})$ [88, Chap. 15],

$$\gamma_{\operatorname{conv}(\mathcal{A})}(\mathbf{x}) := \inf \{ t > 0 : \mathbf{x} \in t \operatorname{conv}(\mathcal{A}) \} = \|\mathbf{x}\|_{\mathcal{A}} .$$

The norm notation $\|\cdot\|_{\mathcal{A}}$ is justified here since we assumed $\mathcal{A}$ to be compact and centrally symmetric with $\operatorname{conv}(\mathcal{A})$ having non-empty interior. This ensures that $\operatorname{conv}(\mathcal{A})$ is a symmetric convex body which contains an open set around the origin in which case $\|\cdot\|_{\mathcal{A}} = \gamma_{\operatorname{conv}(\mathcal{A})}(\cdot)$ defines a norm on $\mathbb{C}^d$. With this definition in place, the general strategy to recover a simple vector $\mathring{\mathbf{x}} \in \mathcal{K} = \operatorname{cone}_k(\mathcal{A})$ from its linear measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ is

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_{\mathcal{A}} \\ \text{s.t.} \quad & \mathbf{y} = \mathbf{A}\mathbf{x}. \end{aligned} \tag{P$_{\mathcal{A}}$}$$

We will discuss in Sect. 4 why Problem (P$_{\mathcal{A}}$), which we will simply refer to as *atomic norm minimization*, allows for the recovery of simple sets from underdetermined linear measurements.

In the remainder of this section, we will introduce some of the most common low-complexity sets discussed in the literature. We limit our discussion to sparse vectors, block- and group-sparse vectors, as well as low-rank matrices. Note, however, that the atomic norm framework allows for modeling many other interesting signal classes beyond the ones discussed here. These include permutation and cut matrices, eigenvalue-constrained matrices, low-rank tensors, and binary vectors. We specifically refer interested readers to [31, Sect. 2.2] for a more comprehensive list of example applications of atomic sets.

3.1 Sparse Vectors

As we highlighted various times at this point, the most widespread notion of low complexity at the heart of CS is the notion of sparsity. Even before the advent of compressed sensing, exploiting low complexities in signals played a key role in the development of most compression technologies such as MP3, JPEG, or H264. Ultimately, all these technologies are based on the idea that most signals of interest usually live in rather low-dimensional subspaces embedded in high-dimensional vector spaces.[4] Two canonical examples of this phenomenon are the superposition

[4] This idea also extends to signals living on low-dimensional manifolds.

of sine waves and natural images. In the former case, it is obvious that we are only able to infer very little information from glancing at a time series plot of a sound wave recorded at a microphone. For instance, we might be able to say when a signal is made up of mostly low-frequency components if its waveform only appears to change very slowly over time, but for most signals we are usually not able to say much beyond that. The situation changes drastically, however, if we instead inspect the signal's Fourier transform. In the example of superimposed sine waves, the inherent simplicity or low complexity of the signal becomes immediately apparent in the form of a few isolated peaks in the Fourier spectrum of the signal, revealing the true low-complexity structure of the signal. A similar observation can be made for natural images where periodic structures—say a picture of a garden fence or a brick wall—or flat, homogeneous textures—say in images featuring a view of the sky or blank walls—lead to sparse representations in a variety of bases such as the discrete Fourier transform (DFT) basis, the discrete cosine transform (DCT) basis or the extended family of x-let systems, e.g., wavelets [68], curvelets [22], noiselets [34], shearlets [65], and so on.

Formally, the set of sparse vectors is simply defined as the set of vectors in $\mathbb{C}^d$ with at most k nonzero coefficients. For convenience, this is mostly defined mathematically with the help of the ℓ_0-pseudonorm

$$\|\mathbf{x}\|_0 := |\operatorname{supp}(\mathbf{x})| = |\{i \in [d] : x_i \neq 0\}| \,.$$

With this definition, the set of all k-sparse vectors can be written as

$$\Sigma_k = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq k\}.$$

As we discussed in the beginning of Sect. 3, the set Σ_k is a collection of $\binom{d}{k}$ k-dimensional subspaces, each one spanned by k canonical basis vectors. Since it is a union and not a sum of subspaces, the set is highly nonlinear in nature, e.g., the sum of two k-sparse vectors is generally $2k$-sparse in case the vectors are supported on disjoint support sets.

Consider again the linear inverse problem in which we are tasked with inferring $\mathring{\mathbf{x}} \in \Sigma_k$ from its measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$. As we motivated before, if the support of the k-sparse vector is known, so is the corresponding subspace, and the signal can be easily recovered via a least-squares projection. If on the other hand we assume that the support is not known, the situation becomes dire as we now have to consider intractably many possible subspaces. To get a feeling for the complexity of the set of sparse vectors, consider for some $c \in \mathbb{R}$ the set $\left\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_0 = k, x_i = c \;\forall i \in \operatorname{supp}(\mathbf{x})\right\} \subset \Sigma_k$, i.e., the set of exactly k-sparse vectors with identical nonzero entries. A random vector uniformly drawn from this set has entropy $\log\binom{d}{k}$, which means that[5] $\log\binom{d}{k} \approx k \log(d/k)$ bits are required for effective compression of this set [90]. As we will see in Sects. 4 and 7, the expression $k\log(d/k)$ plays a key role in the theory of compressed sensing.

[5] This follows from the classical bound $\left(\frac{d}{k}\right)^k \leq \binom{d}{k} \leq \left(\frac{ed}{k}\right)^k$.

To frame the set of sparse vectors in the language of simple sets as established in the beginning of Sect. 3, we note that the atomic set corresponding to the set of sparse vectors in $\mathbb{R}^d$ is simply the set of signed unit vectors, i.e., $\mathcal{A} = \{\pm \mathbf{e}_i\}$.[6] Since the convex hull of $\mathcal{A}$ clearly corresponds to the ℓ_1-unit ball, we have $\Sigma_k(\mathbb{R}^d) = \mathrm{cone}_k(\mathcal{A})$. The atomic norm associated with this set is simply the ℓ_1-norm on $\mathbb{R}^d$. This easily follows from expanding a vector in terms of the elements of $\mathcal{A}$ as

$$\mathbf{x} = \sum_{i=1}^{d} |x_i| \underbrace{\mathrm{sign}(x_i)\mathbf{e}_i}_{\in \mathcal{A}} .$$

Then we have with Definition 5 that

$$\begin{aligned} \|\mathbf{x}\|_{\mathcal{A}} &= \inf \left\{ \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}} : \sum_{\mathbf{a}\in\mathcal{A}} c_{\mathbf{a}}\mathbf{a}, c_{\mathbf{a}} \geq 0 \right\} \\ &= \sum_{i=1}^{d} |x_i| = \|\mathbf{x}\|_1 . \end{aligned}$$

While there are infinitely many ways to express each coordinate x_i in terms of nonnegative linear combinations of the atoms $\mathbf{e}_i$ and $-\mathbf{e}_i$, the infimum in the definition of $\|\cdot\|_{\mathcal{A}}$ is attained when each coordinate is expressed by exactly one element of $\mathcal{A}$. This follows immediately from the triangle inequality.

Compressible Vectors

While the concept of sparsity arises naturally in an abundance of contexts and applications, in many cases it is also a slightly too stringent model for practical purposes. A canonical example is natural images which certainly exhibit a low-complexity structure if expressed in a suitable sparsity basis. However, this basis expansion is usually not perfect. In other words, by close inspection one usually notices that while the majority of the signal energy concentrates in only a limited number of expansion coefficients, there usually also exist many coefficients with non-negligible amplitudes which carry information about fine structures of images. Nevertheless, a histogram of the transform coefficients usually reveals that the negligible coefficients quickly decay such that natural images are still be well approximated by sparse vectors. This concept, which leads us to the class of so-called *compressible vectors*, is also heavily exploited in image compression algorithms which quantize infrequently occurring transform coefficients more aggressively (i.e., more coarsely) than more dominant ones such as DC coefficients.

Formally, let $\mathbf{x} \in \mathbb{C}^d$ be a vector whose k largest components in absolute value are supported on a set $S \subset [d]$ of size k, and define for $p > 0$ the *best k-term approximation error* $\sigma_k(\cdot)_p \colon \mathbb{C}^d \to \mathbb{R}_{\geq 0}$ as

[6] To define the sparse vectors on $\mathbb{C}^d$, simply replace $\{\pm \mathbf{e}_n\}$ by $\{\pm \mathbf{e}_n, \pm i\mathbf{e}_n\}$.

$$\sigma_k(\mathbf{x})_p := \min_{\mathbf{z}\in\Sigma_k} \|\mathbf{x}-\mathbf{z}\|_p. \tag{5}$$

For any $p > 0$, the minimum in Eq. (5) is attained by the vector $\mathbf{z}$ which agrees with $\mathbf{x}$ on S and vanishes identically on $\overline{S}$. The following result characterizes the decay behavior of the approximation error.

Theorem 1 ([54, Theorem 2.5]) *Let $q > p > 0$. Then for any $\mathbf{x} \in \mathbb{C}^d$, the best k-term approximation error w. r. t. the ℓ_q-norm is bounded by*

$$\sigma_k(\mathbf{x})_q \leq \frac{c_{p,q}}{k^{1/p-1/q}} \|\mathbf{x}\|_p \tag{6}$$

with

$$c_{p,q} := \exp\left(-\frac{h_b(p/q)}{p}\right) \leq 1,$$

and $h_b(x) := -x\log(x) - (1-x)\log(1-x)$ denoting the binary entropy function. In particular, we have

$$\sigma_k(\mathbf{x})_2 \leq \frac{1}{2\sqrt{k}} \|\mathbf{x}\|_1 .$$

The set of vectors which can be well approximated in terms of σ_k are called *compressible vectors*. Informally, this means that a vector $\mathbf{x}$ is compressible if $\sigma_k(\mathbf{x})_p$ decays quickly as k increases. One particular set of vectors which exhibit such a rapid error decay is the elements of the ℓ_q-quasinorm balls

$$\mathbb{B}_q^d = \left\{\mathbf{z} \in \mathbb{C}^d : \|\mathbf{z}\|_q \leq 1\right\}$$

with $0 < q \leq 1$. To see why the ℓ_q-quasinorm balls are suitable proxies for sparse vectors, consider the limiting behavior of the quasinorm. For $q \to 0$ we have

$$\begin{aligned}\lim_{q\to 0} \|\mathbf{x}\|_q^q &= \lim_{q\to 0} \sum_{i=1}^d |x_i|^q \\ &= \sum_{i=1}^d \mathbb{1}_{\{x_i \neq 0\}} \\ &= |\{i \in [d] : x_i \neq 0\}| \\ &= \|\mathbf{x}\|_0 .\end{aligned}$$

In the other limiting case, one obtains the set of unit ℓ_1-norm vectors. Moreover, applying Theorem 1 to the case of ℓ_q-norm balls, we find

$$\sigma_k(\mathbf{x})_2 \leq \frac{c_{q,2}}{k^{\frac{1}{q}-\frac{1}{2}}}.$$

Finally, it can be shown that the ith biggest entry of $\mathbf{x}$ decays as $i^{-1/q}$ [37].

3.2 Block- and Group-Sparse Vectors

While the model of sparse and compressible vectors has many interesting and justified applications, many times real-world signals will exhibit even more structure beyond simple sparsity. One of the most common generalizations of sparse vectors is so-called *block-sparse* or more generally *group-sparse* signals. In the former case, we assume that the set $[d]$ is partitioned into L disjoint subsets $B_l \subset [d]$ of possibly different sizes $|B_l| = b_l$ such that $\bigcup_{l=1}^{L} B_l = [d]$, and $\sum_{l=1}^{L} b_l = d$. If the sets B_l are allowed to overlap, we refer to them as *groups* instead. As in the case of sparse vectors, a vector $\mathbf{x} \in \mathbb{C}^d$ is called k-block-sparse or k-group-sparse if its nonzero coefficients are limited to at most k nonzero blocks or groups, respectively. Another closely related cousin of block-sparsity is that of fusion frame sparsity. Assuming equisized blocks B_l with $b_l = b$, one additionally imposes in this model that each subvector $\mathbf{x}_{B_l} \in \mathbb{C}^b$ belongs to some s-dimensional subspace $W_l \subset \mathbb{C}^b$ (see, e.g., [5, 15], for details). Structured sparsity models as outlined above arise in a variety of domains in engineering and biology. Some prominent example applications are audio [1] and image signal processing [102], multi-band reconstruction and spectrum sensing [70, 81], as well as sparse subspace clustering [48]. Further applications in which block- and group-sparse signal structures commonly appear are in the context of measuring gene expression levels [78] and protein mass spectroscopy [93]. For a more thorough treatment of block-sparse signal modeling, we also refer readers to [47, Chap. 2].

In the following, we limit our discussion to the case of block-sparsity. A natural way to express the block-sparsity of a vector mathematically is by introducing for $p, q > 0$ the family of mixed (ℓ_p, ℓ_q)-(quasi)norms

$$\|\mathbf{x}\|_{p,q} := \left(\sum_{l=1}^{L} \left\|\mathbf{x}_{B_l}\right\|_p^q\right)^{1/q},$$

where we denote by $\mathbf{x}_{B_l} \in \mathbb{C}^d$ the subvector of $\mathbf{x}$ restricted to the index set B_l. Extending the notation to include the case $q = 0$, we define additionally the mixed (ℓ_p, ℓ_0)-pseudonorm

$$\begin{aligned}\|\mathbf{x}\|_{p,0} &:= \left|\left\{\left\|\mathbf{x}_{B_l}\right\|_p \neq 0 : l \in [L]\right\}\right| \\ &= \left|\left\{\mathbf{x}_{B_l} \neq \mathbf{0} : l \in [L]\right\}\right|,\end{aligned}$$

which simply counts the number of nonzero blocks of $\mathbf{x}$ w.r.t. $\{B_l\}_{l=1}^L$. With this definition, a vector is called k-block-sparse if $\|\mathbf{x}\|_{p,0} \leq k$. Moreover, the atomic set which gives rise to the set of k-block-sparse vectors can now be defined as

$$\mathcal{A}_p := \bigcup_{l=1}^{L} \left\{ \mathbf{a} \in \mathbb{C}^d : \left\|\mathbf{a}_{B_l}\right\|_p = 1, \mathbf{a}_{\overline{B_l}} = \mathbf{0} \right\}. \tag{7}$$

Note that unlike in the case of sparse vectors where we defined $\tilde{\mathcal{A}} = \{\pm \mathbf{e}_i\}$, the set in Eq. (7) is uncountable. To calculate the atomic norm, recall the definition

$$\|\mathbf{x}\|_{\mathcal{A}_p} = \inf \left\{ \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} : \mathbf{x} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a}, \, c_{\mathbf{a}} \geq 0 \right\}.$$

Since $\operatorname{span}(\mathcal{A}_p) = \mathbb{C}^d$, there exists a $c_{\mathbf{a}} \geq 0$ and $\mathbf{a} \in \mathcal{A}_p$ such that for every $\mathbf{x} \in \mathbb{C}^d$, we may express its coefficients in block B_l as $\mathbf{x}_{B_l} = c_{\mathbf{a}}\mathbf{a}$. Then we have $\left\|\mathbf{x}_{B_l}\right\|_p = \|c_{\mathbf{a}}\mathbf{a}\|_p = |c_{\mathbf{a}}| \cdot \|\mathbf{a}\|_p = c_{\mathbf{a}}$ where the last step simply follows from the fact that $c_{\mathbf{a}} \geq 0$ and $\mathbf{a} \in \mathcal{A}_p$. Again, we have by the triangle inequality that the infimum in the definition of the atomic norm must be attained by a decomposition where each block B_l is represented by exactly one atom. Hence

$$\|\mathbf{x}\|_{\mathcal{A}_p} = \sum_{l=1}^{L} \left\|\mathbf{x}_{B_l}\right\|_p = \|\mathbf{x}\|_{p,1}.$$

Note that a similar argument holds for the group-sparsity case where the sets B_l are not assumed to be disjoint [84, Lemma 2.1].

Clearly, the atomic norm induced by $\mathcal{A}$ is closely related to the ℓ_1-norm as discussed in the previous section. In the edge case with $L = d$, and $|B_l| = 1$, we have $\mathcal{A}_p = \{\pm \mathbf{e}_i\}$ such that we immediately arrive again at the set of sparse vectors.

3.3 Low-Rank Matrices

A slightly different linear inverse problem which can still be conveniently modeled by means of atomic sets is the so-called low-rank matrix recovery problem. Consider a matrix $\mathbf{X} \in \mathbb{C}^{d_1 \times d_2}$ of rank at most r which we observe through the linear operator

$$\mathcal{M} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^m : \mathbf{X} \mapsto \mathcal{M}(\mathbf{X}) = \mathbf{y}.$$

As usual, our task is to infer $\mathbf{X}$ from knowledge of the map $\mathcal{M}$ and the measurements $\mathbf{y}$ by solving the atomic norm minimization problem ($\mathrm{P}_{\mathcal{A}}$). In general, there are of course $d_1 d_2$ unknown entries in $\mathbf{X}$ so that the linear inverse problem is clearly ill-posed as long as $m < d_1 d_2$. However, by exploiting a potential low-rank structure on $\mathbf{X}$, it turns out to be possible to drastically reduce the number of observations needed to allow for faithful estimation of low-rank matrices (cf. Table 1).

A typical example application of low-rank matrix recovery, known as the *matrix completion* problem, is the task of estimating missing entries of a matrix based on partial observations of $\mathbf{X}$ of the form $\mathcal{M}(\mathbf{X})_i = X_{kl}$ for some $(k, l) \in [d_1] \times [d_2]$. As before, this problem is clearly hopelessly ill-posed if $\mathbf{X}$ is a full-rank or close to full-rank matrix. However, in many practical situations in the context of collaborative filtering [56], the low-rank assumption on $\mathbf{X}$ is justified by the problem domain, making low-rank matrix recovery a useful prediction tool. The matrix completion problem was famously popularized by the so-called *Netflix Prize* [11], an open competition in collaborative filtering to predict user ratings of movies based on partial knowledge of ratings about other titles in the portfolio. The underlying assumption is that if two users both share the same opinion about certain titles they saw, then they are likely to share the same opinion about titles so far only seen or rated by one of them. In other words, if we collect the user ratings of all available titles in a database in a matrix $\mathbf{X}$, then we can assume that due to overlapping interests and opinions, the matrix will exhibit a low-rank structure. This reduction in the degrees of freedom therefore allows to accurately predict unknown user ratings which can then be used to provide personalized recommendations on a per-user basis.

To demonstrate how low-rank matrices can be modeled in the context of atomic sets, consider the set of rank-1 matrices of the form

$$\begin{aligned}\mathcal{A} &= \left\{\mathbf{u}\mathbf{v}^* \in \mathbb{C}^{d_1 \times d_2} : \|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1\right\} \\ &= \left\{\mathbf{u}\mathbf{v}^* \in \mathbb{C}^{d_1 \times d_2} : \left\|\mathbf{u}\mathbf{v}^*\right\|_{\mathrm{F}} = 1\right\}.\end{aligned}$$

Clearly, a nonnegative linear combination of r elements of $\mathcal{A}$ forms a matrix of at most rank r so that $\operatorname{cone}_r(\mathcal{A})$ generates the set of rank r matrices. To derive the atomic norm associated with $\mathcal{A}$, consider that for every $\mathbf{X} \in \mathbb{C}^{d_1 \times d_2}$ we have by the singular value decomposition of $\mathbf{X}$ that

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*,$$

Table 1 Mean width estimates for tangent cones

Signal set	Induced norm	Upper bound on $w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1})^2$
Sparse vectors in $\mathbb{R}^d$	$\|\cdot\|_1$	$2k \log(d/k) + 3k/2$
Block-sparse vectors in $\mathbb{R}^d$ with L blocks of size d/L	$\|\cdot\|_{2,1}$	$4k \log(L/k) + (1 + 6d/L)k/2$
Rank r matrices in $\mathbb{R}^{d_1 \times d_2}$	$\|\cdot\|_*$	$3r(d_1 + d_2 - r)$

where $\mathbf{U} \in \mathbb{C}^{d_1 \times d_1}$ and $\mathbf{V} \in \mathbb{C}^{d_2 \times d_2}$ are unitary matrices, and $\Sigma \in \mathbb{C}^{d_1 \times d_2}$ is a matrix containing the real-valued, nonnegative singular values on its main diagonal and zeros otherwise. Hence, we have with $d := \min\{d_1, d_2\}$,

$$\mathbf{X} = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$$

with $\mathbf{u}_i \mathbf{v}_i^* \in \mathcal{A}$. Again, with Definition 5 this yields

$$\begin{aligned} \|\mathbf{X}\|_{\mathcal{A}} &= \inf \left\{ \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} : \mathbf{X} = \sum_{\mathbf{a} \in \mathcal{A}} c_{\mathbf{a}} \mathbf{a},\, c_{\mathbf{a}} \geq 0 \right\} \\ &= \sum_{i=1}^{d} \sigma_i(\mathbf{X}) =: \|\mathbf{X}\|_*, \end{aligned}$$

where in the second step we simply identified $c_{\mathbf{a}}$ with the singular values of the decomposition after using the fact that by the triangle inequality (w. r. t. the Frobenius norm), the infimum must be attained by a decomposition of at most d atoms. While the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ which make up the atoms $\mathbf{a} = \mathbf{u}_i \mathbf{v}_i^* \in \mathcal{A}$ are not necessarily unique, each $\mathbf{X}$ is identified by a unique set of singular values.

