1 Introduction to Compressed Sensing

Mark A. Davenport
Stanford University, Department of Statistics

Marco F. Duarte
Duke University, Department of Computer Science

Yonina C. Eldar
Technion, Israel Institute of Technology, Department of Electrical Engineering
Stanford University, Department of Electrical Engineering (Visiting)

Gitta Kutyniok
University of Osnabrueck, Institute for Mathematics

In recent years, compressed sensing (CS) has attracted considerable attention in areas of applied mathematics, computer science, and electrical engineering by suggesting that it may be possible to surpass the traditional limits of sampling theory. CS builds upon the fundamental fact that we can represent many signals using only a few non-zero coefficients in a suitable basis or dictionary. Nonlinear optimization can then enable recovery of such signals from very few measurements. In this chapter, we provide an up-to-date review of the basic theory underlying CS. After a brief historical overview, we begin with a discussion of sparsity and other low-dimensional signal models. We then treat the central question of how to accurately recover a high-dimensional signal from a small set of measurements and provide performance guarantees for a variety of sparse recovery algorithms. We conclude with a discussion of some extensions of the sparse recovery framework. In subsequent chapters of the book, we will see how the fundamentals presented in this chapter are extended in many exciting directions, including new models for describing structure in both analog and discrete-time signals, new sensing design techniques, more advanced recovery results, and emerging applications.

1.1 Introduction

We are in the midst of a digital revolution that is driving the development and deployment of new kinds of sensing systems with ever-increasing fidelity and resolution. The theoretical foundation of this revolution is the pioneering work of Kotelnikov, Nyquist, Shannon, and Whittaker on sampling continuous-time band-limited signals [162, 195, 209, 247]. Their results demonstrate that signals, images, videos, and other data can be exactly recovered from a set of uniformly spaced samples taken at the so-called *Nyquist rate* of twice the highest frequency present in the signal of interest. Capitalizing on this discovery, much of signal processing has moved from the analog to the digital domain and ridden the wave of Moore's law. Digitization has enabled the creation of sensing and processing systems that are more robust, flexible, cheaper and, consequently, more widely used than their analog counterparts.

As a result of this success, the amount of data generated by sensing systems has grown from a trickle to a torrent. Unfortunately, in many important and emerging applications, the resulting Nyquist rate is so high that we end up with far too many samples. Alternatively, it may simply be too costly, or even physically impossible, to build devices capable of acquiring samples at the necessary rate [146, 241]. Thus, despite extraordinary advances in computational power, the acquisition and processing of signals in application areas such as imaging, video, medical imaging, remote surveillance, spectroscopy, and genomic data analysis continues to pose a tremendous challenge.

To address the logistical and computational challenges involved in dealing with such high-dimensional data, we often depend on compression, which aims at finding the most concise representation of a signal that is able to achieve a target level of acceptable distortion. One of the most popular techniques for signal compression is known as *transform coding*, and typically relies on finding a basis or frame that provides *sparse* or *compressible* representations for signals in a class of interest [31, 77, 106]. By a sparse representation, we mean that for a signal of length n, we can represent it with $k \ll n$ nonzero coefficients; by a compressible representation, we mean that the signal is well-approximated by a signal with only k nonzero coefficients. Both sparse and compressible signals can be represented with high fidelity by preserving only the values and locations of the largest coefficients of the signal. This process is called *sparse approximation*, and forms the foundation of transform coding schemes that exploit signal sparsity and compressibility, including the JPEG, JPEG2000, MPEG, and MP3 standards.

Leveraging the concept of transform coding, *compressed sensing* (CS) has emerged as a new framework for signal acquisition and sensor design. CS enables a potentially large reduction in the sampling and computation costs for sensing signals that have a sparse or compressible representation. While the Nyquist-Shannon sampling theorem states that a certain minimum number of samples is required in order to perfectly capture an arbitrary bandlimited signal, when the signal is sparse in a known basis we can vastly reduce the number of measurements that need to be stored. Consequently, when sensing sparse signals we might be able to do better than suggested by classical results. This is the fundamental idea behind CS: rather than first sampling at a high rate and then compressing the sampled data, we would like to find ways to *directly* sense the data in a compressed form — i.e., at a lower sampling rate. The field of CS grew out of the work of Candès, Romberg, and Tao and of Donoho, who showed that a finite-dimensional signal having a sparse or compressible representation can be recovered from a small set of linear, nonadaptive measurements [3, 33, 40–42, 44, 82]. The design of these measurement schemes and their extensions to practical data models and acquisition systems are central challenges in the field of CS.