The norm $\|\cdot\|_*$ is generally known as the nuclear norm and acts as an analog of the ℓ_1-norm in the case of sparse vectors since $\|\mathbf{X}\|_*$ corresponds to the ℓ_1-norm of the vector of singular values of $\mathbf{X}$. Considering that efficient algorithms for the singular value decomposition exist, the atomic norm minimization for low-rank matrices constitutes a tractable convex optimization problem.

Representability of Atomic Norms

While the examples of atomic sets we presented so far all admitted relatively straightforward representations of their associated atomic norms, efficient computation of $\|\cdot\|_{\mathcal{A}}$ for arbitrary atomic sets $\mathcal{A}$ is by no means guaranteed. A classic example of where the atomic norm framework fails to yield an efficient way to recover elements of a simple set generated by $\mathrm{cone}_k(\mathcal{A})$ is the set

$$\mathcal{A} = \left\{ \mathbf{z}\mathbf{z}^\top : \mathbf{z} \in \{\pm 1\}^d \right\}.$$

Similar to the set of low-rank matrices, the simple set generated by $\mathcal{A}$ consists of low-rank matrices but with its elements restricted to the set ± 1—a model which appears, for instance, in the context of collaborative filtering [73]. Considering that $\mathrm{conv}(\mathcal{A})$ corresponds to the so-called *cut polytope* which does not admit a tractable characterization, there exists no efficient way of computing $\|\cdot\|_{\mathcal{A}}$. In this case, one may turn to a particular approximation scheme of $\mathrm{conv}(\mathcal{A})$ known as *theta bodies* [58] which are closely related to the theory of sum-of-squares (SOS) polynomials. We refer interested readers to [31, Sect. 4].

As another example, consider the atomic set

$$\mathcal{A} = \left\{\mathbf{a}_{f,\phi} \in \mathbb{C}^d : f \in [0, 1], \phi \in [0, 2\pi)\right\}$$

with

$$\mathbf{a}_{f,\phi} := e^{i2\pi\phi} \begin{pmatrix} 1 \\ e^{i2\pi f} \\ \vdots \\ e^{i2\pi f(d-1)} \end{pmatrix}.$$

This set represents a continuous alphabet of atoms which gives rise to the signal set of sampled representations of continuous-time superpositions of complex exponentials [13]. Using results from the theory of SOS polynomials, Bhaskar et al. showed in [13] that the associated atomic norm can be computed as the solution of the program

$$\begin{array}{ll} \underset{\mathbf{x},\mathbf{u},t}{\text{minimize}} & \dfrac{\operatorname{tr} T(\mathbf{u})}{2d} + \dfrac{t}{2} \\ \text{s.t.} & \begin{pmatrix} T(\mathbf{u}) & \mathbf{x} \\ \mathbf{x}^* & t \end{pmatrix} \geq 0 \end{array}$$

where the linear operator $T : \mathbb{C}^d \to \mathbb{C}^{d \times d}$ maps a vector $\mathbf{u}$ to the Toeplitz matrix generated by $\mathbf{u}$. The same representation also appears in the context of *compressed sensing off the grid* where one aims to recover a sampled representation of a superposition of complex exponentials from randomly observed time-domain samples [92].

Both of these examples illustrate that while the atomic norm framework represents a convenient modeling tool for low-complexity signal sets, it may turn out to be a nontrivial or in some cases simply impossible task to actually find efficient ways to compute the atomic norm.

3.4 Low-Complexity Models in Bases and Frames

Up until this point, we have assumed that signals of interest are elements of a simple set $\mathcal{K} = \text{cone}_k(\mathcal{A})$ generated by an atomic set $\mathcal{A}$. Given a vector $\mathring{\mathbf{x}} \in \mathcal{K}$ and its linear measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$, the general task is to infer $\mathring{\mathbf{x}}$ from knowledge of $\mathbf{A}$ and $\mathbf{y}$. In this context, the measurement process is entirely modeled by $\mathbf{A}$. However, oftentimes in practical scenarios, we might not have direct access to the signal exhibiting a low-complexity structure but rather only to its representation in a particular *orthonormal basis* or more generally an *overcomplete dictionary* or *frame*. As a classical example, consider the situation in which $\mathring{\mathbf{x}} \in \mathbb{C}^d$ represents the sampled time-domain representation of a band-limited function. If the continuous-time signal is a superposition of k complex exponentials, the sampled representation $\mathring{\mathbf{x}}$

will generally have dense support. The underlying sparsity structure[7] only reveals itself to us after transforming $\mathring{\mathbf{x}}$ into the frequency domain, i.e., $\mathring{\mathbf{z}} = \mathbf{F}_d\mathring{\mathbf{x}} \in \Sigma_k$ with $\mathbf{F}_d = d^{-1/2}(e^{-i2\pi\mu\nu})_{0\leq\mu,\nu\leq d-1}$ denoting the DFT matrix. We therefore acquire measurements according to $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{F}_d^*\mathring{\mathbf{z}} =: \tilde{\mathbf{A}}\mathring{\mathbf{z}}$. Reconstruction of $\mathring{\mathbf{x}}$ now proceeds in two steps by first reconstructing the vector $\mathring{\mathbf{z}}$, exploiting its underlying low-complexity structure, and then resynthesizing the estimate of $\mathring{\mathbf{x}}$. For this reason, this model is also known as *synthesis model* throughout the literature. In general, one may assume that rather than exhibiting a low-complexity structure in the canonical basis, applications typically either fix or learn a suitable basis change matrix. Moreover, allowing for the transform matrix to be an overcomplete dictionary or frame $\boldsymbol{\Omega} \in \mathbb{C}^{d\times D}$ with $D > d$ such that $\mathring{\mathbf{x}} = \boldsymbol{\Omega}\mathring{\mathbf{z}}$ where $\mathring{\mathbf{z}} \in \mathbb{C}^D$ exhibits a low-complexity structure, one may exploit additional advantages stemming from the redundancy of overcomplete representations [30]. Classical examples of such representation systems are curvelet transforms [23] and time–frequency atoms arising from the Gabor transform [49]. For simplicity of presentation, we will assume in the remainder of this chapter that signals of interest already live in simple sets, i.e., we set $\boldsymbol{\Omega} = \mathbf{I}_d$, and point out that most results presented in the sequel also generalize to low-complexity models in unitary bases and frames. For more details, we refer interested readers to [86].

4 Recovery of Individual Vectors

In this section, we address the recovery of individual signals in simple sets $\mathcal{K}$ generated by $\mathrm{cone}_k(\mathcal{A})$. For simplicity, we limit our discussion to the case where the atomic set $\mathcal{A}$ contains only real elements so that $\mathcal{K} \subset \mathbb{R}^d$.

4.1 Exact Recovery

We begin our discussion by motivating why atomic norm minimization as stated in Problem ($\mathrm{P}_{\mathcal{A}}$) is a suitable strategy for the recovery of simple signals from linear measurements. To that end, consider again the equality-constrained minimization problem

$$\begin{aligned} &\text{minimize} \;\; \|\mathbf{x}\|_{\mathcal{A}} \\ &\text{s.t.} \qquad \mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{x}. \end{aligned} \tag{8}$$

By rewriting the equality constraint in terms of $\mathbf{d} = \mathring{\mathbf{x}} - \mathbf{x} \in \ker(\mathbf{A})$, we may restate the problem as

[7] We assume that the fundamental frequencies of each complex exponential are integer multiples of the frequency resolution f_s/d where f_s denotes the sampling rate.

$$\underset{\mathbf{d}\in\ker(\mathbf{A})}{\text{minimize}} \quad \left\|\mathbf{d}+\mathring{\mathbf{x}}\right\|_{\mathcal{A}}.$$

Of course, the above problem is not of any practical interest as it requires knowledge of the true solution $\mathring{\mathbf{x}}$. However, it immediately follows from this representation that Problem (8) has a unique solution if the null space of $\mathbf{A}$ does not contain any nontrivial directions which reduce the atomic norm anchored at $\mathring{\mathbf{x}}$. More precisely, by introducing the set

$$\mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}}) := \left\{\mathbf{d}\in\mathbb{R}^d : \left\|\mathbf{d}+\mathring{\mathbf{x}}\right\|_{\mathcal{A}} \le \left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}}\right\} = \left\{\mathbf{z}-\mathring{\mathbf{x}} : \|\mathbf{z}\|_{\mathcal{A}} \le \left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}}\right\}$$

of *descent directions* of $\|\cdot\|_{\mathcal{A}}$ at $\mathring{\mathbf{x}}$, we obtain the condition

$$\mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \ker(\mathbf{A}) = \{\mathbf{0}\}, \tag{9}$$

which, if satisfied, guarantees perfect recovery of $\mathring{\mathbf{x}}$ via Problem (8).

Alternatively, one may argue as follows. Let $\mathring{\mathbf{x}} \in \mathrm{cone}_k(\mathcal{A})$ and define the set $\mathcal{X} = \left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}} \mathrm{conv}(\mathcal{A})$ which clearly contains $\mathring{\mathbf{x}}$. Given access to linear measurements of the form $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$, one may then attempt to solve the feasibility problem

$$\begin{aligned} &\text{find } \mathbf{x} \in \mathcal{X} \\ &\text{s.t. } \ \mathbf{y} = \mathbf{A}\mathbf{x} \end{aligned} \tag{10}$$

to recover $\mathring{\mathbf{x}}$. This program has a unique solution if $\mathcal{X}$ intersects the affine subspace $E_{\mathring{\mathbf{x}}} := \left\{\mathbf{z}\in\mathbb{R}^d : \mathbf{A}\mathbf{z} = \mathbf{A}\mathring{\mathbf{x}}\right\}$ only at the solution $\mathring{\mathbf{x}}$, i.e.,

$$\begin{aligned} &\mathcal{X} \cap E_{\mathring{\mathbf{x}}} = \left\{\mathring{\mathbf{x}}\right\} \\ \iff\ &(\mathcal{X} - \mathring{\mathbf{x}}) \cap (E_{\mathring{\mathbf{x}}} - \mathring{\mathbf{x}}) = \{\mathbf{0}\} \\ \iff\ &(\mathcal{X} - \mathring{\mathbf{x}}) \cap \ker(\mathbf{A}) = \{\mathbf{0}\}. \end{aligned} \tag{11}$$

Since Definition 4 required $\mathrm{conv}(\mathcal{A})$ to be a symmetric convex body, it is also a closed star domain.[8] In this case, we may use a well-known result from functional analysis that allows us to express $\mathcal{X}$ in terms of the 1-sublevel set of its Minkowski functional [88]

$$\begin{aligned} \gamma_{\mathcal{X}}(\mathbf{x}) &= \inf\left\{t > 0 : \mathbf{x} \in t\left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}} \mathrm{conv}(\mathcal{A})\right\} \\ &= \frac{1}{\left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}}} \inf\left\{t > 0 : \mathbf{x} \in t\,\mathrm{conv}(\mathcal{A})\right\} = \frac{\|\mathbf{x}\|_{\mathcal{A}}}{\left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}}}. \end{aligned}$$

Thus we have that

[8] A set K is a closed star domain if K is closed, and $tK \subseteq K\ \forall t \in [0, 1]$.

$$\begin{aligned}\mathcal{X} - \mathring{\mathbf{x}} &= \left\{\mathbf{x} \in \mathbb{R}^d : \gamma_{\mathcal{X}}(\mathbf{x}) \leq 1\right\} - \mathring{\mathbf{x}} \\ &= \left\{\mathbf{x} - \mathring{\mathbf{x}} : \|\mathbf{x}\|_{\mathcal{A}} \leq \left\|\mathring{\mathbf{x}}\right\|_{\mathcal{A}}\right\} \\ &= \mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}}),\end{aligned}$$

yielding again the uniqueness condition stated in Eq. (9).

Since $\|\cdot\|_{\mathcal{A}}$ defines a norm on $\mathbb{R}^d$, the set of descent directions is a convex body. We may therefore replace $\mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}})$ in Eq. (9) by its conic hull without changing the statement. This set, denoted by

$$\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) := \operatorname{cone} \mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}}),$$

is usually referred to as the *tangent* or *descent cone* of $\|\cdot\|_{\mathcal{A}}$ at $\mathring{\mathbf{x}}$, and represents a central object in the study of convex analysis. This ultimately leads to the following result.

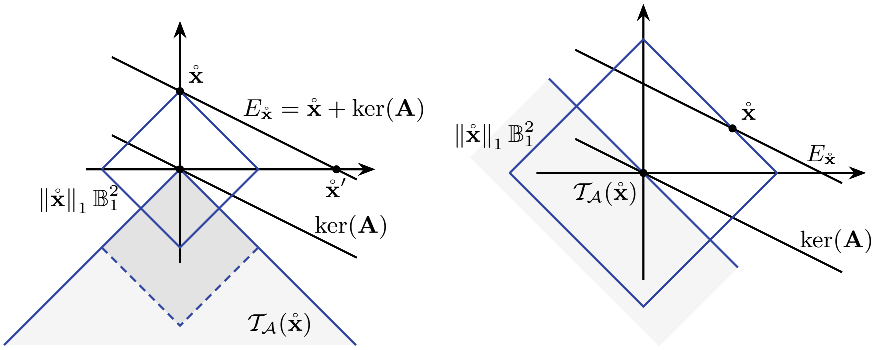
Proposition 1 ([13, Proposition 2.1]) *The vector* $\mathring{\mathbf{x}}$ *is the unique solution of Problem* ($\mathrm{P}_{\mathcal{A}}$) *if and only if*

$$\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \ker(\mathbf{A}) = \{\mathbf{0}\}. \tag{12}$$

As a typical example application of Proposition 1, consider the atomic set $\mathcal{A} = \{\pm \mathbf{e}_i\} \subset \mathbb{R}^d$ of signed unit vectors. The convex hull of this set is the ℓ_1-unit ball in $\mathbb{R}^d$, and hence $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_1$; the conic hull is all of $\mathbb{R}^d$. However, if we restrict attention to nonnegative linear combinations of at most k elements in $\mathcal{A}$, we obtain the set $\mathcal{K} = \operatorname{cone}_k(\mathcal{A}) = \left\{\mathbf{x} \in \mathbb{R}^d : |\operatorname{supp}(\mathbf{x})| \leq k\right\} = \Sigma_k(\mathbb{R}^d)$ of k-sparse vectors. As illustrated in Fig. 2a, the 1-sparse vector $\mathring{\mathbf{x}}$ can be uniquely recovered via ℓ_1-minimization since its tangent cone $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$ intersects the null space of $\mathbf{A}$ only at $\{\mathbf{0}\}$. On the other hand, if $\mathring{\mathbf{x}}$ is as depicted in Fig. 2b, then the tangent cone of $\mathcal{A}$ at $\mathring{\mathbf{x}}$ corresponds to a rotated half-space. Since every 1-dimensional subspace of $\mathbb{R}^2$ clearly intersects this half-space at arbitrarily many points, the only way a vector on a 2-dimensional face of $\left\|\mathring{\mathbf{x}}\right\|_1 \mathbb{B}_1^2$ can be recovered is if $\ker(\mathbf{A})$ is the 0-dimensional subspace $\{\mathbf{0}\}$, i.e., if $\mathbf{A}$ has full-rank. Finally, note that the vector $\mathring{\mathbf{x}}'$ in Fig. 2a cannot be recovered either despite sharing the same sparsity structure as $\mathring{\mathbf{x}}$. Conceptually, this is immediately obvious from the fact that $\|\mathring{\mathbf{x}}\|_1 < \|\mathring{\mathbf{x}}'\|_1$ which implies that even if we were to observe $\mathring{\mathbf{x}}'$, atomic norm minimization would still yield the solution $\mathbf{x}^\star = \mathring{\mathbf{x}}$. In light of Proposition 1, this is explained by the fact that the tangent cone at $\mathring{\mathbf{x}}'$ has the same shape as $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$ but rotated 90° clockwise so that $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}')$ and $\ker(\mathbf{A})$ share a ray, violating the uniqueness condition (12). This example demonstrates the nonuniform character of the recovery condition of Proposition 1 which locally depends on the particular choice of $\mathring{\mathbf{x}}$.

Since the tangent cone is a bigger set than $\mathcal{D}_{\mathcal{A}}(\mathring{\mathbf{x}})$, the condition

$$\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \ker(\mathbf{A}) = \{\mathbf{0}\}$$


$\mathring{\mathbf{x}}$
$E_{\mathring{\mathbf{x}}} = \mathring{\mathbf{x}} + \ker(\mathbf{A})$
$\|\mathring{\mathbf{x}}\|_1 \, \mathbb{B}_1^2$
$\mathring{\mathbf{x}}'$
$\ker(\mathbf{A})$
$\mathcal{T}_\mathcal{A}(\mathring{\mathbf{x}})$
$\|\mathring{\mathbf{x}}\|_1 \, \mathbb{B}_1^2$
$\mathring{\mathbf{x}}$
$E_{\mathring{\mathbf{x}}}$
$\mathcal{T}_\mathcal{A}(\mathring{\mathbf{x}})$
$\ker(\mathbf{A})$

Recovery of 1-sparse vectors

Recovery of a 2-sparse vector

Fig. 2 Recovery of vectors in $\mathbb{R}^2$

in a sense represents a stronger requirement than $\mathcal{D}_\mathcal{A}(\mathring{\mathbf{x}}) \cap \ker(\mathbf{A})$ from before. Moreover, while Proposition 1 provides a necessary and sufficient condition for the successful recovery of individual vectors via Problem ($\mathrm{P}_\mathcal{A}$), testing the condition in practice ultimately requires prior knowledge of the solution $\mathring{\mathbf{x}}$ which we aim to recover. However, as we will see shortly, both issues can be elegantly circumvented by turning to the probabilistic setting where we assume the elements of the measurement matrix are drawn independently from the standard Gaussian distribution. This will allow us to draw on a powerful result from asymptotic convex geometry to assess the success of recovering individual vectors probabilistically. Before stating this result, we first need to introduce the concept of *Gaussian mean width* or *mean width* for short, an important summary parameter of a bounded set.

Definition 6 (*Gaussian mean width*) The Gaussian mean width of a bounded set $\boldsymbol{\Omega}$ is defined as

$$w(\boldsymbol{\Omega}) := \mathbb{E} \sup_{\mathbf{x} \in \boldsymbol{\Omega}} \langle \mathbf{g}, \mathbf{x} \rangle, \tag{13}$$

where $\mathbf{g} \sim \mathsf{N}(\mathbf{0}, \mathrm{Id})$ is an isotropic zero-mean Gaussian random vector.

The Gaussian mean width is closely related to the spherical mean width

$$w_\mathbb{S}(\boldsymbol{\Omega}) := \mathbb{E} \sup_{\mathbf{x} \in \boldsymbol{\Omega}} \langle \boldsymbol{\eta}, \mathbf{x} \rangle,$$

where $\boldsymbol{\eta}$ is a random d-vector drawn uniformly from the Haar measure on the sphere. Since length and direction of a Gaussian random vector are independent by rotation invariance of the Gaussian distribution, we can decompose every standard Gaussian vector $\mathbf{g}$ as $\mathbf{g} = \|\mathbf{g}\|_2 \, \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is again drawn from the uniform Haar measure. The Gaussian and spherical mean width are therefore related by

$$w(\boldsymbol{\Omega}) = \mathbb{E}\, \|\mathbf{g}\|_2\, w_{\mathbb{S}}(\boldsymbol{\Omega}) \leq \sqrt{d} w_{\mathbb{S}}(\boldsymbol{\Omega}),$$

where the last step follows from Jensen's inequality. Intuitively, the mean width of a bounded set measures its average diameter over all directions chosen uniformly at random. Consider for a moment the mean width $w(\boldsymbol{\Omega} - \boldsymbol{\Omega})$ of the Minkowski difference of $\boldsymbol{\Omega}$ with itself. Then we immediately have

$$\begin{aligned} w(\boldsymbol{\Omega} - \boldsymbol{\Omega}) &= \mathbb{E} \sup_{\mathbf{d} \in \boldsymbol{\Omega} - \boldsymbol{\Omega}} \langle \mathbf{g}, \mathbf{d} \rangle \\ &= \mathbb{E} \sup_{\mathbf{x}, \mathbf{z} \in \boldsymbol{\Omega}} \langle \mathbf{g}, \mathbf{x} - \mathbf{z} \rangle \\ &\leq 2\mathbb{E} \sup_{\mathbf{x} \in \boldsymbol{\Omega}} \langle \mathbf{g}, \mathbf{x} \rangle = 2w(\boldsymbol{\Omega}) \end{aligned}$$

with equality if $\boldsymbol{\Omega}$ is origin-symmetric. Given a realization of the random vector $\mathbf{g}$, the term $\sup_{\mathbf{x},\mathbf{z}\in\boldsymbol{\Omega}} \langle \mathbf{g}, \mathbf{x} - \mathbf{z} \rangle$ then corresponds to the distance of two supporting hyperplanes to $\boldsymbol{\Omega}$ with normal $\mathbf{g}$, scaled by $\|\mathbf{g}\|_2$.