While this idea has only recently gained significant attraction in the signal processing community, there have been hints in this direction dating back as far as the eighteenth century. In 1795, Prony proposed an algorithm for the estimation of the parameters associated with a small number of complex exponentials sampled in the presence of noise [201]. The next theoretical leap came in the early 1900's, when Carathéodory showed that a positive linear combination of *any* k sinusoids is uniquely determined by its value at $t = 0$ and at *any* other $2k$ points in time [46, 47]. This represents far fewer samples than the number of Nyquist-rate samples when k is small and the range of possible frequencies is large. In the 1990's, this work was generalized by George, Gorodnitsky, and Rao, who studied sparsity in biomagnetic imaging and other contexts [134–136, 202]. Simultaneously, Bresler, Feng, and Venkataramani proposed a sampling scheme for acquiring certain classes of signals consisting of k components with nonzero bandwidth (as opposed to pure sinusoids) under restrictions on the possible spectral supports, although exact recovery was not guaranteed in general [29, 117, 118, 237]. In the early 2000's Blu, Marziliano, and Vetterli developed sampling methods for certain classes of parametric signals that are governed by only k parameters, showing that these signals can be sampled and recovered from just $2k$ samples [239].

A related problem focuses on recovery of a signal from partial observation of its Fourier transform. Beurling proposed a method for extrapolating these observations to determine the entire Fourier transform [22]. One can show that if the signal consists of a finite number of impulses, then Beurling's approach will correctly recover the entire Fourier transform (of this non-bandlimited signal) from *any* sufficiently large piece of its Fourier transform. His approach — to find the signal with smallest ℓ_1 norm among all signals agreeing with the acquired Fourier measurements — bears a remarkable resemblance to some of the algorithms used in CS.

More recently, Candès, Romberg, Tao [33, 40–42, 44], and Donoho [82] showed that a signal having a sparse representation can be recovered *exactly* from a small set of linear, nonadaptive measurements. This result suggests that it may be possible to sense sparse signals by taking far fewer measurements, hence the name *compressed* sensing. Note, however, that CS differs from classical sampling in three important respects. First, sampling theory typically considers infinite length, continuous-time signals. In contrast, CS is a mathematical theory focused on measuring finite-dimensional vectors in $\mathbb{R}^n$. Second, rather than sampling the signal at specific points in time, CS systems typically acquire measurements in the form of inner products between the signal and more general test functions. This is in fact in the spirit of modern sampling methods which similarly acquire signals by more general linear measurements [113, 230]. We will see throughout this book that *randomness* often plays a key role in the design of these test functions. Thirdly, the two frameworks differ in the manner in which they deal with *signal recovery*, i.e., the problem of recovering the original signal from the compressive measurements. In the Nyquist-Shannon framework, signal recovery is achieved through sinc interpolation — a linear process that requires little computation and has a simple interpretation. In CS, however, signal recovery is typically achieved using highly nonlinear methods.[1] See Section 1.6 as well as the survey in [226] for an overview of these techniques.

CS has already had notable impact on several applications. One example is medical imaging [178–180, 227], where it has enabled speedups by a factor of seven in pediatric MRI while preserving diagnostic quality [236]. Moreover, the broad applicability of this framework has inspired research that extends the CS framework by proposing practical implementations for numerous applications, including sub-Nyquist sampling systems [125, 126, 186–188, 219, 224, 225, 228], compressive imaging architectures [99, 184, 205], and compressive sensor networks [7, 72, 141].