With the definition of the mean width in place, we are now ready to state the following result known as *Gordon's escape through a mesh* or simply *Gordon's escape theorem*. We present here a version of the theorem adopted from [31, Corollary 3.3]. The original result was first presented in [57].

Theorem 2 (Gordon's escape through a mesh) *Let $S \subset \mathbb{S}^{d-1}$, and let E be a random $(d - m)$-dimensional subspace of $\mathbb{R}^d$ drawn uniformly from the Haar measure on the Grassmann manifold $\mathcal{G}(d, d - m)$.*[9] *Then*

$$\mathbb{P}(S \cap E = \emptyset) \geq 1 - \exp\left(-\frac{1}{2}\left[\frac{m}{\sqrt{m+1}} - w(S)\right]^2\right)$$

provided

$$m \geq w(S)^2 + 1.$$

In words, Gordon's escape through a mesh phenomenon asserts that a randomly drawn subspace misses a subset of the Euclidean unit sphere with overwhelmingly high probability if the codimension m of the subspace is on the order of $w(S)^2$. Moreover, the probability of this event only depends on the codimension m of the subspace, as well as on the Gaussian width of the sphere patch S. In order to apply this result to the situation of Proposition 1 in the context of the standard Gaussian measurement ensemble, we merely need to restrict the tangent cone $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$ to the sphere, i.e., $S = \mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1}$, and choose $E = \ker(\mathbf{A})$. This immediately yields the following straightforward specialization of Theorem 2.

[9] The *Grassmann manifold* or *Grassmannian* $\mathcal{G}(d, s)$ is an abstract Riemannian manifold containing all s-dimensional subspaces of $\mathbb{R}^d$.

Corollary 1 (Exact recovery from Gaussian observations) *Let* $\mathbf{A} \in \mathbb{R}^{m \times d}$ *be a matrix populated with independent standard Gaussian entries, and let* $\mathring{\mathbf{x}} \in \operatorname{cone}_k(\mathcal{A})$. *Then* $\mathring{\mathbf{x}}$ *can be perfectly recovered from its measurements* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ *via atomic norm minimization with probability at least* $1 - \eta$ *if*

$$m \geq \left(w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1}) + \sqrt{2 \log(\eta^{-1})} \right)^2.$$

So far, we have only concerned ourselves with establishing conditions under which an arbitrary vector could be uniquely recovered from its linear measurements by solving Problem (8). In fact, nothing in our discussion so far precludes that this undertaking might require us to take at least as many measurements as the linear algebraic dimension of the vector space containing $\mathring{\mathbf{x}}$. The power of the presented approach lies in the fact that for many signal models of interest such as sparse vectors, group-sparse vectors, and low-rank matrices, the tangent cone at points $\mathring{\mathbf{x}}$ lying on low-dimensional faces of a scaled version of $\operatorname{conv}(\mathcal{A})$ is narrow (cf. Fig. 2), and therefore exhibit small mean widths. Coming back to the canonical example of sparse vectors as discussed before, it can be shown that $w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1})$ roughly scales like $\sqrt{k \log(d/k)}$ for any $\mathring{\mathbf{x}} \in \Sigma_k(\mathbb{R}^d)$ (see, for instance, [31, 89]). In light of Corollary 1, this requires m to scale linearly in k, and only logarithmically in the ambient dimension d. For convenience, we list some of the best known bounds for the mean widths of tangent cones associated with the signal models introduced in Sect. 3 in Table 1 [55].

Without going into too much detail, we want to briefly comment on a few natural extensions of Corollary 1.

Extensions to Noisy Recovery and Subgaussian Observations

An obvious question to ask at this point is what kind of recovery performance we might expect if we extend our sensing model to include additive noise of the form $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{w}$ with $\|\mathbf{w}\|_2 \leq \sigma$ as a more realistic model of observation. Naturally, we cannot hope to ever recover $\mathring{\mathbf{x}}$ exactly in that case unless $\sigma = 0$. Nevertheless, one should still expect to be able to control the recovery quality in terms of the mean width of the tangent cone and the noise level σ by an appropriate choice of m. The following result, which was adapted from [31, Corollary 3.3], demonstrates that this is in fact the case if we solve the noise-constrained atomic norm minimization problem

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{x}\|_{\mathcal{A}} \\ \text{s.t.} \quad & \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \sigma. \end{aligned} \tag{14}$$

Proposition 2 (Robust recovery from Gaussian observations) *Let* $\mathbf{A}$ *and* $\mathring{\mathbf{x}}$ *be as in Corollary 1. Assume we observe* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{w}$ *with* $\|\mathbf{w}\|_2 \leq \sigma$. *Then with probability at least* $1 - \eta$, *the solution* $\mathbf{x}^\star$ *of Problem* (14) *satisfies*

$$\left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_2 \le \nu$$

provided

$$m \ge \left(\frac{w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1}) + \sqrt{2\log(\eta^{-1})}}{1 - 2\sigma/\nu} \right)^2 .$$

Note that the reconstruction fidelity ν in Proposition 2 is inherently limited by the noise level σ since we require $\nu > 2\sigma$ for the bound on m to yield sensible values.

In closing, we also want to mention a recent extension of Gordon's escape theorem to measurement matrices whose rows are independent copies of subgaussian isotropic random vectors $\mathbf{a}_i \in \mathbb{R}^d$ with subgaussian parameter τ, i.e.,

$$\mathbb{E}(\mathbf{a}_i \mathbf{a}_i^\top) = \mathrm{Id}, \quad \|\mathbf{a}_i\|_{\psi_2} = \sup_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} \|\left\langle \boldsymbol{\theta}, \mathbf{a}_i \right\rangle\|_{\psi_2} \le \tau. \tag{15}$$

Based on a concentration result for such matrices acting on bounded subsets of $\mathbb{R}^d$ [66, Corollary 1.5], Liaw et al. proved a general version of the following result which we state here in the context of signal recovery in the same vein as Corollary 1.

Theorem 3 (Exact recovery from subgaussian observations) *Let* $\mathbf{A} \in \mathbb{R}^{m \times d}$ *be a matrix whose rows are independent subgaussian random vectors satisfying Eq.* (15)*, and let* $\mathring{\mathbf{x}} \in \mathrm{cone}_k(\mathcal{A})$*. Then with probability at least* $1 - \eta$*,* $\mathring{\mathbf{x}}$ *is the unique minimizer of Problem* (8) *with* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ *if*

$$m \gtrsim \tau^4 \left(w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1}) + \sqrt{\log(\eta^{-1})} \right)^2 .$$

Surprisingly, this bound suggests almost the same scaling behavior as in the Gaussian case (cf. Corollary 1), barring the dependence on the subgaussian parameter τ, as well as an absolute constant hidden in the notation.

The results mentioned so far are not without their own set of drawbacks. While robustness against noise was established in Proposition 2, the tangent cone characterization is inherently susceptible to model deficiencies. For instance, consider again the example $\mathcal{A} = \{\pm \mathbf{e}_i\}$ giving rise to the set of $\Sigma_k(\mathbb{R}^d)$. If $\mathring{\mathbf{x}}$ is not a sparse linear combination of elements in $\mathcal{A}$ (e.g., $\mathring{\mathbf{x}}$ may only be compressible rather than exactly sparse), then the tangent cone of $\|\cdot\|_{\mathcal{A}}$ at $\mathring{\mathbf{x}}$ may not have a small mean width at all as we saw in Fig. 2. In fact, in this case, $w(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}}) \cap \mathbb{S}^{d-1})^2$ is usually on the order of the ambient dimension d [80]. Moreover, as we also demonstrated graphically in Fig. 2, the recovery guarantees presented in this section only apply to individual vectors. Such results are customarily referred to as nonuniform guarantees in the compressed sensing literature. Before moving on to the uniform recovery case which provides recovery conditions for *all* vectors in a signal class simultaneously, we want to briefly comment on an important line of work connecting sparse recovery with the field of conic integral geometry. This is the subject of the next section.

4.2 Connections to Conic Integral Geometry

In an independent line of research [4], the sparse recovery problem was recently approached from the perspective of conic integral geometry. At the heart of this field lies the study of the so-called *intrinsic volumes of cones*. We limit our discussion to the important class of polyhedral cones[10] here, and refer interested readers to [4] for a treatment of general convex cones.

Definition 7 (*Intrinsic volumes*) Let $\mathcal{C}$ be a polyhedral cone in $\mathbb{R}^d$, and denote by $\mathbf{g}$ a standard Gaussian random vector. Then for $i = 0, \dots, d$, the ith intrinsic volume of $\mathcal{C}$ is defined as

$$v_i(\mathcal{C}) := \mathbb{P}(\Pi_{\mathcal{C}}(\mathbf{g}) \in \mathcal{F}_i(\mathcal{C})),$$

where $\Pi_{\mathcal{C}}$ denotes the orthogonal projector on $\mathcal{C}$, and $\mathcal{F}_i(\mathcal{C})$ denotes the union of relative interiors of all i-dimensional faces of $\mathcal{C}$.

If we are given two non-empty convex cones $\mathcal{C}, \mathcal{D} \subset \mathbb{R}^d$, one of which is not a subspace, and we draw an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$ from the uniform Haar measure, then the probability that $\mathcal{C}$ and the randomly rotated cone $\mathbf{Q}\mathcal{D}$ intersect nontrivially is fully determined by the intrinsic volumes of $\mathcal{C}$ and $\mathcal{D}$. The precise statement of this result is known as the *conic kinematic formula*.

Theorem 4 (Conic kinematic formula, [4, Fact 2.1]) *Let $\mathcal{C}$ and $\mathcal{D}$ be two non-empty closed convex cones in $\mathbb{R}^d$ of which at most one is a subspace. Denote by $\mathbf{Q} \in \mathrm{O}(d)$ a matrix drawn uniformly from the Haar measure on the orthogonal group. Then*

$$\mathbb{P}(\mathcal{C} \cap \mathbf{Q}\mathcal{D} \neq \{\mathbf{0}\}) = \sum_{i=0}^{d} (1 + (-1)^{i+1}) \sum_{j=1}^{d} v_i(\mathcal{C}) v_{d+i-j}(\mathcal{D}).$$

To apply this result to the context of sparse recovery as discussed in the previous section, one simply chooses $\mathcal{C} = \mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$, and $\mathcal{D} = \ker(\mathbf{A})$, similar to the situation of Gordon's escape theorem. While the intrinsic volumes of $\ker(\mathbf{A})$, a $(d-m)$-dimensional linear subspace, are easily determined by[11]

$$v_i(\ker(\mathbf{A})) = \begin{cases} 1, & i = d - m, \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

the calculation of the intrinsic volumes of tangent cones is much less straightforward. Fortunately, there is an elegant way out of this situation which was first demonstrated

[10] A cone $\mathcal{C} \subset \mathbb{R}^d$ is called *polyhedral* if it can be expressed as the intersection of finitely many half-spaces.

[11] This follows from the fact that $\ker(\mathbf{A})$ only has a single face on which $\Pi_{\ker(\mathbf{A})}$ projects every point $\mathbf{x} \in \mathbb{R}^d$, namely, $\ker(\mathbf{A})$ itself.

in [4]. Since any vector $\mathbf{x} \in \mathbb{R}^d$ projected on a closed convex cone $\mathcal{C}$ must belong to exactly one of the $d+1$ sets $\mathcal{F}_i(\mathcal{C})$ defined in Definition 7, the collection $\{v_i(\mathcal{C})\}_{i=0}^d$ of intrinsic volumes defines a discrete probability distribution on $\{0, 1, \ldots, d\}$. Moreover, the distribution can be shown to concentrate sharply around its expectation

$$\delta(\mathcal{C}) := \sum_{i=0}^{d} i v_i(\mathcal{C}),$$

known as the *statistical dimension* of $\mathcal{C}$, which in turn can be tightly estimated in many cases of interest by appealing to techniques from convex analysis. In fact, the same technique was previously used in [31] to derive tight estimates of the mean width of various tangent cones. Note, however, that this work merely exploited a numerical relation between the Gaussian mean width and the statistical dimension which we will comment on below but was not generally motivated by conic integral geometry. The concentration behavior of intrinsic volumes ultimately allowed Amelunxen et al. to derive the following remarkable pair of bounds which constitute a breakthrough result in the theory of sparse recovery.

Theorem 5 (Approximate conic kinematic formula, [4, Theorem II]) *Let* $\mathring{\mathbf{x}} \in \operatorname{cone}_k(\mathcal{A})$*, and denote by* $\mathbf{A} \in \mathbb{R}^{m \times d}$ *a standard Gaussian matrix with independent entries as usual. Given the linear observations* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$*, and denoting by* $\mathbf{x}^\star$ *the optimal solution of Problem* (8)*, the following two statements hold for* $\eta \in (0, 1]$:

$$\begin{aligned} \mathbb{P}(\mathbf{x}^\star = \mathring{\mathbf{x}}) \geq 1 - \eta \quad &\textit{if} \quad m \geq \delta(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})) + c_\eta \sqrt{d}, \\ \mathbb{P}(\mathbf{x}^\star \neq \mathring{\mathbf{x}}) \leq \eta \quad &\textit{if} \quad m \leq \delta(\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})) - c_\eta \sqrt{d} \end{aligned}$$

with $c_\eta = \sqrt{8 \log(4/\eta)}$*.*

Before addressing the problem of estimating the statistical dimension δ of the tangent cone $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$, let us briefly comment on the above result first. Theorem 5 is remarkable for a variety of reasons. First, as was demonstrated numerically in [4], the two bounds correctly predict the position of the so-called phase transition. Such results were previously only known in the asymptotic large-system limit (cf. [43, 45]) where one considers for $d, m, k \to \infty$ the fixed ratios $\delta := m/d$, and $\rho := k/m$ over the open unit square $(0, 1)^2$. The phase-transition phenomenon describes a particular behavior of the system which exhibits a certain critical line $\rho^\star = \rho^\star(\delta)$ that partitions $(0, 1)^2$ into two distinct regions: one where recovery almost certainly succeeds, and one where it almost certainly fails. The transition line then corresponds to the 50th percentile. Second, it represents the first non-asymptotic result which correctly predicts a fundamental limit below which sparse recovery will fail with high probability. This is in stark contrast to previous results based on Gordon's escape theorem which were only able to predict that recovery would succeed above a certain threshold but could not make any assessment of the behavior below it. Finally, as a result of the second point, Theorem 5 represents the first result which quantifies the width of the transition region where the probability of exact recovery will change from almost certain failure to almost certain success. Once again we refer interested readers to the excellent exposition [4], particularly Sect. 10, for a thorough comparison of their results to the pertinent literature on the existence of phase transitions in compressed sensing.

The key ingredient in the application of Theorem 5 is the statistical dimension δ of the tangent cone $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$. As mentioned above, the statistical dimension is defined as the expected value of the distribution defined by the intrinsic volumes of $\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$. However, it admits two alternative representations which can be leveraged to estimate $\delta(\mathcal{C})$, especially when $\mathcal{C}$ corresponds to a tangent cone. This is the content of the following result.

Proposition 3 (Statistical dimension, [4, Proposition 3.1]) *Let $\mathcal{C}$ be a closed convex cone in $\mathbb{R}^d$, and let $\mathbf{g}$ be a standard Gaussian d-vector. Then*

$$\delta(\mathcal{C}) = \sum_{i=0}^{d} i v_i(\mathcal{C}) = \mathbb{E}\big[\|\Pi_{\mathcal{C}}(\mathbf{g})\|_2^2\big] = \mathbb{E}\big[\operatorname{dist}(\mathbf{g}, \mathcal{C}^\circ)^2\big],$$

where $\mathcal{C}^\circ := \big\{\mathbf{z} \in \mathbb{R}^d : \langle \mathbf{x}, \mathbf{z}\rangle \leq 0 \ \forall \mathbf{x} \in \mathcal{C}\big\}$ denotes the polar cone of $\mathcal{C}$.

In particular, we want to focus on the last identity when $\mathcal{C} = \mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})$. In fact, in this situation one may exploit a well-known fact from convex geometry that states that the polar cone of the tangent cone corresponds to the normal cone [88]

$$\begin{aligned}\mathcal{N}_{\mathcal{A}}(\mathring{\mathbf{x}}) &:= \big\{\mathbf{v} \in \mathbb{R}^d : \big\langle \mathbf{v}, \mathbf{x} - \mathring{\mathbf{x}}\big\rangle \leq 0 \ \forall \mathbf{x}\colon \ \|\mathbf{x}\|_{\mathcal{A}} \leq \|\mathring{\mathbf{x}}\|_{\mathcal{A}}\big\} \\ &= \big\{\mathbf{v} \in \mathbb{R}^d : \langle \mathbf{v}, \mathbf{d}\rangle \leq 0 \ \forall \mathbf{d} \in \mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})\big\},\end{aligned}$$

which in turn can be expressed as the conic hull of the subdifferential of the atomic norm at $\mathring{\mathbf{x}}$,

$$\mathcal{T}_{\mathcal{A}}(\mathring{\mathbf{x}})^\circ = \mathcal{N}_{\mathcal{A}}(\mathring{\mathbf{x}}) = \operatorname{cone}(\partial \|\mathring{\mathbf{x}}\|_{\mathcal{A}}) = \bigcup_{t \geq 0} t\partial \|\mathring{\mathbf{x}}\|_{\mathcal{A}}.$$

The last identity follows from the fact that the subdifferential of a convex function is always a convex set. In other words, given a recipe for the subdifferential of the atomic norm, the statistical dimension of its associated tangent cone can be estimated by bounding the expected distance of a Gaussian vector to its convex hull. In many cases of interest, this turns out to be a comparatively easy task (see, e.g., [31, Appendix C], [55, Appendix A] and [4, Sect. 4]).

As alluded to before, the statistical dimension also shares a close connection to the Gaussian mean width. In particular, we have the following two inequalities (cf. [4, Proposition 10.2]):

$$w(\mathcal{C} \cap \mathbb{S}^{d-1})^2 \leq \mathbb{E}\big[\operatorname{dist}(\mathbf{g}, \mathcal{C}^\circ)^2\big] = \delta(\mathcal{C}) \leq w(\mathcal{C} \cap \mathbb{S}^{d-1})^2 + 1.$$

This shows that estimating the mean width is qualitatively equivalent to estimating δ. As previously mentioned, this connection was used in [31] to derive precise bounds for the mean widths of the tangent cones for sparse vectors, and low-rank matrices, as well as for block- and group-sparse signals in [55] and [84], respectively. Note that the connection between mean width and statistical dimension was already used in the pioneering works of Stojnic [91], as well as Oymak and Hassibi [76], even if the term *statistical dimension* was originally coined in [4] where the connection between the probability distribution induced by the intrinsic volumes and its projective characterization in Proposition 3 was first established. We want to emphasize again that the fundamental significance of the statistical dimension in the context of sparse recovery did not become clear until the seminal work of Amelunxen, Lotz, McCoy, and Tropp who rigorously demonstrated the concentration behavior of intrinsic volumes, culminating in the breakthrough result stated in Theorem 5. In the same context, the authors argued that the statistical dimension generally represents a more appropriate measure of "dimension" of cones than the mean width. For instance, if $\mathcal{C}$ is an n-dimensional linear subspace L_n of $\mathbb{R}^d$, then it immediately follows from Eq. (16) that $\delta(L_n) = \dim(L_n) = n$. Moreover, given a closed convex cone $\mathcal{C} \subset \mathbb{R}^d$, we have $\delta(\mathcal{C}) + \delta(\mathcal{C}^\circ) = d$ (cf. [4, Proposition 3.1]) which generalizes the property $\dim(L_n) + \dim(L_n^\perp) = d$ from linear subspaces to convex cones since $L_n^\circ = L_n^\perp$, i.e., the polar cone of a subspace is its orthogonal complement.

The concepts discussed in this section all addressed the problem of recovering or estimating individual vectors with a low-complexity structure from low-dimensional linear measurements. In other words, given two vectors $\mathring{\mathbf{x}}$ and $\mathring{\mathbf{x}}'$ with the same low-complexity structure, and the knowledge that $\mathring{\mathbf{x}}$ can be estimated with a particular accuracy, we are not able to infer that the same accuracy also holds when we try to recover $\mathring{\mathbf{x}}'$ given a fixed choice of $\mathbf{A}$. Recall, for example, the situation illustrated in Fig. 2a. If instead of $\mathring{\mathbf{x}}$ we observe a vector $\mathring{\mathbf{x}}'$ positioned on the rightmost vertex of the scaled ℓ_1-ball, the tangent cone at $\mathring{\mathbf{x}}'$ now corresponds to the tangent cone at $\mathring{\mathbf{x}}$ rotated 90° clockwise around the origin. However, since this cone intersects the null space of $\mathbf{A}$ at arbitrarily many points, we are not able to recover $\mathring{\mathbf{x}}$ and $\mathring{\mathbf{x}}'$ simultaneously. In the parlance of probability theory, we might say that the results presented in this section are conditioned on a particular choice of $\mathring{\mathbf{x}}$. Such results are therefore known as *nonuniform* guarantees as they do not hold uniformly for all signals in a particular class at once.

In contrast, in the next section, we will introduce a variety of properties of measurement matrices which will allow us to characterize the recovery behavior uniformly over all elements in a signal class given the same choice of measurement matrix. Most importantly, we will focus on a particularly important property which not only yields a sufficient condition for perfect recovery of sparse vectors but one which has also proven an indispensable tool in providing stability and robustness conditions in situations where we are tasked with the recovery of signals from corrupted measurements.

5 Exact Recovery of Sparse Vectors

In this section, we consider conditions under which the sparse linear inverse problem, in which we are to infer a d-dimensional vector $\mathring{\mathbf{x}} \in \Sigma_k$ from its linear measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} \in \mathbb{R}^m$, admits a unique solution. In contrast to the situation discussed in Sect. 4, we are now specifically interested in conditions under which the entire set Σ_k can be recovered or at least well approximated by a single measurement matrix $\mathbf{A}$.

Consider two vectors $\mathbf{x}, \mathbf{z} \in \Sigma_k$, and suppose that both vectors are mapped to the same point $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$ such that $\mathbf{x} - \mathbf{z} \in \ker(\mathbf{A})$. Obviously, unless we specifically ask that $\mathbf{x} \neq \mathbf{z}$, there is absolutely no chance that we would ever be able to decide which element in Σ_k generated the measurements $\mathbf{y}$. In other words, if there is to be any hope to ever uniquely identify sparse vectors from their image under $\mathbf{A}$, the most fundamental condition we must impose is that no two vectors in Σ_k are mapped to the same point $\mathbf{y}$ in $\mathbb{R}^m$. However, since the difference of two k-sparse vectors is $2k$-sparse, this immediately yields the condition $\ker(\mathbf{A}) \cap \Sigma_{2k} = \{\mathbf{0}\}$. In words, the linear inverse problem for sparse vectors is well-posed if and only if the only $2k$-sparse vector contained in the null space of $\mathbf{A}$ is the zero vector.

Note that this viewpoint differs from the way we approached the recovery problem earlier in Sect. 4 where we merely asked for a particular optimization problem defined in terms of a fixed vector $\mathring{\mathbf{x}} \in \mathcal{K}$ to have a unique solution which ultimately lead us to the local tangent cone condition in Proposition 1. This also explains why, in the example depicted in Fig. 2a, we were able to recover the 1-sparse vector $\mathring{\mathbf{x}} \in \mathbb{R}^2$ but not the 1-sparse vector $\mathring{\mathbf{x}}'$. As the considerations above show, there simply is no circumstance under which we would ever be able to uniquely recover every 1-sparse vector in $\mathbb{R}^2$ from scalar measurements $y \in \mathbb{R}$. This is due to the fact that the null space of any matrix $\mathbf{A} \in \mathbb{R}^{1\times 2}$ (a row vector) either corresponds to a line through the origin or the entire plane $\mathbb{R}^2$ itself if $\mathbf{A} = \mathbf{0}$. However, since the set of 2-sparse vectors in $\mathbb{R}^2$ also corresponds to $\mathbb{R}^2$, the subspace $\ker(\mathbf{A})$ intersects Σ_2 at arbitrarily many points regardless of the choice of $\mathbf{A}$, violating the condition $\ker(\mathbf{A}) \cap \Sigma_2 = \{\mathbf{0}\}$.

The following theorem, which constitutes a key result in compressed sensing, formalizes the observations above.

Theorem 6 ([54, Theorem 2.13]) *Given a matrix* $\mathbf{A} \in \mathbb{C}^{m\times d}$*, the following statements are equivalent:*

1. *Given a vector* $\mathring{\mathbf{x}} \in \mathbb{C}^d$ *supported on a set of size at most k, the problem*

$$\begin{aligned} &\textit{minimize} \;\; \|\mathbf{x}\|_0 \\ &\quad\textit{s.t.} \quad\;\; \mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{x} \end{aligned} \tag{P$_0$}$$

   has a unique k-sparse minimizer, namely, $\mathbf{x}^\star = \mathring{\mathbf{x}}$.
2. *Every vector* $\mathring{\mathbf{x}}$ *is the unique k-sparse solution of the system* $\mathbf{A}\mathbf{z} = \mathbf{A}\mathring{\mathbf{x}}$.
3. *The only* $2k$*-sparse vector contained in the null space of* $\mathbf{A}$ *is the zero vector, i.e.,* $\ker(\mathbf{A}) \cap \Sigma_{2k} = \{\mathbf{0}\}$.

The key insight of the above result is the equivalence between the condition $\ker(\mathbf{A}) \cap \Sigma_{2k} = \{\mathbf{0}\}$, and the existence of sparse minimizers of a particularly important nonconvex optimization problem. More precisely, we have by Theorem 1 that a natural strategy to recover a sparse vector $\mathring{\mathbf{x}} \in \Sigma_k$ given $\mathbf{y}$ and $\mathbf{A}$ corresponds to a search for the sparsest element in the affine space $\left\{\mathbf{x} \in \mathbb{C}^d : \mathbf{A}\mathbf{x} = \mathbf{y}\right\}$.

One immediate question arising from Theorem 6 is "how underdetermined" the system $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ is allowed to become for there to still be a unique solution. Remarkably, Problem (P_0) can be shown to uniquely recover the original vector $\mathring{\mathbf{x}}$ as soon as the rank of the measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ exceeds the critical threshold $\operatorname{rank} \mathbf{A} \geq 2k$ [54]. In other words, every $2k$ columns of $\mathbf{A}$ must be linearly independent. Motivated by this observation, some authors refer to the so-called spark of a matrix—a portmanteau of the words "sparse" and "rank"—as the smallest number of linearly dependent columns of $\mathbf{A}$ [41]. With this definition, the rank constraint can be equivalently stated as $\operatorname{spark}(\mathbf{A}) > 2k$. Given a measurement matrix $\mathbf{A}$ of size $m \times d$ in the regime $m < d$, perfect recovery of any k-sparse vector is therefore guaranteed as soon as $\operatorname{spark} \mathbf{A} > 2k$. Moreover, since $\operatorname{rank}(\mathbf{A}) \leq m$, the rank requirement $\operatorname{rank}(\mathbf{A}) \geq 2k$ ultimately yields the necessary condition $m \geq 2k$ for perfect recovery of all k-sparse vectors via ℓ_0-minimization.

As alluded to before, an important distinction between the rank characterization above, and the tangent cone condition from Proposition 1 is that the latter only applies to individual elements of Σ_k while the requirement $\operatorname{rank}(\mathbf{A}) \geq 2k$ implies perfect recovery of every k-sparse vector via ℓ_0-minimization. If we are only interested in a nonuniform recovery condition, it turns out that we already get by with $m \geq k + 1$ measurements [54, Sect. 2.2]. Note, however, that the condition in Proposition 1 is based on a tractable optimization problem. This stands in stark contrast to the ℓ_0-minimization problem (P_0) which is provably NP-hard as it can be reduced to the so-called *exact 3-set cover* problem which in turn is known to belong to the class of NP-complete problems [72]. As a result, solving Problem (P_0) requires a combinatorial search over all $\sum_{i=0}^{d} \binom{d}{i}$ possible subproblems if k is unknown and $\binom{d}{k}$ otherwise, both of which are intractable for even moderately sized problems. While there exist certain deterministic matrices which satisfy the rank condition such as Vandermonde matrices, as well as tractable algorithms such as *Prony's method* to solve the associated ℓ_0-minimization problem, the solution of the general problem remains out of reach unless $\mathrm{P} = \mathrm{NP}$. Moreover, another drawback of attempting to solve the ℓ_0-minimization problem directly is that it can be shown to be highly sensitive to measurement noise and sparsity defects [54, Chap. 2].

While Theorem 6 in and of itself already represents a fascinating result in the field of linear algebra, the story does not end there. Despite the seemingly dire situation we find ourselves in when attempting to find minimizers of Problem (P_0), one of the key insights in the theory of compressed sensing is that there is a convenient escape hatch in the form of convex relaxations. In fact, it turns out that under slightly more demanding conditions on the null space of $\mathbf{A}$, we are still able to faithfully recover sparse or approximately sparse vectors by turning to a particular relaxation of Problem (P_0). We are, of course, talking about the infamous ℓ_1-minimization problem which we already discussed implicitly in the context of atomic norm minimization

w. r. t. the atomic set $\mathcal{A} = \{\pm \mathbf{e}_i\}$ generating the set of sparse vectors. It is this insight which elevates the field of compressed sensing from a purely mathematical theory to a highly desirable tool with far-reaching implications in countless domains of engineering, physics, chemistry, and biology. Before discussing the particular conditions on $\mathbf{A}$ which allow for robust and most importantly efficient recovery of sparse vectors from underdetermined linear measurements, let us first state and briefly comment on what is by now probably one of the most well-known and well-studied optimization problems in mathematics to date.

In light of our discussion of compressible vectors in Sect. 3.1, the following optimization problem, famously known as the basis pursuit (BP) problem, naturally represents the closest convex relaxation of the nonconvex ℓ_0-minimization problem (P_0):

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{A}\mathring{\mathbf{x}}. \end{aligned} \tag{P$_1$}$$

Ignoring for a moment any structural properties on the vector $\mathring{\mathbf{x}}$ we aim to recover, as well as the properties of the measurement matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$, the program can be shown to yield m-sparse minimizers [54, Theorem 3.1]. This observation alone already serves as a strong indicator of the deep connection between ℓ_1-minimization and sparse recovery. Moreover, the relaxation can be solved in polynomial time by so-called interior-point methods, a class of algorithms which is by now considered a standard tool in the field of convex optimization. In particular, in the real setting, Problem (P_1) belongs to the class of linear programs (LPs), while in the complex case the problem can be transformed into a second-order cone program (SOCP) over the Cartesian product of d Lorentz cones $\mathcal{K}_\mathrm{L} := \left\{(\mathbf{z}, t) \in \mathbb{R}^2 \times \mathbb{R}_{\geq 0} : \|\mathbf{z}\|_2 \leq t\right\}$.

6 Characterization of Measurement Matrices

At the beginning of Sect. 4, we presented a necessary and sufficient condition for the exact recovery of vectors in simple sets from underdetermined linear measurements (cf. Proposition 1). This condition is very much local in nature as it depends on the particular choice of the vector one aims to recover. To circumvent this issue, we turned to random matrices which allowed us to draw on powerful probabilistic methods to bound the probability that, conditioned on the choice of a particular vector, we would be able to recover it via atomic norm minimization.

It turns out that in a sense, this strategy can be mirrored in the case of uniform recovery of sparse vectors. However, rather than directly estimating the probability that the condition in Theorem 3 as established in the previous section holds for a particular choice of random matrix, we first introduce a few common properties of general measurement matrices, some of which will enable us to state powerful recovery guarantees which hold over entire signal classes rather than individual vectors. In Sect. 7, we will then present a series of results which assert that for many different choices of random measurement ensembles, such properties can be shown to be satisfied with overwhelmingly high probability, provided the number of measurements is chosen appropriately.

6.1 Null Space Property

As alluded to before, the relaxation of the original ℓ_0-minimization problem to a tractable convex program comes at the price of a critical difference to Problem (P$_0$). While the only requirement for Problem (P$_0$) to recover the original vector $\mathring{\mathbf{x}} \in \Sigma_k$ was for the number of measurements to exceed $2k$, perfect recovery will now be dependent on a certain structural property of the null space of $\mathbf{A}$, aptly referred to as the null space property (NSP), which was first introduced in [33].

Definition 8 (*Null space property*) A matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is said to satisfy the NSP of order k if, for any set $S \subset [d]$ with $|S| \leq k$, we have

$$\|\mathbf{v}_S\|_1 < \left\|\mathbf{v}_{\overline{S}}\right\|_1 \qquad \forall \mathbf{v} \in \ker \mathbf{A} \backslash \{\mathbf{0}\}.$$

The definition of the null space property admits a few additional observations for vectors in the null space of $\mathbf{A}$. Consider again an index set $S \subset [d]$ of size at most k. Then for $\mathbf{v} \in \ker \mathbf{A} \setminus \{\mathbf{0}\}$ we have

$$\begin{aligned} \|\mathbf{v}\|_1 = \left\|\mathbf{v}_S + \mathbf{v}_{\overline{S}}\right\|_1 &= \|\mathbf{v}_S\|_1 + \left\|\mathbf{v}_{\overline{S}}\right\|_1 \\ &< \left\|\mathbf{v}_{\overline{S}}\right\|_1 + \left\|\mathbf{v}_{\overline{S}}\right\|_1 \\ &= 2\left\|\mathbf{v}_{\overline{S}}\right\|_1 . \end{aligned}$$

Moreover, if S is the set supporting the largest components of $\mathbf{v}$ in absolute value, one has with the definition of the best k-term approximation error in Eq. (5),

$$\|\mathbf{v}\|_1 < 2\sigma_k(\mathbf{v})_1 .$$

Finally, by the Cauchy–Schwarz inequality, we have that for any $\mathbf{v} \in \mathbb{C}^d$, it holds that $\|\mathbf{v}\|_1^2 \leq \|\mathbf{v}\|_0 \cdot \|\mathbf{v}\|_2^2$. Therefore, one often alternatively finds the condition

$$\|\mathbf{v}_S\|_2 < \frac{1}{\sqrt{k}} \left\|\mathbf{v}_{\overline{S}}\right\|_1$$

in the definition of the null space property.

Given a matrix that satisfies the null space property, we can now state the general result for the recovery of any k-sparse vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ from its linear measurements by solving the basis pursuit (BP) problem (BP) below. In particular, consider a vector $\mathbf{v} \in \ker \mathbf{A} \cap \Sigma_{2k}$ supported on an index set $S \subset [d]$ of size $2k$, and assume further that $\mathbf{v} \neq \mathbf{0}$. Then for two disjoint sets $S_1, S_2 \subset S$ with $S = S_1 \cup S_2$ and $|S_1| = |S_2| = k$, by the null space property we have $\|\mathbf{v}_{S_1}\|_1 < \|\mathbf{v}_{\overline{S_1}}\|_1 = \|\mathbf{v}_{S \setminus S_1}\|_1 = \|\mathbf{v}_{S_2}\|_1$ and $\|\mathbf{v}_{S_2}\|_1 < \|\mathbf{v}_{S_1}\|_1$ which is a contradiction, and hence $\mathbf{v} = \mathbf{0}$. In other words, the null space property implies that the null space of $\mathbf{A}$ only contains a single $2k$-sparse vector: the zero vector. This implies the condition we previously stated in Theorem 3 which said that ℓ_0-minimization can recover any k-sparse vector as long as the null space of the measurement matrix contains no $2k$-sparse vectors save for the zero vector. Amazingly, the null space property provides a necessary and sufficient condition for the following recovery guarantee for sparse vectors.

Theorem 7 *Let $\mathbf{A} \in \mathbb{C}^{m \times d}$ and $k \in [d]$. Then every k-sparse vector $\mathring{\mathbf{x}}$ is the unique minimizer of the basis pursuit problem*

$$\begin{aligned} \textit{minimize} \ & \|\mathbf{x}\|_1 \\ \textit{s.t.} \quad & \mathbf{y} = \mathbf{A}\mathbf{x} \end{aligned} \tag{BP}$$

with $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ iff $\mathbf{A}$ satisfies the null space property of order k.

Proof If $\mathbf{A}\mathring{\mathbf{x}} = \mathbf{A}\mathbf{z}$, then $\mathbf{d} := \mathring{\mathbf{x}} - \mathbf{z} \in \ker(\mathbf{A})$ with $\mathbf{d}_S = \mathring{\mathbf{x}} - \mathbf{z}_S$ and $\mathbf{d}_{\overline{S}} = \mathbf{z}_{\overline{S}}$. Invoking the null space property we have

$$\begin{aligned} \left\|\mathring{\mathbf{x}}\right\|_1 &= \left\|\mathring{\mathbf{x}} - \mathbf{z}_S + \mathbf{z}_S\right\|_1 \\ &\leq \|\mathbf{d}_S\|_1 + \|\mathbf{z}_S\|_1 \\ &< \left\|\mathbf{d}_{\overline{S}}\right\|_1 + \|\mathbf{z}_S\|_1 \\ &= \left\|\mathbf{z}_{\overline{S}}\right\|_1 + \|\mathbf{z}_S\|_1 = \|\mathbf{z}\|_1 . \end{aligned}$$

This means that $\mathring{\mathbf{x}}$ is the unique minimizer of (BP). For the other direction, every $\mathbf{v} \in \ker(\mathbf{A})$ satisfies $\mathbf{A}\mathbf{v}_S = \mathbf{A}(-\mathbf{v}_{\overline{S}})$. Since $\mathbf{v}_S$ is the unique minimizer of (BP), we have $\|\mathbf{v}_S\|_1 < \| - \mathbf{v}_{\overline{S}}\|_1$ which is the null space property. □

Two situations are of particular importance in linear inverse problems, namely, situations in which $\mathring{\mathbf{x}}$ is only approximately sparse, and when the measurements are corrupted by additive noise. It is therefore generally desirable for a recovery algorithm to be both robust to noise and stable w. r. t. to so-called sparsity defect. To that end, one can extend the definition of the null space property to provide similar guarantees to the one stated in Theorem 7. We first consider the so-called stable null space property which can be used to account for sparsity defects of vectors.

Definition 9 (*Stable null space property*) A matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is said to satisfy the stable null space property of order k with constant $0 < \rho < 1$ w. r. t. any set $S \subset [d]$ if

$$\|\mathbf{v}_S\|_1 \leq \rho \left\|\mathbf{v}_{\overline{S}}\right\|_1 \quad \forall \mathbf{v} \in \ker \mathbf{A}$$

with $|S| \leq k$.

With this definition in place, the following result characterizes the impact of sparsity defects on the recovery error of the basis pursuit problem.

Theorem 8 ([54, Theorem 4.12]) *Let* $\mathbf{A} \in \mathbb{C}^{m \times d}$ *and* $k \in [d]$. *Then with* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$, *the solution* $\mathbf{x}^\star$ *of Problem* (BP) *satisfies*

$$\left\|\mathbf{x}^\star - \mathring{\mathbf{x}}\right\|_2 \leq \frac{2(1+\rho)}{(1-\rho)} \sigma_k(\mathring{\mathbf{x}})_1$$

if $\mathbf{A}$ *satisfies the stable null space property of order k. In particular, if* $\mathring{\mathbf{x}} \in \Sigma_k$ *then* $\mathbf{x}^\star = \mathring{\mathbf{x}}$.

We can extend the definition of the stable null space property once more to also account for additive noise in the measurements. For reference, we state here the most general form of the so-called ℓ_q-robust null space property. However, instead of using this definition to state a stable, noise-robust counterpart to Theorem 8, we will instead turn to a more commonly used property of measurement matrices in the next section to state a guarantee of this type.

Definition 10 (*ℓ_q-robust null space property*) Let $q \geq 1$, and denote by $\|\cdot\|$ an arbitrary norm on $\mathbb{C}^m$. Then the matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfies the ℓ_q-robust null space property of order k with constants $0 < \rho < 1$ and $\tau > 0$ if for all $\mathbf{v} \in \mathbb{C}^d$,

$$\|\mathbf{v}_S\|_q \leq \frac{\rho}{k^{1-1/q}} \left\|\mathbf{v}_{\overline{S}}\right\|_1 + \tau \|\mathbf{A}\mathbf{v}\|$$

for all $S \subset [d]$, $|S| \leq k$.