The aim of this book is to provide an up-to-date review of some of the important results in CS. Many of the results and ideas in the various chapters rely on the fundamental concepts of CS. Since the focus of the remaining chapters is on more recent advances, we concentrate here on many of the basic results in CS that will serve as background material to the rest of the book. Our goal in this chapter is to provide an overview of the field and highlight some of the key technical results, which are then more fully explored in subsequent chapters. We begin with a brief review of the relevant mathematical tools, and then survey many of the low-dimensional models commonly used in CS, with an emphasis on sparsity and the union of subspaces models. We next focus attention on the theory and algorithms for sparse recovery in finite dimensions. To facilitate our goal of providing both an elementary introduction as well as a comprehensive overview of many of the results in CS, we provide proofs of some of the more technical lemmas and theorems in the Appendix.

1.2 Review of Vector Spaces

For much of its history, signal processing has focused on signals produced by physical systems. Many natural and man-made systems can be modeled as linear. Thus, it is natural to consider signal models that complement this kind of linear structure. This notion has been incorporated into modern signal processing by modeling signals as *vectors* living in an appropriate *vector space*. This captures

[1] It is also worth noting that it has recently been shown that nonlinear methods can be used in the context of traditional sampling as well, when the sampling mechanism is nonlinear [105].

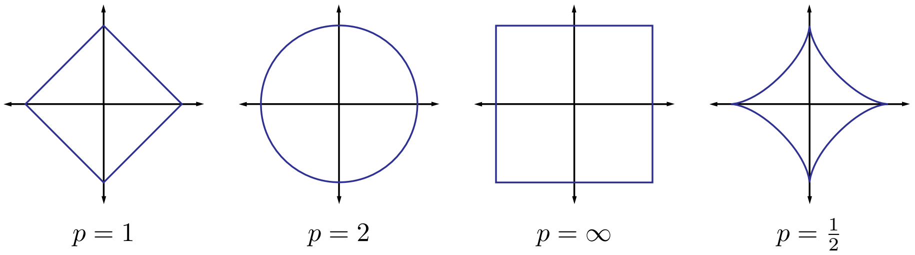


Figure 1.1 Unit spheres in $\mathbb{R}^2$ for the ℓ_p norms with $p = 1, 2, \infty$, and for the ℓ_p quasinorm with $p = \frac{1}{2}$.

the linear structure that we often desire, namely that if we add two signals together then we obtain a new, physically meaningful signal. Moreover, vector spaces allow us to apply intuitions and tools from geometry in $\mathbb{R}^3$, such as lengths, distances, and angles, to describe and compare signals of interest. This is useful even when our signals live in high-dimensional or infinite-dimensional spaces. This book assumes that the reader is relatively comfortable with vector spaces. We now provide only a brief review of some of the key concepts in vector spaces that will be required in developing the CS theory.

1.2.1 Normed vector spaces

Throughout this book, we will treat signals as real-valued functions having domains that are either continuous or discrete, and either infinite or finite. These assumptions will be made clear as necessary in each chapter. We will typically be concerned with *normed vector spaces*, i.e., vector spaces endowed with a *norm.*

In the case of a discrete, finite domain, we can view our signals as vectors in an n-dimensional Euclidean space, denoted by $\mathbb{R}^n$. When dealing with vectors in $\mathbb{R}^n$, we will make frequent use of the ℓ_p norms, which are defined for $p \in [1, \infty]$ as

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, & p \in [1, \infty); \\ \max\limits_{i=1,2,\ldots,n} |x_i|, & p = \infty. \end{cases} \tag{1.1}$$

In Euclidean space we can also consider the standard *inner product* in $\mathbb{R}^n$, which we denote

$$\langle x, z \rangle = z^T x = \sum_{i=1}^n x_i z_i.$$

This inner product leads to the ℓ_2 norm: $\|x\|_2 = \sqrt{\langle x, x \rangle}$.

In some contexts it is useful to extend the notion of ℓ_p norms to the case where $p < 1$. In this case, the "norm" defined in (1.1) fails to satisfy the triangle inequality, so it is actually a quasinorm. We will also make frequent use of the notation $\|x\|_0 := |\mathrm{supp}(x)|$, where $\mathrm{supp}(x) = \{i : x_i \neq 0\}$ denotes the support of