Theorem 7 yields a necessary and sufficient condition for the matrix $\mathbf{A}$ that answers the central question when minimizers of (P_0) and (P_1) coincide. While this represents an invaluable result, Theorem 7 makes no statement regarding the actual existence of such matrices. As it turns out, constructing deterministic matrices which directly satisfy the null space property (or its stable or noise-robust variants) constitutes a highly nontrivial problem. In fact, even verifying whether a given matrix satisfies the null space property was eventually shown to be an NP-hard decision problem [95]. Fortunately, it can be shown that matrices satisfying the null space property still exist in abundance if one turns to random measurement ensembles. While it is possible to directly establish the existence of such matrices probabilistically,[12] it has become common practice in the compressed sensing literature to mainly consider an alternative property of measurement matrices to establish recovery guarantees. The property in question is of course the infamous restricted isometry property (RIP) which was introduced in one of the very first papers on compressed sensing [27], and by now constitutes one of the most well-studied objects in the theory.

[12] In fact, as we will briefly discuss in Sect. 7, such random constructions are often characterized by more well-behaved scaling constants.

6.2 Restricted Isometry Property

The restricted isometry property (RIP) was first introduced in the seminal work by Candes and Tao [27], and shown in [21] to allow for robust recovery of approximately sparse vectors in the presence of measurement noise. While this property only yields a sufficient condition implying the null space property, matrices of this type can be found—at least in a probabilistic sense—in abundance as various random measurement ensembles can be shown to satisfy the RIP with high probability (cf. Sect. 7). The property is defined as follows.

Definition 11 A matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is said to satisfy the RIP of order k if

$$(1-\delta)\, \|\mathbf{x}\|_2^2 \le \|\mathbf{A}\mathbf{x}\|_2^2 \le (1+\delta)\, \|\mathbf{x}\|_2^2$$

for all $\mathbf{x} \in \Sigma_k$ with $\delta \ge 0$. The smallest $\delta_k \le \delta$ satisfying this condition is called the restricted isometry constant (RIC) of $\mathbf{A}$.

Intuitively, this definition states that for any $S \subset [d]$ with $|S| \le k$ the submatrix $\mathbf{A}_S$ obtained by retaining only the columns indexed by S approximately acts like an isometry on the set of k-sparse vectors which admits an alternative characterization of the restricted isometry constant δ_k as

$$\delta_k = \max_{\substack{S \subset [d],\\ |S|=k}} \left\| \mathbf{A}_S^* \mathbf{A}_S - \mathrm{Id} \right\|_{2 \to 2}.$$

This definition of the restricted isometry constant is commonly used in proofs establishing the restricted isometry property in a probabilistic setting by showing that δ_k concentrates sharply around its expectation.

In light of the importance and popularity of the restricted isometry property in compressed sensing, we will state most recovery conditions of the various algorithms introduced in Sect. 8 exclusively in terms of the restricted isometry constants associated with the RIP matrices in question.

The restricted isometry property admits a particularly short and concise proof of why k-sparse vectors have unique measurement vectors $\mathbf{y}$ under projections through $\mathbf{A}$. Assume the matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfies the RIP condition of order $2k$ with constant $\delta_{2k} < 1$, and consider two distinct k-sparse vectors $\mathbf{x}, \mathbf{z} \in \mathbb{C}^d$ with $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{z}$. Define now $\mathbf{v} := \mathbf{x} - \mathbf{z} \in \Sigma_{2k}$, i.e., $\mathbf{A}\mathbf{v} = \mathbf{0}$. Then we have by the restricted isometry property,

$$0 < (1-\delta_{2k})\, \|\mathbf{v}\|_2^2 \le \|\mathbf{A}\mathbf{v}\|_2^2 = 0.$$

Since this only holds for $\mathbf{v} = \mathbf{0}$, we must have $\mathbf{x} = \mathbf{z}$. In other words, if $\mathbf{A}$ is an RIP matrix of order $2k$, no two k-sparse vectors are mapped to the same measurement vector $\mathbf{y}$ through $\mathbf{A}$.

In the following, we consider noisy measurements of the form $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ where the additive noise term $\mathbf{e} \in \mathbb{C}^m$ is assumed to be bounded according to $\|\mathbf{e}\|_2 \le \eta$.

Under assumption of the restricted isometry property, one may then establish the following stable and robust recovery result.

Theorem 9 ([54, Theorem 6.12]) *Let* $\mathbf{A} \in \mathbb{C}^{m \times d}$ *be a matrix satisfying the RIP of order* $2k$ *with restricted isometry constant* $\delta_{2k} < 4/\sqrt{41}$. *For* $\mathring{\mathbf{x}} \in \mathbb{C}^d$, *and* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ *with* $\|\mathbf{e}\|_2 \leq \eta$, *denote by* $\mathbf{x}^\star$ *the solution of the quadratically constrained basis pursuit problem*

$$\begin{aligned} \textit{minimize} \quad & \|\mathbf{x}\|_1 \\ \textit{s.t.} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \eta. \end{aligned} \tag{QCBP}$$

Then

$$\begin{aligned} \left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_1 &\leq C\sigma_k(\mathring{\mathbf{x}})_1 + D\sqrt{k}\eta, \\ \left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_2 &\leq \frac{C}{\sqrt{k}}\sigma_k(\mathring{\mathbf{x}})_1 + D\eta, \end{aligned}$$

where $C, D > 0$ *depend only on* δ_{2k}.

This result is both stable w. r. t. sparsity defect and robust against additive noise as the error bounds only depend on the model mismatch quantified by the best k-term approximation error of $\mathring{\mathbf{x}}$, as well as on the extrinsic noise level η. In case of exact k-sparsity of $\mathring{\mathbf{x}}$, and in the absence of measurement noise, Theorem 9 immediately implies perfect recovery.

6.3 Mutual Coherence

Despite the fact that both NSP and RIP allow for the derivation of very strong results in terms of stability and robustness of general recovery algorithms, checking either of them in practice remains an NP-hard decision problem [95]. One alternative property of a measurement matrix $\mathbf{A}$ that can easily be checked in practice is the so-called *mutual coherence*.

Definition 12 Let $\mathbf{A} \in \mathbb{C}^{m \times d}$. Then the mutual coherence $\mu = \mu(\mathbf{A})$ is defined as

$$\mu(\mathbf{A}) := \max_{1 \leq i \neq j \leq d} \frac{|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|}{\|\mathbf{a}_i\|_2 \left\|\mathbf{a}_j\right\|_2},$$

where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$. Assuming ℓ_2-normalized columns of $\mathbf{A}$, this corresponds to the largest off-diagonal element in absolute value of the Gramian $\mathbf{A}^*\mathbf{A}$ of $\mathbf{A}$.

The following proposition presents a fundamental limit on the mutual coherence of a matrix known as the *Welch bound*.

Proposition 4 ([101]) *The coherence of a matrix* $\mathbf{A} \in \mathbb{C}^{m \times d}$ *with* ℓ_2*-normalized columns satisfies*

$$\mu(\mathbf{A}) \geq \sqrt{\frac{d-m}{m(d-1)}}.$$

The equality is attained for every matrix whose columns form an equiangular tight frame.

Unfortunately, coherence-based analyses are rather pessimistic in terms of the number of measurements required to establish robust and stable recovery guarantees. In fact, it can be shown that conditions for perfect recovery in terms of the mutual coherence dictate a quadratic scaling $m = \boldsymbol{\Omega}(k^2)$ of the number of measurements [96], which is only of interest in practice at low sparsity levels.

6.4 Quotient Property

One drawback of the quadratically constrained basis pursuit (QCBP) problem (QCBP) is the fact that one has to have access to an estimate of the noise parameter $\eta \geq \|\mathbf{e}\|_2$, which is often not available in practice. Surprisingly, it can be shown, however, that under an additional condition on the measurement matrix stable and robust recovery of compressible vectors is still possible without any prior knowledge of $\|\mathbf{e}\|_2 \in \mathbb{C}^m$ by means of solving the equality-constrained basis pursuit problem. This condition is given in the form of the so-called *quotient property* of $\mathbf{A}$.

Definition 13 A matrix $\mathbf{A} \in \mathbb{C}^{m \times d}$ is said to satisfy the ℓ_1-quotient property with constant ν if for any $\mathbf{e} \in \mathbb{C}^m$ there exists a vector $\mathbf{u} \in \mathbb{C}^d$ such that

$$\mathbf{e} = \mathbf{A}\mathbf{u} \quad \text{with} \quad \|\mathbf{u}\|_1 \leq \nu\sqrt{k_*}\,\|\mathbf{e}\|_2\,,$$

where $k_* := m/\log(ed/m)$.

If a matrix satisfies both the robust null space property and the quotient property, this allows one to establish the following remarkable result.

Theorem 10 ([54, Theorem 11.12]) *Let* $\mathbf{A} \in \mathbb{C}^{m \times d}$ *be a matrix satisfying the* ℓ_2*-robust null space property as in Definition 10, as well as the* ℓ_1*-quotient property as in Definition 13. Let further* $\mathring{\mathbf{x}} \in \mathbb{C}^d$, $\mathbf{e} \in \mathbb{C}^m$, *and denote by* $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ *the noisy linear measurements of* $\mathring{\mathbf{x}}$*. Then the solution* $\mathbf{x}^\star$ *of the basis pursuit problem (*BP*) satisfies for* $k \leq ck_*$*,*

$$\left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_2 \leq \frac{C_1}{\sqrt{k}}\sigma_k(\mathring{\mathbf{x}})_1 + C_2\,\|\mathbf{e}\|\,,$$

where $\|\cdot\|$ denotes the norm assumed in the ℓ_2-robust null space property. The constants C_1 and C_2 only depend on ρ, τ, c, and ν, i.e., the parameters of the null space and quotient property, respectively.

In the next section, we will address the construction of random measurement matrices which, with high probability, satisfy either the restricted isometry property and/or null space property, respectively. Note that similar probabilistic results can also be shown to hold for the quotient property as introduced above. However, we skip the discussion of this topic for brevity and refer interested readers to [54, Sect. 11.3] instead.

7 Probabilistic Constructions of Measurement Matrices

In this section, we present a series of results which establish the existence of suitable measurement matrices for compressed sensing in the sense that they satisfy the restricted isometry property and consequently the null space property with high probability.

7.1 Restricted Isometries

The first remarkable result we look at in this section concerns the class of subgaussian ensembles which encompasses many important instances of random measurement matrices such as Gaussian and Bernoulli matrices, as well as any matrix populated with independent copies of bounded random variables.

Theorem 11 (Subgaussian restricted isometries, [64, Theorem C.1]) *Let the rows of the $m \times d$ matrix $\mathbf{A}$ be distributed according to an independent isotropic subgaussian distribution. Then the matrix $\frac{1}{\sqrt{m}}\mathbf{A}$ satisfies the restricted isometry property of order k with constant $\delta_k \leq \delta$ if*

$$m \geq C\delta^{-2}k \log\left(\frac{ed}{k}\right)$$

with probability at least $1 - 2\exp(-\delta^2 m/C)$ where the constant C only depends on the subgaussian norm of the rows of $\mathbf{A}$.

A similar theorem can be stated for the case where the columns instead of rows of $\mathbf{A}$ follow a subgaussian distribution. Due to the isotropy assumption of the distribution, the random matrix $m^{-1/2}\mathbf{A}$ acts as an isometry in expectation as we would expect from an RIP matrix, i.e., $\mathbb{E}\|m^{-1/2}\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$. The exponential decay of the failure probability in the above theorem therefore indicates that $\|m^{-1/2}\mathbf{A}\mathbf{x}\|_2^2$ concentrates sharply around its mean $\|\mathbf{x}\|_2^2$ as intended for $\mathbf{A}$ to behave like an isometry.

The original proof of the restricted isometry property for Gaussian random matrices goes back to the work of Candès and Tao [27, 28]. As hinted at above, the restricted isometry property is usually established by means of concentration inequalities that control the deviation of $m^{-1/2}\mathbf{A}$ from its mean. In particular, such concentration results are usually based on Bernstein's inequality for subexponential random variables. In the case of Gaussian random matrices, one can appeal to slightly simpler methods that characterize the smallest and largest singular values of the Gaussian random matrices to establish the RIP in that way.

Another possible proof strategy is based on a result due to Gordon which bounds the expected minimum and maximum gain of a Gaussian random matrix acting on subsets of the sphere ([57, Corollary 1.2]). This result also lies at the heart of the proof of Gordon's escape theorem. Combined with Gaussian concentration of measure, and a simple bound on the mean width of the set of sparse vectors restricted to the unit sphere (see, for instance, [79, Lemma 2.3]), these arguments admit a simple concentration bound which implies the restricted isometry property.

Yet another proof of the restricted isometry property for Gaussian matrices is based on the famous Johnson–Lindenstrauss (JL) lemma [62] (see also [36]). Given a finite collection of points $P := \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{R}^d$, and a random matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ populated with independent zero-mean Gaussian random variables with standard deviation $1/\sqrt{m}$, the JL lemma establishes a bound on the probability that the pairwise distances between the projected points $\mathbf{A}P$ and P deviate at most by a factor of $\pm\epsilon$. A matrix $\mathbf{A}$ that satisfies the property

$$(1-\epsilon)\left\|\mathbf{x}-\mathbf{y}\right\|_2 \le \left\|\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{y}\right\|_2 \le (1+\epsilon)\left\|\mathbf{x}-\mathbf{y}\right\|_2 \quad \forall \mathbf{x}, \mathbf{y} \in P$$

is therefore called a Johnson–Lindenstrauss embedding of P. Note that while this property looks very similar to the definition of the restricted isometry property, it only holds for finite point sets. The JL lemma now asserts that the dimension m of the space has to be at least $m \gtrsim \log(N)$ for the above property to hold with high probability. In [8], this result was used in combination with a covering argument for the set of sparse vectors to provide an alternative RIP proof.

The statement of Theorem 11 depends on a yet unspecified constant C that effects the number of measurements required for a matrix to be an RIP matrix. For Gaussian matrices, the constant can be explicitly characterized (see [54, Chap. 9]). For example, in the asymptotic regime when $d/k \to \infty$, the RIP constant $\delta_{2k} \le 0.6129$ can be achieved with probability at least $1-\epsilon$ if

$$m \ge 54.868\left(k \log\left(\frac{ed}{2k}\right) + \frac{1}{2}\log(2\epsilon^{-1})\right). \tag{17}$$

Finally, it can be shown using tight bounds on the Gelfand widths of ℓ_1-balls that this bound on m is in fact optimal up to a constant [53, 67].

7.1.1 Bounded Orthonormal Systems

The random matrices discussed so far did not possess any discernible structure. However, in many domains of engineering, this assumption would be quite restrictive as the type of measurement matrix is often in part dictated by the specific application, be it due to the particular structure of the problem or for computational purposes. A typical example is structured random matrices involving the DFT or the Hadamard transform. In such situations, we may aim to exploit the existence of highly efficient numerical implementations such as fast Fourier transform (FFT) routines which might prevent us from incorporating a mixing stage involving random matrices into the acquisition system. Moreover, if fast implementations of the measurement operator are available, we can often exploit the operator in the decoding stage to drastically improve the efficiency of the employed recovery procedure. A canonical example of where structured random matrices emerge is when a band-limited function is to be constructed from random time-domain samples. In this case, we consider functions of the form

$$f(t) = \sum_{i=1}^{d} x_i \phi_i(t), \tag{18}$$

where $t \in \mathcal{D} \subset \mathbb{R}$ and the collection $\{\phi_i\}_i$ of functions from $\mathcal{D}$ to $\mathbb{C}$ forms a bounded orthonormal system (BOS) according to the following definition.[13]

Definition 14 (*Bounded orthonormal systems*) A collection of complex-valued functions $\{\phi_i\}_{i=1}^{d}$ defined on a set $\mathcal{D} \subset \mathbb{R}$ equipped with a probability measure μ is called a bounded orthonormal system with constant K if

$$\int_{\mathcal{D}} \phi_i(t)\phi_j(t)\mathrm{d}\mu(t) = \delta_{i,j}$$

and

$$\|\phi_i\|_\infty := \sup_{t \in \mathcal{D}} |\phi_i(t)| \leq K \; \forall i \in [d].$$

Let f be a function with a basis expansion as in Eq. (18) w. r. t. a bounded orthonormal system defined by the collection $\{\phi_i\}_i$. If we sample f at m points $t_1, \ldots, t_m \in \mathcal{D}$, we obtain the system of equations

$$y_j := f(t_j) = \sum_{i=1}^{d} x_i \phi_i(t_j), \quad j \in [m].$$

[13] The definition can easily be extended to the case where $\mathcal{D} \subset \mathbb{R}^n$, but we restrict our discussion to the scalar case here.

Collecting the samples $\{\phi_i(t_j)\}_j$ of the ith basis function in a vector $\boldsymbol{\phi}_i = (\phi_i(t_1), \ldots, \phi_i(t_m))^\top$ forming a column of the matrix $\mathbf{A} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_d]$ of size $m \times d$, we immediately obtain the familiar form

$$\mathbf{y} = \mathbf{Ax},$$

where $\mathbf{y} = (y_1, \ldots, y_m)^\top$ and $\mathbf{x} = (x_1, \ldots, x_d)^\top$. As usual, we assume that $\mathbf{x}$ is sparse or compressible. In this case, the same recovery guarantees w.r.t. to the equality- or quadratically constrained basis pursuit problem can be established as soon as $\mathbf{A}$ or a scaled version of $\mathbf{A}$ can be shown to satisfy the restricted isometry property as before.

The reason why we endow $\mathcal{D}$ with a probability measure is of course that it allows us to draw the sampling points t_j from μ at random to establish the restricted isometry property of matrices defined w.r.t. subsampled bounded orthonormal systems probabilistically. Such results were first demonstrated in [28] for the case of the partial random Fourier matrix which satisfies the restricted isometry property with high probability provided we record $\boldsymbol{\Omega}(k \log^6(d))$ measurements. A nonuniform version of this result, which reduced the power of the log-term from 6 to 4, was shortly after proven by Rudelson and Vershynin in [89]. Another improvement was recently presented in [61] where the required number of measurements was further reduced to $\boldsymbol{\Omega}(k \log^2(k) \log(d))$ for randomly subsampled Fourier matrices. Under certain conditions, this bound can further be reduced. For instance, if the dimension d is an integer multiple of the sparsity level k, Bandeira et al. managed to remove the second log-factor in the previous bound, proving that $\boldsymbol{\Omega}(k \log(d))$ measurements suffice to establish the restricted isometry property for partial Fourier matrices [7]. In case the measurement matrix corresponds to a subsampled Hadamard matrix, Bourgain demonstrated in [17] the sufficiency of $\boldsymbol{\Omega}(k \log(k) \log^2(d))$ measurements to establish the restricted isometry property. A similar bound had previously been shown to hold by Nelson et al. in [74]. The best general bound to date asserts that $m = \boldsymbol{\Omega}(k \log^3(k) \log(d))$ measurements are required to establish the restricted isometry property for arbitrary subsampled bounded orthonormal systems where the sampling points are drawn from a discrete measure [32, Theorem 4.6]. This includes all measurement matrices formed by randomly selecting rows of a unitary matrix such as the DCT or DFT matrix, a Hadamard matrix, etc.

The following theorem records a modern general version of the RIP characterization for measurement matrices based on randomly subsampled bounded orthonormal systems.

Theorem 12 (BOS-RIP, [87, Theorem 4]) *Consider a set of complex-valued bounded orthonormal basis functions* $\{\phi_j\}_{j=1}^d$ *defined on a measure space* $\mathcal{D} \subset \mathbb{R}$ *equipped with the probability measure* μ. *Define a matrix* $\mathbf{A} \in \mathbb{C}^{m \times d}$ *with entries*

$$a_{ij} := \phi_j(t_i), \quad i \in [m],\ j \in [d],$$

constructed by independently drawing the sampling points t_i *from the measure* μ. *Then with probability at least* $1 - d^{-c \log^3(k)}$, *the matrix* $\frac{1}{\sqrt{m}}\mathbf{A}$ *is an RIP matrix of order* k *with constant* $\delta_k \leq \delta$ *provided*

$$m \geq C\delta^{-2}K^2 k \log^3(k) \log(d).$$

The positive constants C *and* c *are universal.*

For the existing bounds, the number of necessary measurements m scales with K^2. For the bound on m in Theorem 12 to be meaningful, the constant K should therefore either be independent of the dimension d or at least only scale with lower powers of d.

Finally, let us highlight that results as stated above can be extended to even more restrictive structured random matrices [6, 85]. For instance, the authors of [64] applied a novel technique to bound the suprema of chaos processes to obtain conditions under which random partial circulant matrices would satisfy the RIP. In this situation, the measurement procedure is of the form

$$\mathbf{Ax} = \frac{1}{\sqrt{m}}\mathbf{R}_{\boldsymbol{\Omega}}(\boldsymbol{\epsilon} * \mathbf{x}),$$

where $\mathbf{R}_{\boldsymbol{\Omega}} : \mathbb{C}^d \to \mathbb{C}^m$ denotes the operator restricting the entries of a vector to the set $\boldsymbol{\Omega} \subset [d]$ of cardinality m, $\boldsymbol{\epsilon}$ is a Rademacher vector of length d, and $*$ denotes the circular convolution operator. In general, if $m \geq C\delta^{-2}k \log^2(k) \log^2(d)$, then with probability at least $1 - d^{-\log(d) \log^2(k)}$ the partial random circulant matrix $\mathbf{A}$ satisfies the RIP of order k with constant $\delta_k \leq \delta$.

7.2 Random Matrices and the Null Space Property

While probabilistic constructions of RIP matrices have been established for a variety of random ensembles such as subgaussian distributions, as well as measurement matrices defined by randomly subsampled basis functions of bounded orthonormal systems as discussed in the previous section, there are some shortcomings to RIP-based recovery guarantees. For instance, the leading constants involved in the required scaling for Gaussian matrices to satisfy the RIP are often quite large. While these constants are usually due to artifacts of the proof strategy, analyses which establish stable and robust recovery by directly appealing to the null space property for Gaussian matrices often have much nicer constants. For instance, for large d and d/k with moderately large k, establishing the null space property requires $m \geq 8k \log(ed/k)$ measurements (cf. [54, Theorem 9.29]) which is much smaller than the constant involved in Eq. (17).

Another shortcoming in RIP-based analyses becomes evident when one tries to obtain recovery guarantees of the form

$$\left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_q \leq C_{k,p}\sigma_k(\mathring{\mathbf{x}})_1 + D_k \left\|\mathbf{e}\right\|_p ,$$

where we aim to characterize the reconstruction performance in the presence of ℓ_p-bounded measurement noise for cases other than $(p, q) \in \{1, 2\}^2$. Note that we still measure the sparsity mismatch in terms of best k-term approximation error w. r. t. the ℓ_1-norm.[14] Such guarantees based on restricted isometries require a generalization of the restricted isometry property as stated in Definition 11. In particular, if one is interested in the recovery of a vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ from compressive measurements of the form $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ with $\left\|\mathbf{e}\right\|_p \leq \varepsilon$, we may solve the program

$$\begin{aligned} \text{minimize } & \left\|\mathbf{x}\right\|_1 \\ \text{s.t. } \quad & \left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_p \leq \varepsilon. \end{aligned}$$

In order to characterize the reconstruction quality of a minimizer $\mathbf{x}^\star$ of this program, one may turn to the mixed (ℓ_p, ℓ_q)-RIP of the form

$$c\left\|\mathbf{x}\right\|_q \leq \left\|\mathbf{A}\mathbf{x}\right\|_p \leq C\left\|\mathbf{x}\right\|_q \qquad \forall \mathbf{x} \in \Sigma_k.$$

However, as was recently addressed in [39], the best known probability bounds to establish the existence of such matrices for $p \neq 1, 2$ exhibit significantly worse scaling in the number of required measurements than $k\log(d/k)$. In their work, Dirksen et al. therefore derive concentration results which instead establish the ℓ_q-robust null space property (Definition 10), providing near-optimal scaling behavior of m (up to possible log-factors) [39] for more general heavy-tailed random matrices. In other words, they demonstrate that recovery guarantees as outlined above, which require similar scaling compared to the provably optimal regime in the case of the (ℓ_2, ℓ_2)-RIP, are not in general outside the realm of possibility. However, their work demonstrates that one may have to move away from RIP-type conditions, and consider stronger concepts such as the null space property and its generalizations to establish similar guarantees. Note that to the best of our knowledge, there currently do not exist any results which establish probabilistic bounds that directly assert the null space property of subsampled BOS matrices without first establishing the RIP to imply the null space property.

Finally, we want to point out two examples of measurement ensembles which provably require more than $k\log(d/k)$ measurements to satisfy the RIP but which nevertheless allow for typical recovery guarantees from $k\log(d/k)$ measurements. The first example is random matrices whose rows follow an isotropic log-concave distribution. Such matrices satisfy the canonical restricted isometry property, i.e., the (ℓ_2, ℓ_2)-RIP, only if $m \gtrsim k\log^2(ed/k)$ but provably allow for exact recovery as soon as $m \gtrsim k\log(ed/k)$ [2, 3, 63]. The second example concerns a certain combinatorial construction of sensing matrices based on the adjacency matrix of random left k-regular bipartite graphs with d left and m right vertices [12]. The corresponding

[14] This avoids another issue regarding the so-called instance optimality of pairs $(\mathbf{A}, \Delta)$ where $\Delta \colon \mathbb{C}^m \to \mathbb{C}^d$ denotes an arbitrary reconstruction algorithm (see [54, Chap. 11] for details).

graph is called a lossless expander and its normalized adjacency matrix $\frac{1}{s}\mathbf{A}$ can be shown to provide typical recovery guarantees with probability at least $1 - \eta$ if $s \gtrsim \log(ed/(k\eta))$ and $m \gtrsim k \log(ed/(k\eta))$. However, the matrix $\frac{1}{s}\mathbf{A}$ does not satisfy the (ℓ_2, ℓ_2)-RIP even though it satisfies the (ℓ_1, ℓ_1)-RIP.

8 An Algorithmic Primer

In the remainder of this introduction to compressed sensing, we want to turn our attention to the practical aspects of signal recovery. To that end, we decided to include a whirlwind tour of recovery algorithms that go beyond the scope of the quadratically constrained basis pursuit problem. Note, however, that the selection of algorithms chosen for this survey is not even close to exhaustive, and really only scratches the surface of what the literature holds in store. An informal search on the IEEE Xplore database produces upward of 1600 search results for the query “compressed sensing recovery algorithm.” Naturally, there is no doubt that this list includes a huge volume of work on specialized algorithms which go beyond the simple sparsity case that we will discuss in this section, as well as survey papers and works which simply benchmark the performance of existing algorithms in the context of specific problems. Nevertheless, this informal experiment still demonstrates the incredibly lively research activity in the field of recovery algorithms in compressed sensing and related domains. For that reason, we limit attention to only a handful of some of the most popular methods found in the pertinent literature and leave it up to the reader to inform him or herself beyond the methods surveyed in this section.

In general, there are multiple criteria by which authors have historically grouped different recovery algorithms for compressed sensing. The most generic classification usually considers three (mostly) distinct classes: convex optimization-based formulations,[15] so-called greedy methods, and iterative thresholding algorithms. Another possible classification could be based on the amount of prior knowledge required to run a particular algorithm. The most coarse classification in this regard takes the form of algorithms which require an explicit estimate of the sparsity level, and those which do not. As is the case for most other surveys on CS recovery algorithms, we decided to opt for the former here.

Before moving on to more efficient recovery methods (at least from a run time and computational complexity perspective), we first state some of the most common variants of convex problems one predominantly finds presented in the relevant literature.

[15] We are careful not to call this an algorithm class as optimization programs are technically just descriptions of problems which still require specialized algorithms such as interior-point methods to actually solve them.

8.1 Convex Programming

As usual, we model the measurement process of a perfectly sparse or compressible signal $\mathring{\mathbf{x}} \in \mathbb{C}^d$ via the affine model $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{C}^m$ is a norm-constrained noise term, i.e., $\|\mathbf{e}\|_p \leq \eta$ with $\eta \geq 0$ and $p \geq 1$. If an upper bound, say w. r. t. the ℓ_2-norm, of this error term $\mathbf{e}$ is known, we naturally consider the quadratically constrained basis pursuit problem that we already discussed in Sect. 6.2:

$$\begin{aligned} &\text{minimize} \ \|\mathbf{x}\|_1 \\ &\quad \text{s.t.} \quad \ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \eta. \end{aligned} \tag{QCBP}$$

For $\eta = 0$, this immediately reduces to the original basis pursuit problem.

Even though we already characterized the recovery behavior of this problem when we introduced the restricted isometry property, we state the result here again for completeness. If $\mathring{\mathbf{x}} \in \mathbb{C}^d$ is merely approximately sparse, one obtains the following characterization for minimizers $\mathbf{x}^\star$ of Problem (QCBP): if $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfies the restricted isometry property of order $2k$ with constant $\delta_{2k} < 4/\sqrt{41}$, one has [54, Theorem 6.12]

$$\left\|\mathbf{x}^\star - \mathring{\mathbf{x}}\right\|_2 \leq C_1 k^{-1/2} \sigma_k(\mathring{\mathbf{x}})_1 + C_2 \eta, \tag{19}$$

where $C_1, C_2 > 0$ only depend on δ_{2k}. Clearly, this result implies perfect recovery in the case where we measure strictly k-sparse vectors in a noise-free environment.

For completeness, we also want to briefly highlight a few alternative convex programming formulations closely related to Problem (QCBP). A very common variant of the quadratically constrained basis pursuit program is the following unconstrained problem:

$$\text{minimize} \quad \|\mathbf{x}\|_1 + \lambda \, \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \tag{BPDN}$$

with $\lambda > 0$, often referred to as basis pursuit denoising (BPDN). The BPDN problem is particularly interesting in situations where no sensible estimate for the noise level η is available. In this case, one may instead use the parameter λ to control the trade-off between sparsity and data fidelity. Depending on the type of method used to solve this unconstrained problem, it might be helpful to replace the data penalty term $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2$ with its squared version to remove the differentiability issue. Of course, the nondifferentiability of the objective function of Problem (BPDN) remains unchanged by this step. However, if one employs a splitting-type algorithm where one alternates between optimizing over individual parts of the objective function, considering a squared ℓ_2-penalty enables us to use gradient-based techniques to deal with the smooth part of the problem. We will discuss an example of such an approach in Sect. 8.2.2 where we present a well-known iterative algorithm to solve a particular variation of Problem (BPDN).

Another important formulation is the so-called *least-absolute shrinkage selection operator* (LASSO) which was originally proposed in the context of sparse model selection in statistics:

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_1 \leq \sigma. \end{aligned} \tag{LASSO}$$

Since the ℓ_1-norm generally functions as a sparsity prior, this formulation might be of interest in situations where rather than an estimate of the noise level η we might have access to a suitable estimate of the sparsity level. Recall that for $\mathring{\mathbf{x}} \in \Sigma_k$ we have by the Cauchy–Schwarz inequality that $\|\mathring{\mathbf{x}}\|_1 \leq \sqrt{k}\, \|\mathring{\mathbf{x}}\|_2$. Depending on the application of interest, an upper bound on the energy of the original signal $\mathring{\mathbf{x}}$ might be naturally available so that one may simply choose $\sigma = \sqrt{k}\, \|\mathring{\mathbf{x}}\|_2$.

Finally, the following program is known as the *Dantzig selector*:

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{x}\|_1 \\ \text{s.t.} \quad & \|\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y})\|_\infty \leq \tau. \end{aligned} \tag{DS}$$

The key idea here is to impose a maximum tolerance on the worst-case correlation between the residuum $\mathbf{r} := \mathbf{A}\mathbf{x} - \mathbf{y}$ and the columns $\{\mathbf{a}_i\}_{i=1}^d$ of $\mathbf{A}$. In the extreme case $\tau = 0$, the Dantzig selector reduces to the classic basis pursuit problem since $\ker(\mathbf{A}^*) = \{\mathbf{0}\}$, and thus $\|\mathbf{A}^*(\mathbf{A}\mathbf{x} - \mathbf{y})\|_\infty = 0$ if and only if $\mathbf{x}$ belongs to the affine space $\left\{\mathbf{z} \in \mathbb{C}^d : \mathbf{A}\mathbf{z} = \mathbf{y}\right\}$.

Conveniently, despite their different formulations and use cases, the problems (BPDN), (LASSO), and (DS) all share the same recovery guarantee from Eq. (19) up to nonlinear transformation of the parameters η, λ, and σ [54, Proposition 3.2]. While the Dantzig selector is the odd one out, similar guarantees can still be derived with relative ease. We refer interested readers to [20].

8.2 Thresholding Algorithms

While the recovery guarantees in the literature are usually strongest for convex optimization-based recovery procedures, generic solving algorithms based on interior-point methods [18, Chap. 11] as employed by popular optimization toolboxes like CVX [59] or CVXPY [38], as well as implementations more specialized to the particular nature of ℓ_1-minimization problems such as ℓ_1- MAGIC [19], SPGL1 [97] and YALL1 [105], become less and less practical if problem sizes increase. The class of thresholding algorithms represents an attractive compromise between strong theoretical guarantees and highly efficient and predictable running times.

Thresholding algorithms can generally be further subdivided into so-called *hard* and *soft-thresholding algorithms*. In the following, we present the most popular representatives from each class, namely, iterative hard thresholding (IHT) and hard thresholding pursuit (HTP) for the former, and the iterative soft-thresholding algorithm (ISTA) and the fast iterative soft-thresholding algorithm (FISTA) for the latter. Other popular thresholding-based algorithms include subspace pursuit [35], NESTA [10], and SpaRSA [103].

8.2.1 Hard Thresholding

At the heart of any hard thresholding algorithm lies the so-called *hard thresholding operator* $H_k\colon \mathbb{C}^d \to \Sigma_k$ defined as

$$H_k(\mathbf{x}) := \underset{\mathbf{z}\in\Sigma_k}{\operatorname{argmin}} \|\mathbf{x}-\mathbf{z}\|_p,$$

for $p \geq 1$ which projects an arbitrary d-vector on the set of k-sparse vectors. The value $H_k(\mathbf{x})$ is constructed by identifying the index set $G \subset [d]$ of size $|G| = k$ which supports the largest values of $\mathbf{x}$ (in absolute value), and zeroing out any values supported on $\overline{G}$. In other words, the vector $H_k(\mathbf{x})$ achieves the best k-term approximation error $\sigma_k(\mathbf{x})_p$ for any $p \geq 1$. For convenience, we also define the set-valued operator $L_k\colon \mathbb{C}^d \to 2^{[d]}$ with $L_k := \operatorname{supp} \circ H_k$ yielding the support set of the best k-term approximation of $\mathbf{x} \in \mathbb{C}^d$. Here, 2^G denotes the power set of G.

With these definitions in place, we now turn to the first hard thresholding algorithm.

Iterative Hard Thresholding

The key idea of iterative hard thresholding is to reduce the smooth loss function $g(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2$ with gradient $\nabla g(\mathbf{x}) = \mathbf{A}^*(\mathbf{A}\mathbf{x}-\mathbf{y})$ at every iteration by means of a gradient descent update before pruning the solution to the set of k-sparse vectors by means of the hard thresholding operator. The full listing of the algorithm is given in Algorithm 1.

Algorithm 1: Iterative Hard Thresholding (IHT)

```
Input: $\mathbf{A} \in \mathbb{C}^{m\times d}$, $\mathbf{y} \in \mathbb{C}^m$, $k \in [d]$
Initialization: $\mathbf{x}^0 \leftarrow \mathbf{0}$, $n \leftarrow 0$
while halting condition is not satisfied do
    $\mathbf{v}^{n+1} \leftarrow \mathbf{x}^n - \mathbf{A}^*(\mathbf{A}\mathbf{x}^n - \mathbf{y})$        Gradient descent step
    $\mathbf{x}^{n+1} \leftarrow H_k(\mathbf{v}^{n+1})$        Projection on $\Sigma_k$
    $n \leftarrow n + 1$
end
Output: $\mathbf{x}^n$
```

Considering the nonlinearity of the operator H_k, it is not immediately obvious that Algorithm 1 even converges, let alone to the true solution $\mathring{\mathbf{x}}$. The following result demonstrates both robustness w. r. t. sparsity defect and stability w. r. t. measurement noise. Consider an arbitrary vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ which we measure according to the model $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$. If $\mathbf{A}$ satisfies the RIP condition with constant $\delta_{6k} < 1/\sqrt{3}$, Algorithm 1 produces iterates $(\mathbf{x}^n)_{n\geq 0}$ satisfying [54, Theorem 6.21]

$$\left\|\mathbf{x}^n - \mathring{\mathbf{x}}\right\|_2 \leq 2\rho^n \left\|\mathring{\mathbf{x}}\right\|_2 + C_1 k^{-1/2}\sigma_k(\mathring{\mathbf{x}})_1 + C_2 \|\mathbf{e}\|_2,$$

where $C_1, C_2 > 0$, and $0 < \rho < 1$ are constants which only depend on δ_{6k}. For $n \to \infty$, this sequence converges to a cluster point $\mathbf{x}^\star$ satisfying

$$\left\|\mathbf{x}^\star - \mathring{\mathbf{x}}\right\|_2 \leq C_1 k^{-1/2} \sigma_k(\mathring{\mathbf{x}})_1 + C_2 \left\|\mathbf{e}\right\|_2 . \tag{20}$$

If the vector $\mathring{\mathbf{x}}$ we wish to recover is in reality supported on an index set $S \subset [d]$ of size k, and measurements are not disturbed by noise ($\mathbf{e} = \mathbf{0}$), one has $\sigma_k(\mathring{\mathbf{x}})_1 = 0$, and therefore $\left\|\mathbf{x}^\star - \mathring{\mathbf{x}}\right\|_2 \leq 0$, implying perfect recovery with $\mathbf{x}^\star = \mathring{\mathbf{x}}$.

Hard Thresholding Pursuit

The fundamental difference between IHT and HTP is the fact that HTP merely uses hard thresholded gradient descent updates to estimate the support set of $\mathring{\mathbf{x}}$. In particular, it propagates least-squares solutions of $\mathbf{y} = \mathbf{A}\mathbf{x}$ w. r. t. to a submatrix of $\mathbf{A}$ obtained by pursuing the active support set of coefficients in each iteration based on the operator $L_k = \operatorname{supp} \circ H_k$. A full algorithm listing is given in Algorithm 2.

Surprisingly, the stability and robustness analyses are identical for IHT and HTP barring a change of parameters (C_1, C_2, ρ) for HTP. Most importantly, this change results in a faster rate of convergence for the HTP algorithm [54].

Algorithm 2: Hard Thresholding Pursuit (HTP)

```
Input: $\mathbf{A} \in \mathbb{C}^{m \times d}, \mathbf{y} \in \mathbb{C}^m, k \in [d]$
Initialization: $\mathbf{x}^0 \leftarrow \mathbf{0}, n \leftarrow 0$
while halting condition is not satisfied do
    $\mathbf{v}^{n+1} \leftarrow \mathbf{x}^n - \mathbf{A}^*(\mathbf{A}\mathbf{x}^n - \mathbf{y})$        Gradient descent step
    $G_{n+1} \leftarrow L_k(\mathbf{v}^{n+1})$                  Support identification
    $\mathbf{x}^{n+1} \leftarrow \mathbf{0}$
    $\mathbf{x}^{n+1}_{G_{n+1}} \leftarrow \mathbf{A}^\dagger_{G_{n+1}} \mathbf{y}$               Least-squares update
    $n \leftarrow n + 1$
end
Output: $\mathbf{x}^n$
```

8.2.2 Soft Thresholding

While the algorithms described in Sect. 8.2.1 rely on explicit hard thresholding to guarantee a certain sparsity level of solutions, soft-thresholding methods (also referred to as shrinkage thresholding for reasons which will become clear shortly) promote sparsity by incorporating an ℓ_1-prior in their objective functions, and applying the so-called *proximal gradient algorithm* or a variant thereof. In particular, we aim to solve the unconstrained regularized problem

$$\text{minimize} \;\; \lambda \left\|\mathbf{x}\right\|_1 + \frac{1}{2} \left\|\mathbf{A}\mathbf{x} - \mathbf{y}\right\|_2^2 , \tag{21}$$

with $\lambda > 0$. Up to rescaling of the objective function, and squaring of the ℓ_2-penalty, this is identical to Problem (BPDN) introduced earlier.