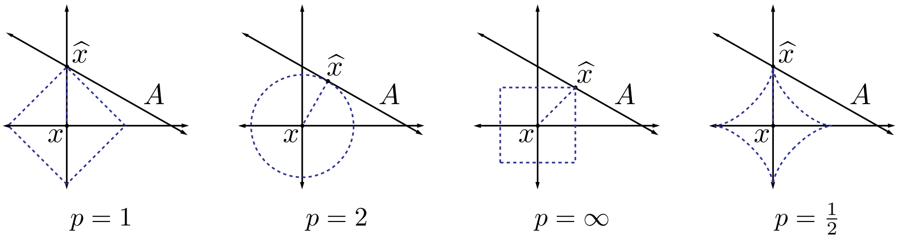


Figure 1.2 Best approximation of a point in $\mathbb{R}^2$ by a one-dimensional subspace using the ℓ_p norms for $p = 1, 2, \infty$, and the ℓ_p quasinorm with $p = \frac{1}{2}$.

x and $|\mathrm{supp}(x)|$ denotes the cardinality of $\mathrm{supp}(x)$. Note that $\|\cdot\|_0$ is not even a quasinorm, but one can easily show that

$$\lim_{p\to 0} \|x\|_p^p = |\mathrm{supp}(x)|,$$

justifying this choice of notation. The ℓ_p (quasi-)norms have notably different properties for different values of p. To illustrate this, in Fig. 1.1 we show the unit sphere, i.e., $\{x : \|x\|_p = 1\}$, induced by each of these norms in $\mathbb{R}^2$.

We typically use norms as a measure of the strength of a signal, or the size of an error. For example, suppose we are given a signal $x \in \mathbb{R}^2$ and wish to approximate it using a point in a one-dimensional affine space A. If we measure the approximation error using an ℓ_p norm, then our task is to find the $\widehat{x} \in A$ that minimizes $\|x - \widehat{x}\|_p$. The choice of p will have a significant effect on the properties of the resulting approximation error. An example is illustrated in Fig. 1.2. To compute the closest point in A to x using each ℓ_p norm, we can imagine growing an ℓ_p sphere centered on x until it intersects with A. This will be the point $\widehat{x} \in A$ that is closest to x in the corresponding ℓ_p norm. We observe that larger p tends to spread out the error more evenly among the two coefficients, while smaller p leads to an error that is more unevenly distributed and tends to be sparse. This intuition generalizes to higher dimensions, and plays an important role in the development of CS theory.

1.2.2 Bases and frames

A set $\{\phi_i\}_{i=1}^n$ is called a basis for $\mathbb{R}^n$ if the vectors in the set span $\mathbb{R}^n$ and are linearly independent.[2] This implies that each vector in the space has a unique representation as a linear combination of these basis vectors. Specifically, for any $x \in \mathbb{R}^n$, there exist (unique) coefficients $\{c_i\}_{i=1}^n$ such that

$$x = \sum_{i=1}^n c_i \phi_i.$$

[2] In any n-dimensional vector space, a basis will always consist of exactly n vectors. Fewer vectors are not sufficient to span the space, while additional vectors are guaranteed to be linearly dependent.

Note that if we let Φ denote the $n \times n$ matrix with columns given by ϕ_i and let c denote the length-n vector with entries c_i, then we can represent this relation more compactly as

$$x = \Phi c.$$

An important special case of a basis is an orthonormal basis, defined as a set of vectors $\{\phi_i\}_{i=1}^n$ satisfying

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

An orthonormal basis has the advantage that the coefficients c can be easily calculated as

$$c_i = \langle x, \phi_i \rangle,$$

or

$$c = \Phi^T x$$

in matrix notation. This can easily be verified since the orthonormality of the columns of Φ means that $\Phi^T \Phi = I$, where I denotes the $n \times n$ identity matrix.

It is often useful to generalize the concept of a basis to allow for sets of possibly linearly dependent vectors, resulting in what is known as a *frame* [48, 55, 65, 163, 164, 182]. More formally, a frame is a set of vectors $\{\phi_i\}_{i=1}^n$ in $\mathbb{R}^d$, $d < n$ corresponding to a matrix $\Phi \in \mathbb{R}^{d \times n}$, such that for all vectors $x \in \mathbb{R}^d$,

$$A \|x\|_2^2 \leq \left\|\Phi^T x\right\|_2^2 \leq B \|x\|_2^2$$

with $0 < A \leq B < \infty$. Note that the condition $A > 0$ implies that the rows of Φ must be linearly independent. When A is chosen as the largest possible value and B as the smallest for these inequalities to hold, then we call them the *(optimal) frame bounds*. If A and B can be chosen as $A = B$, then the frame is called *A-tight*, and if $A = B = 1$, then Φ is a *Parseval frame*. A frame is called *equal-norm*, if there exists some $\lambda > 0$ such that $\|\phi_i\|_2 = \lambda$ for all $i = 1, \ldots, n$, and it is *unit-norm* if $\lambda = 1$. Note also that while the concept of a frame is very general and can be defined in infinite-dimensional spaces, in the case where Φ is a $d \times n$ matrix A and B simply correspond to the smallest and largest eigenvalues of $\Phi\Phi^T$, respectively.

Frames can provide richer representations of data due to their redundancy [26]: for a given signal x, there exist infinitely many coefficient vectors c such that $x = \Phi c$. In order to obtain a set of feasible coefficients we exploit the *dual frame* $\widetilde{\Phi}$. Specifically, any frame satisfying

$$\Phi \widetilde{\Phi}^T = \widetilde{\Phi} \Phi^T = I$$

is called an (alternate) dual frame. The particular choice $\tilde{\Phi} = (\Phi\Phi^T)^{-1}\Phi$ is referred to as the *canonical dual frame*. It is also known as the Moore-Penrose pseudoinverse. Note that since $A > 0$ requires Φ to have linearly independent rows, this also ensures that $\Phi\Phi^T$ is invertible, so that $\tilde{\Phi}$ is well-defined. Thus, one way to obtain a set of feasible coefficients is via

$$c_d = (\Phi\Phi^T)^{-1}\Phi x.$$

One can show that this sequence is the smallest coefficient sequence in ℓ_2 norm, i.e., $\|c_d\|_2 \leq \|c\|_2$ for all c such that $x = \Phi c$.

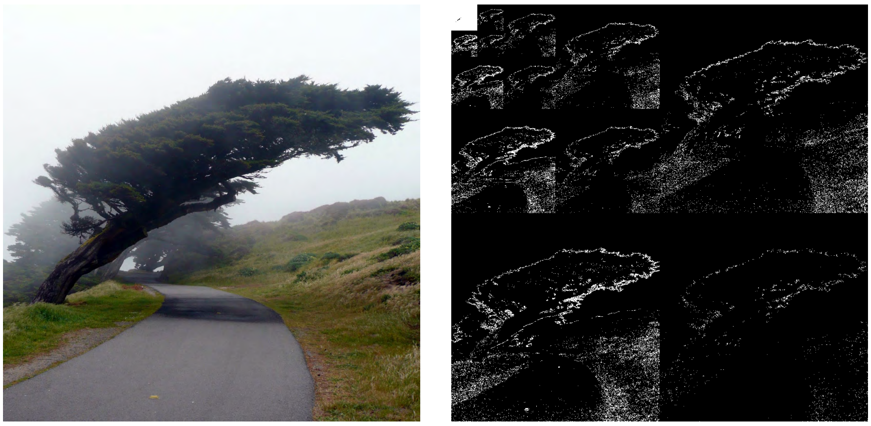
Finally, note that in the sparse approximation literature, it is also common for a basis or frame to be referred to as a *dictionary* or *overcomplete dictionary* respectively, with the dictionary elements being called *atoms*.

1.3 Low-Dimensional Signal Models

At its core, signal processing is concerned with efficient algorithms for acquiring, processing, and extracting information from different types of signals or data. In order to design such algorithms for a particular problem, we must have accurate *models* for the signals of interest. These can take the form of generative models, deterministic classes, or probabilistic Bayesian models. In general, models are useful for incorporating *a priori* knowledge to help distinguish classes of interesting or probable signals from uninteresting or improbable signals. This can help in efficiently and accurately acquiring, processing, compressing, and communicating data and information.