To explain the general idea behind soft thresholding, consider a loss function of the form $f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})$ where $g\colon \mathbb{R}^d \cup \{-\infty, \infty\} \to \mathbb{R}$ is a (possibly) nonsmooth lower semi-continuous extended value function and $h\colon \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function. If g were smooth, this problem could be solved by standard optimization tools such as (conjugate) gradient descent or Newton's method. However, in order to promote sparsity one will often choose $g = \lambda \|\cdot\|_1$, meaning that such a simple approach is not applicable. In the proximal gradient method, one therefore replaces the smooth part h of f by means of a second-order approximation, i.e., one considers an iterative approach of the form

$$\mathbf{x}^+ := \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^d} \left\{ g(\mathbf{v}) + \hat{h}_t(\mathbf{x}, \mathbf{v}) \right\},$$

where $\mathbf{x}$ and $\mathbf{x}^+$ denote the current and next iterate, respectively, and

$$\hat{h}_t(\mathbf{x}, \mathbf{v}) := h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{v} - \mathbf{x} \rangle + \frac{1}{2t} \|\mathbf{v} - \mathbf{x}\|_2^2 \tag{22}$$

with $t > 0$ is a second-order approximation of h around the point $\mathbf{x}$. It is easily verified that the expression for $\mathbf{x}^+$ can be rewritten as

$$\begin{aligned} \mathbf{x}^+ &= \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^d} \left\{ g(\mathbf{v}) + h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{v} - \mathbf{x} \rangle + \frac{1}{2t} \|\mathbf{v} - \mathbf{x}\|_2^2 \right\} \\ &= \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^d} \left\{ g(\mathbf{v}) + \frac{1}{2t} \|\mathbf{v} - (\mathbf{x} - t\nabla h(\mathbf{x}))\|_2^2 \right\}. \end{aligned} \tag{23}$$

While this formulation might give the impression that we merely traded one difficult optimization problem for another, it turns out that the operator in Eq. (23) corresponds to the so-called *proximal operator* [77]

$$\operatorname{prox}_{tg}(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^d} \left\{ g(\mathbf{v}) + \frac{1}{2t} \|\mathbf{v} - \mathbf{x}\|_2^2 \right\},$$

applied to the gradient descent update $\mathbf{x} - t\nabla h(\mathbf{x})$. Conveniently, this operator has a closed-form solution for a variety of different nonsmooth functions g. In particular, it is easy to check via subdifferential calculus over its individual entries that $\operatorname{prox}_{\alpha\|\cdot\|_1}(\mathbf{x}) = S_\alpha(\mathbf{x})$ where

$$S_\alpha(x) := \begin{cases} \operatorname{sign}(x)(|x| - \alpha), & |x| \geq \alpha, \\ 0, & \text{otherwise}, \end{cases}$$

is the so-called shrinkage operator that is applied element-wise to $\mathbf{x}$.[16] Overall, we obtain the iteration

$$\mathbf{x}^{+} = S_{\lambda t}(\mathbf{x} - t\mathbf{A}^{\top}(\mathbf{A}\mathbf{x} - \mathbf{y})) \tag{24}$$

if we apply this method to the basis pursuit denoising Problem (21). In this particular formulation, the parameter t acts as a step size which we may choose (e.g.,) via backtracking line search, while $\lambda > 0$ can be used to control the trade-off between sparsity of the solution $\mathbf{x}^{\star}$ and the data fidelity term $\|\mathbf{A}\mathbf{x}^{\star} - \mathbf{y}\|_2$.

This algorithm requires on the order of $\mathcal{O}(1/\epsilon)$ iterations to come within an ϵ-range $|f(\mathring{\mathbf{x}}) - f(\mathbf{x}^n)| \leq \epsilon$ of optimality, implying a convergence rate of $\mathcal{O}(1/n)$ [9]. According to a celebrated result by Nesterov [75], the best achievable convergence rate in the class of nonsmooth first-order methods[17] is $\mathcal{O}(1/n^2)$. This rate is achievable by Nesterov's acceleration method, resulting in the well-known *fast iterative soft-thresholding algorithm* (FISTA) due to Beck and Teboulle when applied to the iterative soft-thresholding algorithm [9]. Informally, the key idea of FISTA is to add a momentum term depending on the last two iterates to avoid erratic changes in the search direction, i.e., one updates the iterates according to

$$\begin{aligned}\mathbf{v}^{n+1} &:= \mathbf{x}^n + \frac{n-2}{n+1}(\mathbf{x}^n - \mathbf{x}^{n-1}),\\ \mathbf{x}^{n+1} &:= S_{t_n}(\mathbf{v}^{n+1} - t_n\mathbf{A}^{\top}(\mathbf{A}\mathbf{x}^n - \mathbf{y}))\end{aligned}$$

with $t_n > 0$ the step size at iteration n. Note that this formulation, taken from [77], differs from the original one given in [9] which explicitly depends on the Lipschitz constant of the gradient of the smooth part of (21). Also note that while this algorithm obtains the desired convergence rate of $\mathcal{O}(1/n^2)$, it is not a descent method. In practice, this means that additional book keeping is required to keep track of the best current iterate. However, considering that this accelerated scheme virtually comes at the same computational cost as Eq. (24), the impact of book keeping is negligible if weighed against the greatly improved convergence behavior.

Both ISTA and FISTA solve the unconstrained problem (21), and provably converge to the global optimum at a linear and super-linear rate, respectively, where convergence without step size adaptation is determined by the Lipschitz constant $L := \|\mathbf{A}^{\top}\mathbf{A}\|_{2\to 2}$ of the gradient of $h(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$. Since our main objective is the recovery of sparse or more generally compressible vectors from noisy measurements, we still have to answer the question how closely these algorithms approximate the true solution $\mathring{\mathbf{x}}$, and under which conditions recovery is exact. Conveniently, these recovery guarantees can be expressed in terms of the guarantees obtained for the quadratically constrained basis pursuit problem stated in Sect. 8.1.

[16] Hence the name *shrinkage* thresholding.

[17] Note that while we used a second-order approximation of h in Eq. (22), we did so by approximating the Hessian $\nabla^2 h(\mathbf{x})$ as a scaled identity matrix, thereby ignoring the true second-order information of h.

This holds because—given a minimizer $\mathbf{x}^\star_{\text{QCBP}}$ of (QCBP)—we can always find a transformation $T(\mathbf{x}^\star_{\text{QCBP}}, \eta) = \lambda$ of the parameter $\eta \geq 0$ of (QCBP) and the parameter $\lambda > 0$ of the unconstrained problem (21) such that both convex problems have the same optimal value $f^\star$ [10]. Note, however, that explicitly finding the mapping T is generally a nontrivial problem [98].

It remains to show when Problem (21) has a unique minimizer such that the correspondence between the solutions $\mathbf{x}^\star_{\text{QCBP}}$ and $\mathbf{x}^\star_{\text{BPDN}}$ is one-to-one given an appropriate choice of parameters η and λ. To that end, one seeks conditions when minimizers of (21) are unique. While there are various publications that address the issue of uniqueness of solutions to this problem, e.g., [24, 94], none of them is immediately guaranteed by the RIP or NSP. For instance, [104, Theorem 4.1] establishes the following condition for minimizers of (21) to be unique.

Theorem 13 *Let* $\mathbf{x}^\star$ *be a minimizer of the basis pursuit denoising problem, and define* $S := \operatorname{supp}(\mathbf{x}^\star)$. *Then* $\mathbf{x}^\star$ *is a unique minimizer iff*

1. $\mathbf{A}_S$ *has full column-rank,*
2. $\exists \mathbf{u} \in \mathbb{R}^m$ *such that* $\mathbf{A}_S^\top \mathbf{u} = \operatorname{sign}(\mathbf{x}^\star_S)$ *and* $\left\| \mathbf{A}_{\bar{S}}^\top \mathbf{y} \right\|_\infty < 1$.

Approximate Message Passing

Due to the structural similarity to the iterative soft-thresholding algorithm, we briefly touch upon another popular development in the field of iterative thresholding algorithms, namely, the so-called *approximate message passing* (AMP) method. Pioneered by Donoho et al. in [44], the general formulation of approximate message passing (AMP) closely resembles the basic form of ISTA. The difference amounts to a correction term of the residuum $\mathbf{r}^n = \mathbf{A}\mathbf{x}^n - \mathbf{y}$ stemming from the interpretation of the measurement model $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}}$ in terms of *loopy belief propagation* in graphical models. Based on a slight reformulation of Eq. (24), approximate message passing proceeds via the iterations

$$\mathbf{x}^{n+1} := S_{\mu_n}(\mathbf{A}^\top \mathbf{r}^n + \mathbf{x}^n), \tag{25}$$

$$\mathbf{r}^n := \mathbf{y} - \mathbf{A}\mathbf{x}^n + \frac{1}{\delta}\mathbf{r}^{n-1}\left\langle \mathbf{1}, S'_{\mu_n}(\mathbf{A}^\top \mathbf{r}^{n-1} + \mathbf{x}^{n-1})\right\rangle, \tag{26}$$

where $\delta := m/d$ and $S'_\mu(x)$ denotes the derivative of $S_\mu(x)$ ignoring the nondifferentiability at $|x| = \mu$. Despite this innocent looking correction term in Eq. (26) (also known as *Onsager correction*), which barely increases the computational complexity over ISTA, the performance of this algorithm in terms of the observed phase-transition diagrams turns out to be highly competitive with the de facto gold standard of ℓ_1-minimization and in certain situations even manages to outperform it [43].

The key ingredient to the success of AMP is the observation that in the large-system limit $m, d \to \infty$ with δ fixed, and $A_{ij} \sim_{\text{i.i.d.}} \mathsf{N}(0, 1/m)$, one has $\mathbf{A}^\top \mathbf{r}^n + \mathbf{x}^n = \mathring{\mathbf{x}} + \mathbf{v}^n$ for the argument of S_{μ_n} in Eq. (25) where $\mathbf{v}^n$ is an i.i.d. zero-mean Gaussian random vector whose variance σ_n^2—and hence the mean squared error (MSE) of the reconstruction—can be predicted by a state evolution formalism.

Since its original introduction, a variety of modifications and improvements have been proposed for the AMP algorithm. These include the denoising-based AMP (D-AMP) [69] which generalizes the state evolution formalism to general Lipschitz continuous denoisers other than the soft-thresholding function, vector AMP (V-AMP) [83] which extends AMP to more general classes of measurement matrices, and generalized AMP (GAMP) [82] which extends AMP to arbitrary input and output distributions and allows for dealing with nonlinearities in the measurement process. While the general versions of most of these AMP variants require some statistical knowledge about the parameters involved, there exist several modifications which estimate these parameters online via expectation maximization (EM).

In closing, we mention that Problem (21) can be tackled by a variety of related methods such as alternating direction method of multipliers (ADMM), forward–backward splitting, Douglas–Rachford splitting, or homotopy methods. We refer the interested reader to the excellent survey [50], as well as to the notes in [54, Chap. 15].

8.3 Greedy Methods

Greedy algorithms are generally characterized by their tendency to act according to locally optimal decision rules in hopes of eventually arriving at a global optimal solution. In particular, they never explicitly aim at minimizing a particular (non-)convex objective. Instead, they treat the collection of columns of the measurement matrix $\mathbf{A}$ as a dictionary of atoms $\{\mathbf{a}_i\}_{i=1}^d$ and first try to identify the atoms which likely contributed to the measurement vector $\mathbf{y}$, before estimating the associated weighting factors. Despite the fact that algorithms of this type had been in use long before the advent of compressed sensing, particularly in the image processing community, research into greedy algorithms for sparse recovery experienced a resurgence ever since the rise of compressed sensing. In this section, we will look at two of the most popular representatives in this particular class of algorithms, namely, the so-called orthogonal matching pursuit and compressive sampling matching pursuit methods.

Orthogonal Matching Pursuit

While technically a successor to the lesser used matching pursuit algorithm, orthogonal matching pursuit (OMP) remains to this day one of the most popular greedy algorithms due to the fact that it is one of the methods with the lowest footprint in terms of computational complexity. As can be seen from Algorithm 3, OMP updates its estimated support set one atom at a time by identifying the atom $\mathbf{a}_i$ that exhibits the strongest correlation with the residuum $\mathbf{r}^n = \mathbf{A}\mathbf{x}^n - \mathbf{y}$ as measured by the inner product $|\langle \mathbf{a}_i, \mathbf{r}^n \rangle|$.

The atom selection step in each OMP iteration can be interpreted as identifying the component of $\mathbf{x}^n$ w.r.t. which the function $f(\mathbf{x}^n) := \frac{1}{2} \|\mathbf{A}\mathbf{x}^n - \mathbf{y}\|_2^2$ varies the most. This is due to the fact that the gradient of f at $\mathbf{x}^n$ reads $\nabla(\frac{1}{2} \|\mathbf{A}\mathbf{x}^n - \mathbf{y}\|_2^2) = \mathbf{A}^*(\mathbf{A}\mathbf{x}^n - \mathbf{y}) = \mathbf{A}^*\mathbf{r}^n$. The update step $\mathbf{x}^n \to \mathbf{x}^{n+1}$ on the other hand corresponds

Algorithm 3: Orthogonal Matching Pursuit (OMP)

```
Input: $\mathbf{A} \in \mathbb{C}^{m \times d}$, $\mathbf{y} \in \mathbb{C}^m$, $k \in [d]$
Initialization: $\mathbf{x}^0 \leftarrow \mathbf{0}$, $G^0 \leftarrow \emptyset$, $n \leftarrow 0$, $\mathbf{r}^0 \leftarrow -\mathbf{A}^*\mathbf{y}$
while halting condition is not satisfied do
    $j_{n+1} \leftarrow \operatorname{argmin}_{j \in [d]} |(\mathbf{A}^*\mathbf{r}^n)_j|$        Atom identification
    $G_{n+1} \leftarrow G_n \cup \{j_{n+1}\}$        Support extension
    $\mathbf{x}^{n+1} \leftarrow \mathbf{A}^\dagger_{G_{n+1}}\mathbf{y}$        Least-squares projection
    $\mathbf{r}^{n+1} \leftarrow \mathbf{A}\mathbf{x}^{n+1} - \mathbf{y}$        Calculation of residuum
    $n \leftarrow n + 1$
end
Output: $\mathbf{x}^n$
```

to a projection of $\mathbf{y}$ on the subspace spanned by the columns of $\mathbf{A}$ indexed by the updated index set G_{n+1}.

While theoretical guarantees in the noise-free and exactly sparse case exist in abundance for OMP, robust and stable recovery guarantees are not as well-developed as one might expect given the maturity of the theory and the popularity of OMP in general. Oftentimes such results depend on additional regularity conditions on the class of vectors one aims to recover.

In general, OMP does not require an estimate of the sparsity level of the vector one aims to recover. The algorithm naturally terminates as soon as the same atom is selected twice in subsequent iterations. Other halting conditions include the relative change of estimates $\mathbf{x}^n$ between iterations and tolerance criteria of data fidelity measures w. r. t. $\mathbf{r}^n$. Considering that OMP updates the support set one index at a time per iteration, OMP requires at least k iterations to find a k-sparse candidate vector. If the sparsity level is known a priori, another natural termination condition is therefore simply given by the number of iterations.

One of the earliest recovery guarantees for OMP was the coherence-based condition $(2k - 1)\mu < 1$ which allows OMP to recover any k-sparse vector from noiseless linear measurements in k iterations [42]. In light of the Welch bound (cf. Proposition 4)

$$\mu \geq \sqrt{\frac{d-m}{m(d-1)}},$$

this implies the quadratic scaling in the number of measurements announced in Sect. 6.3. Currently, one of the best known sufficiency conditions for exact k-sparse recovery in the noiseless setting in terms of the restricted isometry property requires $\delta_{k+1} < 1/\sqrt{k+1}$ [71, Theorem III.1].

In the general noise-corrupted setting with $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$, one obtains the RIP-based bound [54, Theorem 6.25]

$$\left\|\mathbf{x}^{24k} - \mathring{\mathbf{x}}\right\|_2 \leq C_1 k^{-1/2} \sigma_k(\mathring{\mathbf{x}})_1 + C_2 \|\mathbf{e}\|_2 \tag{27}$$

for iterates of OMP after $24k$ iterations, where the constants $C_1, C_2 > 0$ only depend on the RIP constant $\delta_{26k} < 1/6$ of the associated measurement matrix $\mathbf{A}$. In the noiseless and exactly sparse case, Eq. (27) guarantees perfect recovery after $24k$ iterations. Note, however, that in this case OMP will already reach the global optimum after k iterations since the algorithm selects one atom per iteration, after which it will stall due to the fact that $\mathbf{r}^n = \mathbf{0}$ for $n > k$. Otherwise, the solution returned by OMP after $24k$ iterations could not be k-sparse.

These guarantees are a far cry from the recovery conditions one obtains for methods such as QCBP or IHT seeing how RIP matrices of order $26k$ are much harder to construct than matrices of order $2k$ and $3k$, respectively. One possible explanation for the demanding requirement on the RIP order of $\mathbf{A}$ is the fact that OMP in its presented form has no way to correct possibly erroneous choices of atoms made in previous iterations. In a sense, this observation can be seen as one of the main motivations of the compressive sampling matching pursuit algorithm we will introduce in the next section.

Compressive Sampling Matching Pursuit

The compressive sampling matching pursuit (CoSaMP) algorithm shares a lot of similarities both with the OMP algorithm and the hard thresholding pursuit algorithm described in Sect. 8.2. While technically also an iterative algorithm that relies on hard thresholding, it is usually considered an instance of the class of greedy algorithms. The full procedure is given in Algorithm 4. Given a current estimate $\mathbf{x}^n$ of $\mathring{\mathbf{x}}$, CoSaMP proceeds by first identifying the $2k$ columns of $\mathbf{A}$ which best correlate with the residuum $\mathbf{r}^n = \mathbf{A}\mathbf{x}^n - \mathbf{y}$ at iteration n. The algorithm then continues to solve a least-squares problem w. r. t. to column submatrix defined by the support of $\mathbf{x}^n$ and the $2k$ column indices identified in the previous step. Since the algorithm ultimately aims to obtain strictly k-sparse solutions, the next estimate $\mathbf{x}^{n+1}$ is finally found via hard thresholding of the least-squares update $\mathbf{v}^{n+1}$.

Algorithm 4: Compressive Sampling Matching Pursuit (CoSaMP)

```
Input: $\mathbf{A} \in \mathbb{C}^{m\times d}$, $\mathbf{y} \in \mathbb{C}^m$, $k \in [d]$
Initialization $\mathbf{x}^0 \leftarrow \mathbf{0}$, $n \leftarrow 0$, $\mathbf{r}^0 \leftarrow -\mathbf{A}^*\mathbf{y}$
while halting condition is not satisfied do
    $G_{n+1} \leftarrow \mathrm{supp}(\mathbf{x}^n) \cup L_{2k}(\mathbf{A}^*\mathbf{r}^n)$          Support overestimation
    $\mathbf{v}^{n+1} \leftarrow \mathbf{0}$
    $\mathbf{v}^{n+1}_{G_{n+1}} \leftarrow \mathbf{A}^{\dagger}_{G_{n+1}}\mathbf{y}$                  Least-squares projection
    $\mathbf{x}^{n+1} \leftarrow H_k(\mathbf{v}^{n+1})$                        "Projection" on $\Sigma_k$
    $\mathbf{r}^{n+1} \leftarrow \mathbf{A}\mathbf{x}^{n+1} - \mathbf{y}$                      Calculation of residuum
    $n \leftarrow n + 1$
end
Output: $\mathbf{x}^n$
```

Solving the least-squares problem over a column index set of size at most $3k$ effectively allows CoSaMP to adaptively correct previous choices of the support set of its estimate of $\mathring{\mathbf{x}}$. This is one of the main drawbacks of the OMP algorithm, which will never remove a previously selected atom $\mathbf{a}_i$ from its dictionary once column i of $\mathbf{A}$ was identified as an element contributing to $\mathbf{y}$.

In accordance with the previous algorithms, we once again state available stability and robustness results for CoSaMP. Consider a vector $\mathring{\mathbf{x}} \in \mathbb{C}^d$ which we aim to recover from its linear measurements $\mathbf{y} = \mathbf{A}\mathring{\mathbf{x}} + \mathbf{e}$ where $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfies the RIP of order $8k$ with $\delta_{8k} < 0.4782$. Then the sequence $(\mathbf{x}^n)_{n \geq 0}$ generated by Algorithm 4 satisfies [54, Theorem 6.28]

$$\left\|\mathbf{x}^n - \mathring{\mathbf{x}}\right\|_2 \leq 2\rho^n \left\|\mathring{\mathbf{x}}\right\|_2 + C_1 k^{-1/2} \sigma_k(\mathring{\mathbf{x}})_1 + C_2 \left\|\mathbf{e}\right\|_2, \tag{28}$$

where $C_1, C_2 > 0$, and $0 < \rho < 1$ only depend on δ_{8k}. Once again, Eq. (28) establishes the existence of cluster points $\mathbf{x}^\star$ satisfying

$$\left\|\mathbf{x}^\star - \mathring{\mathbf{x}}\right\|_2 \leq C_1 k^{-1/2} \sigma_k(\mathring{\mathbf{x}})_1 + C_2 \left\|\mathbf{e}\right\|_2,$$

which implies perfect recovery by convergence to the unique vector $\mathring{\mathbf{x}}$ once $\mathring{\mathbf{x}} \in \Sigma_k$ and $\mathbf{e} = \mathbf{0}$.

8.4 Iteratively Reweighted Least-Squares

Another popular method which does not quite fit into any of the categories discussed so far is the so-called iteratively reweighted least-squares (IRLS) algorithm. At its core, IRLS is motivated by the observation that

$$|x| = |x|^{-1}|x|^2$$

for $0 \neq x \in \mathbb{C}$. Assuming for the moment that $\mathring{\mathbf{x}} \in \Sigma_k$ were known, we could rewrite the basis pursuit problem as

$$\min\left\{\sum_{i=1}^{d} |x_i| : \mathbf{y} = \mathbf{A}\mathbf{x}\right\} = \min\left\{\sum_{i \in \operatorname{supp}(\mathring{\mathbf{x}})} |\mathring{x}_i|^{-1}|x_i|^2 : \mathbf{y} = \mathbf{A}\mathbf{x}\right\}. \tag{29}$$

The idea now is to treat the term $|\mathring{x}_i|^{-1}$ as a weighting factor that we iteratively update in an alternating fashion in between updates of the variables x_i. To that end, we define the weighting factors as a smooth approximation

$$w_i^{n+1} := |x_i^2 + \tau_{n+1}^2|^{-1/2}, \tag{30}$$

where we require $0 < \tau_{n+1} \leq \tau_n$ so that $w_i^{n+1} \to |x_i|^{-1}$ as $\tau_{n+1} \to 0$. Considering that $\operatorname{supp}(\mathring{\mathbf{x}})$ is unknown, this approximation has the added advantage that we can let the summation on the right-hand side of Eq. (29) run through all indices in $[d]$ as the regularization parameter τ_n avoids divisions by zero. To proceed, we now define the functional

$$\mathcal{F}(\mathbf{x}, \mathbf{w}, \tau) := \frac{1}{2}\left[\sum_{i=1}^{d} |x_i|^2 w_i + \sum_{i=1}^{d} (\tau^2 w_i + w_i^{-1})\right]. \tag{31}$$

This definition is motivated by the following observations. Given a fixed weight vector $\mathbf{w}$ and regularizer τ, Eq. (31) corresponds to Eq. (29) with $|\mathring{x}_i|^{-1}$ replaced by w_i. Defining $\mathbf{D}_\mathbf{w} := \operatorname{diag}\{\mathbf{w}\}$, this constitutes a least-squares minimization problem w. r. t. the induced norm $\|\mathbf{x}\|_{\mathbf{D}_\mathbf{w}} := \sqrt{\mathbf{x}^* \mathbf{D}_\mathbf{w} \mathbf{x}}$, i.e.,

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{x}\|_{\mathbf{D}_\mathbf{w}} \\ \text{s.t.} \quad & \mathbf{y} = \mathbf{A}\mathbf{x}, \end{aligned}$$

which admits the closed-form solution

$$\mathbf{x}^\star = \mathbf{D}_\mathbf{w}^{-1/2} (\mathbf{A}\mathbf{D}_\mathbf{w}^{-1/2})^\dagger \mathbf{y}.$$

The second observation concerns the update of the weighting vector $\mathbf{w}$ given a fixed $\mathbf{x}$ and τ. In that case, it is easily verified for $i \in [d]$ that

$$w_i^\star = \operatorname*{argmin}_{w_i > 0} \mathcal{F}(\mathbf{x}, \mathbf{w}, \tau) = \frac{1}{\sqrt{|x_i|^2 + \tau^2}},$$

which corresponds to the regularization of w_i in terms of x_i and τ as motivated by Eq. (30). The full algorithm is listed in Algorithm 5. Note that the update rule for τ is chosen in such a way that τ_n is a nonincreasing sequence in n as motivated above.

The following recovery guarantee for the IRLS algorithm is based on [54, Theorem 15.15]. Let $\mathbf{A} \in \mathbb{C}^{m \times d}$ satisfy the restricted isometry property of order $2k$ with $\delta_{2k} < 7/(4\sqrt{41}) \approx 0.2733$, and define[18] for $\alpha_\delta := \sqrt{1 - \delta_{2k}^2} - \delta_{2k}/4$,

$$\rho := \frac{\delta_{2k}}{\alpha_\delta} \quad \text{and} \quad \tau := \frac{\sqrt{1 + \delta_{2k}}}{\alpha_\delta}.$$

Then the sequence $(\mathbf{x}^n)_{n \geq 0}$ generated by the IRLS algorithm converges to a point $\mathbf{x}^\star$, and

[18] Note that this choice amounts to $\mathbf{A}$ satisfying the ℓ_2-robust null space property (cf. Definition 10) of order k with constants $\rho < 1/3$ and $\tau > 0$ [54, Theorem 6.13].

$$\left\|\mathring{\mathbf{x}} - \mathbf{x}^\star\right\|_1 \leq \frac{2(3+\rho)}{1-3\rho}\sigma_k(\mathring{\mathbf{x}})_1$$

which implies perfect recovery via the IRLS algorithm if $\mathring{\mathbf{x}}$ is k-sparse.

Algorithm 5: Iteratively Reweighted Least-Squares (IRLS)

```
Input: $\mathbf{A} \in \mathbb{C}^{m\times d}, \mathbf{y} \in \mathbb{C}^m, k \in [d]$
Initialization: $\mathbf{w}^0 \leftarrow \mathbf{1}, n \leftarrow 0, \tau_0 \leftarrow 1$
while halting condition is not satisfied do
    $\mathbf{x}^{n+1} \leftarrow \mathbf{D}_{\mathbf{w}^n}^{-1/2}(\mathbf{A}\mathbf{D}_{\mathbf{w}^n}^{-1/2})^\dagger \mathbf{y}$
    $\tau_{n+1} \leftarrow \min\left\{\tau_n, (\mathbf{x}^n)^*_{k+1}/(2d)\right\}$
    $w_i^{n+1} \leftarrow \left(|x_i^{n+1}|^2 + \tau_{n+1}^2\right)^{-1/2} \quad \forall i \in [d]$
    $n \leftarrow n+1$
end
Output: $\mathbf{x}^n$
```

9 Conclusion

In the years since its inception, the field of compressed sensing has steadily developed into a mature theory at the intersection of applied mathematics and engineering. With numerous applications in various domains of science and engineering, it now constitutes an indispensable tool in the toolbox of signal processing engineers who are faced with the problem of sampling high-dimensional signals in resource-constrained environments.

In this chapter, we reviewed some of the basic concepts of the theory, focusing on large part on the problem of nonuniform recovery of low-complexity signals from linear observations. In particular, we want to highlight the inclusion of a discussion on the connection between sparse recovery and conic integral geometry, a rather young development in the field, as well as a broader discussion of several efficient recovery algorithms and associated performance guarantees. We hope that the selection of topics featured in this introduction serves as a useful starting point in the further study of the theory of compressed sensing and its extensions.

Acknowledgements We would like to thank the anonymous reviewers and contributors to this book for their invaluable comments regarding this introduction.

References

1. S.I. Adalbjörnsson, A. Jakobsson, M.G. Christensen. Estimating multiple pitches using block sparsity, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (May 2013), pp. 6220–6224
2. R. Adamczak, R. Latała, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann, Geometry of log-concave ensembles of random matrices and approximate reconstruction. C. R. Math. **349**(13), 783–786 (2011)
3. R. Adamczak, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann, Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. Constr. Approx. **34**(1), 61–88 (2011)
4. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: phase transitions in convex programs with random data. Inf. Inference **3**(3), 224–294 (2014)
5. U. Ayaz, S. Dirksen, H. Rauhut, Uniform recovery of fusion frame structured sparse signals. Appl. Comput. Harmon. Anal. **41**(2), 341–361 (2016)
6. W.U. Bajwa, J.D. Haupt, G.M. Raz, S.J. Wright, R.D. Nowak, Toeplitz-structured compressed sensing matrices, in *2007 IEEE/SP 14th Workshop on Statistical Signal Processing* (Aug. 2007), pp. 294–298
7. A.S. Bandeira, M.E. Lewis, D.G. Mixon, Discrete Uncertainty Principles and Sparse Signal Processing. J. Fourier Anal. Appl. **24**(4), 935–956 (2018)
8. R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008)
9. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
10. S. Becker, J. Bobin, E.J. Candès, Nesta: A fast and accurate first-order method for sparse recovery. SIAM J. Imaging Sci. **4**, 1–39 (2011)
11. J. Bennett, S. Lanning, The netflix prize (2007)
12. R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, M.J. Strauss, Combining geometry and combinatorics: a unified approach to sparse signal recovery, in *2008 46th Annual Allerton Conference on Communication, Control, and Computing* (Sept. 2008), pp. 798–805
13. B.N. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimation. IEEE Trans. Signal Process. **61**(23), 5987–5999 (2011)
14. H. Boche, *Compressed Sensing and its Applications* (Springer Science+Business Media, New York, 2015)
15. P. Boufounos, G. Kutyniok, H. Rauhut, Sparse recovery from combined fusion frame measurements. IEEE Trans. Inf. Theory **57**(6), 3864–3876 (2011)
16. P.T. Boufounos, L. Jacques, F. Krahmer, R. Saab, Quantization and compressive sensing, in *Compressed Sensing and its Applications: MATHEON Workshop 2013*, Applied and Numerical Harmonic Analysis, ed. by H. Boche, R. Calderbank, G. Kutyniok, J. Vybíral (Springer International Publishing, Cham, 2015), pp. 193–237
17. J. Bourgain, An Improved Estimate in the Restricted Isometry Problem, in *Geometric Aspects of Functional Analysis*, vol. 2116, ed. by B. Klartag, E. Milman (Springer International Publishing, Cham, 2014), pp. 65–70
18. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004)
19. E. Candes, J. Romberg, l1-magic: recovery of sparse signals via convex programming, vol. 4 (2005), p. 14. www.acm.caltech.edu/l1magic/downloads/l1magic.pdf
20. E. Candes, T. Tao, The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)
21. E.J. Candès, The restricted isometry property and its implications for compressed sensing. C. R. Math. **346**(9), 589–592 (2008)
22. E.J. Candes, D.L. Donoho, Curvelets-a surprisingly effective nonadaptive representation for objects with edges, in *Curves and Surfaces Fitting*, ed. by L.L. Schumaker, A. Cohen, C. Rabut (Vanderbilt University Press, Nashville, TN, 1999), p. 16

23. E.J. Candès, D.L. Donoho, New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. Commun. Pure Appl. Math. J. Issued Courant Inst. Math. Sci. **57**(2), 219–266 (2004)
24. E.J. Candes, Y. Plan, Near-ideal model selection by ℓ_1 minimization. Ann. Stat. **37**, 2145–2177 (2009)
25. E.J. Candès, J.K. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
26. E.J. Candès, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
27. E.J. Candes, T. Tao, Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
28. E.J. Candès, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
29. A.Y. Carmi, L. Mihaylova, S.J. Godsill, *Compressed Sensing & Sparse Filtering* (Springer, 2016)
30. P.G. Casazza, G. Kutyniok, F. Philipp, Introduction to finite frame theory, in *Finite Frames* (Springer, 2013), pp. 1–53
31. V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
32. M. Cheraghchi, V. Guruswami, A. Velingker, Restricted isometry of Fourier matrices and list decodability of random linear codes. SIAM J. Comput. **42**(5), 1888–1914 (2013)
33. A. Cohen, W. Dahmen, R. Devore, Compressed sensing and best k-term approximation. J. Am. Math. Soc. 211–231 (2009)
34. R. Coifman, F. Geshwind, Y. Meyer, Noiselets. Appl. Comput. Harmon. Anal. **10**(1), 27–44 (2001)
35. W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans. Inf. Theory **55**, 2230–2249 (2009)
36. S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms **22**(1), 60–65 (2003)
37. R.A. DeVore, Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
38. S. Diamond, S. Boyd, Cvxpy: a python-embedded modeling language for convex optimization. J. Mach. Learn. Res. **17**(1), 2909–2913 (2016)
39. S. Dirksen, G. Lecué, H. Rauhut, On the gap between restricted isometry properties and sparse recovery conditions. IEEE Trans. Inf. Theory **64**(8), 5478–5487 (2018)
40. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
41. D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003)
42. D.L. Donoho, M. Elad, V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**, 6–18 (2006)
43. D.L. Donoho, I. Johnstone, A. Montanari, Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. IEEE Trans. Inf. Theory **59**, 3396–3433 (2013)
44. D.L. Donoho, A. Maleki, A. Montanari, Message passing algorithms for compressed sensing. Proc. Natl. Acad. Sci. U. S. A. **106**(45), 18914–9 (2009)
45. D.L. Donoho, J. Tanner, Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. Philos. Trans. Ser. A Math. Phys. Eng. Sci. **367** (1906), 4273–4293 (2009)
46. M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. (Springer, New York, 2010). OCLC: ocn646114450
47. Y.C. Eldar, G. Kutyniok (eds.), *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, 2012)
48. E. Elhamifar, R. Vidal, Sparse subspace clustering, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (June 2009), pp. 2790–2797

49. H.G. Feichtinger, T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications* (Springer Science & Business Media, 2012)
50. M. Fornasier, S. Peter, An overview on algorithms for sparse recovery, in *Sparse Reconstruction and Compressive Sensing in Remote Sensing*, ed. by X. Zhu, R. Bamler (Springer, June 2015), p. 76
51. M. Fornasier, H. Rauhut, Compressive sensing, in *Handbook of Mathematical Methods in Imaging*, ed. by O. Scherzer (Springer, New York, 2011), pp. 187–228. https://doi.org/10.1007/978-0-387-92920-0_6
52. S. Foucart, Flavors of compressive sensing, in *Approximation Theory XV: San Antonio 2016*, ed. by G.E. Fasshauer, L.L. Schumaker (Springer International Publishing, Cham, 2017), pp. 61–104
53. S. Foucart, A. Pajor, H. Rauhut, T. Ullrich, The Gelfand widths of ℓ_p-balls for $0<p \leq 1$. J. Complex. **26**(6), 629–640 (2010)
54. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Birkhäuser, Basel, 2013)
55. R. Foygel, L.W. Mackey, Corrupted sensing: novel guarantees for separating structured signals. IEEE Trans. Inf. Theory **60**, 1223–1247 (2014)
56. D. Goldberg, D. Nichols, B.M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry. Commun. ACM **35**(12), 61–70 (1992)
57. Y. Gordon, On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, in *Geometric Aspects of Functional Analysis*, ed. by J. Lindenstrauss, V.D. Milman (Springer, Berlin, 1988), pp. 84–106
58. J. Gouveia, P.A. Parrilo, R.R. Thomas, Theta bodies for polynomial ideals. SIAM J. Optim. **20**, 2097–2118 (2010)
59. M. Grant, S. Boyd, Y. Ye, CVX: Matlab software for disciplined convex programming (2008)
60. Z. Han, H. Li, W. Yin, *Compressive Sensing for Wireless Networks* (Cambridge University Press, 2013)
61. I. Haviv, O. Regev, The restricted isometry property of subsampled fourier matrices, in *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics (Springer, Cham, 2017), pp. 163–179
62. W.B. Johnson, J. Lindenstrauss, Extensions of lipschitz mappings into a hilbert space. Contemp. Math. **26**(189–206), 1 (1984)
63. V. Koltchinskii, *Oracle inequalities in empirical risk minimization and sparse recovery problems: École d'été de probabilités de Saint-Flour XXXVIII-2008*. Number 2033 in Lecture notes in mathematics. (Springer, Berlin, 2011). OCLC: ocn733246860
64. F. Krahmer, S. Mendelson, H. Rauhut, Suprema of chaos processes and the restricted isometry property. Commun. Pure Appl. Math. **67**(11), 1877–1904 (2014)
65. G. Kutyniok, D. Labate (eds.), *Shearlets: multiscale analysis for multivariate data*. Applied and Numerical Harmonic Analysis (Birkhäuser, New York, 2012). OCLC: ocn794844320
66. C. Liaw, A. Mehrabian, Y. Plan, R. Vershynin, A simple tool for bounding the deviation of random matrices on geometric sets (2016). CoRR, arXiv:1603.00897
67. G.G. Lorentz, M.V. Golitschek, Y. Makovoz, *Constructive Approximation: Advanced Problems* (Springer, Berlin, 2005). OCLC: 903339623
68. S.G. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd edn. (Elsevier/Academic Press, Amsterdam, 2009)
69. C.A. Metzler, A. Maleki, R.G. Baraniuk, From denoising to compressed sensing. IEEE Trans. Inf. Theory **62**, 5117–5144 (2016)
70. M. Mishali, Y.C. Eldar, Blind multiband signal reconstruction: compressed sensing for analog signals. IEEE Trans. Signal Process. **57**(3), 993–1009 (2009)
71. Q. Mo, A sharp restricted isometry constant bound of orthogonal matching pursuit (2015). CoRR, arXiv:1501.01708
72. B.K. Natarajan, Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
73. S. Nathan, A. Shraibman, Rank, trace-norm and max-norm, in *COLT* (2005)

74. J. Nelson, E. Price, M. Wootters, New constructions of rip matrices with fast multiplication and fewer rows, in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics (2014), pp. 1515–1528
75. Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st edn. (Springer Publishing Company, Incorporated, 2014)
76. S. Oymak, B. Hassibi, New null space results and recovery thresholds for matrix rank minimization (Nov. 2010). arXiv:1011.6326 [cs, math, stat]
77. N. Parikh, S.P. Boyd, Proximal algorithms. Found. Trends Optim. **1**, 127–239 (2014)
78. F. Parvaresh, H. Vikalo, S. Misra, B. Hassibi, Recovering sparse signals using sparse measurement matrices in compressed dna microarrays. IEEE J. Sel. Top. Signal Process. **2**(3), 275–285 (2008)
79. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inf. Theory **59**(1), 482–494 (2013)
80. Y. Plan, R. Vershynin, The generalized Lasso with non-linear observations. IEEE Trans. Inf. Theory **62**(3), 1528–1537 (2016)
81. Y.L. Polo, Y. Wang, A. Pandharipande, G. Leus, Compressive wide-band spectrum sensing, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (Apr. 2009), pp. 2337–2340
82. S. Rangan, Generalized approximate message passing for estimation with random linear mixing, in*2011 IEEE International Symposium on Information Theory Proceedings* (2011), pp. 2168–2172
83. S. Rangan, P. Schniter, A.K. Fletcher, Vector approximate message passing, in *2017 IEEE International Symposium on Information Theory (ISIT)* (2017), pp. 1588–1592
84. N.S. Rao, B. Recht, R.D. Nowak, Universal measurement bounds for structured sparse signal recovery, in *AISTATS* (2012)
85. H. Rauhut, Circulant and Toeplitz matrices in compressed sensing, in *SPARS 09-Signal Processing with Adaptive Sparse Structured Representations* (Saint Malo, France, Apr. 2009), p. 7
86. H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theory **54**(5), 2210–2219 (2008)
87. H. Rauhut, R. Ward, Sparse recovery for spherical harmonic expansions, in *Proceedings of the SampTA 2011* (2011)
88. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, 2015)
89. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**(8), 1025–1045 (2008)
90. S. Sarvotham, D. Baron, R.G. Baraniuk, Measurements vs. bits: compressed sensing meets information theory, in *Allerton Conference on Communication, Control and Computing* (2006)
91. M. Stojnic, ℓ_1 optimization and its various thresholds in compressed sensing, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (2010), pp. 3910–3913
92. G. Tang, B.N. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid. IEEE Trans. Inf. Theory **59**(11), 7465–7490 (2013)
93. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **67**(1), 91–108 (2005)
94. R.J. Tibshirani, The lasso problem and uniqueness (2012)
95. A.M. Tillmann, M.E. Pfetsch, The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Trans. Inf. Theory **60**, 1248–1259 (2014)
96. J.A. Tropp, Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)
97. E. van den Berg, M.P. Friedlander, Spgl1: a solver for large-scale sparse reconstruction (2007)
98. E. van den Berg, M.P. Friedlander, Probing the pareto frontier for basis pursuit solutions. SIAM J. Sci. Comput. **31**(2), 890–912 (2008)
99. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing, Theory and Applications* (Cambridge University Press, Cambridge, 2012), pp. 210–268

100. R. Vershynin, *Estimation in High Dimensions: A Geometric Perspective* (Springer International Publishing, Cham, 2015), pp. 3–66
101. L. Welch, Lower bounds on the maximum cross correlation of signals (corresp.). IEEE Trans. Inf. Theory **20**(3), 397–399 (1974)
102. J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009)
103. S.J. Wright, R.D. Nowak, M.A.T. Figueiredo, Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**, 2479–2493 (2008)
104. H. Zhang, W. Yin, L. Cheng, Necessary and sufficient conditions of solution uniqueness in 1-norm minimization. J. Optim. Theory Appl. **164**, 109–122 (2015)
105. Y. Zhang, J. Yang, W. Yin, Yall1: your algorithms for l1 (2011). http://yall1.blogs.rice.edu