As noted in the introduction, much of classical signal processing is based on the notion that signals can be modeled as vectors living in an appropriate vector space (or subspace). To a large extent, the notion that any possible vector is a valid signal has driven the explosion in the dimensionality of the data we must sample and process. However, such simple linear models often fail to capture much of the structure present in many common classes of signals — while it may be reasonable to model signals as vectors, in many cases not all possible vectors in the space represent valid signals. In response to these challenges, there has been a surge of interest in recent years, across many fields, in a variety of *low-dimensional signal models* that quantify the notion that the number of degrees of freedom in high-dimensional signals is often quite small compared to their ambient dimensionality.

In this section we provide a brief overview of the most common low-dimensional structures encountered in the field of CS. We will begin by considering the traditional sparse models for finite-dimensional signals, and then discuss methods for generalizing these classes to infinite-dimensional (continuous-time) signals. We will also briefly discuss low-rank matrix and manifold models and describe some interesting connections between CS and some other emerging problem areas.



(a) (b)

Figure 1.3 Sparse representation of an image via a multiscale wavelet transform. (a) Original image. (b) Wavelet representation. Large coefficients are represented by light pixels, while small coefficients are represented by dark pixels. Observe that most of the wavelet coefficients are close to zero.

1.3.1 Sparse models

Signals can often be well-approximated as a linear combination of just a few elements from a known basis or dictionary. When this representation is exact we say that the signal is *sparse*. Sparse signal models provide a mathematical framework for capturing the fact that in many cases these high-dimensional signals contain relatively little information compared to their ambient dimension. Sparsity can be thought of as one incarnation of *Occam's razor* — when faced with many possible ways to represent a signal, the simplest choice is the best one.

Sparsity and nonlinear approximation

Mathematically, we say that a signal x is *k-sparse* when it has at most k nonzeros, i.e., $\|x\|_0 \leq k$. We let

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

denote the set of all k-sparse signals. Typically, we will be dealing with signals that are not themselves sparse, but which admit a sparse representation in some basis Φ. In this case we will still refer to x as being k-sparse, with the understanding that we can express x as $x = \Phi c$ where $\|c\|_0 \leq k$.

Sparsity has long been exploited in signal processing and approximation theory for tasks such as compression [77, 199, 215] and denoising [80], and in statistics and learning theory as a method for avoiding overfitting [234]. Sparsity also figures prominently in the theory of statistical estimation and model selection [139, 218], in the study of the human visual system [196], and has been exploited heavily in image processing tasks, since the multiscale wavelet trans-

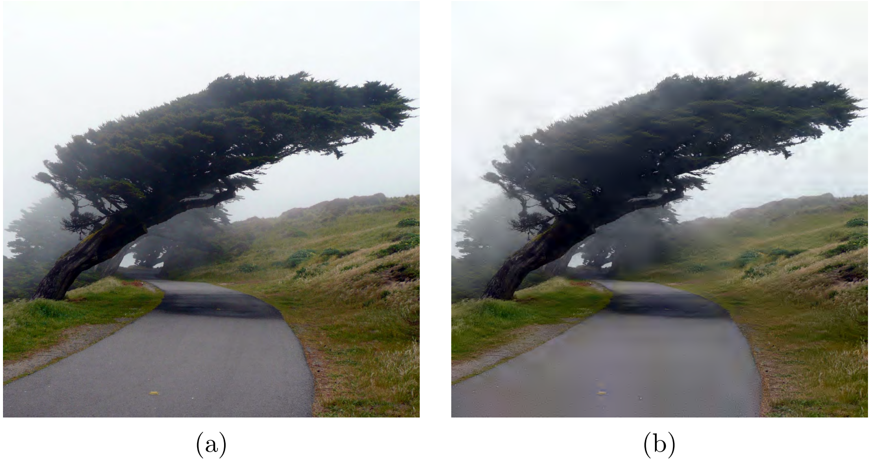


Figure 1.4 Sparse approximation of a natural image. (a) Original image. (b) Approximation of image obtained by keeping only the largest 10% of the wavelet coefficients.

form [182] provides nearly sparse representations for natural images. An example is shown in Fig. 1.3.

As a traditional application of sparse models, we consider the problems of image compression and image denoising. Most natural images are characterized by large smooth or textured regions and relatively few sharp edges. Signals with this structure are known to be very nearly sparse when represented using a multiscale wavelet transform [182]. The wavelet transform consists of recursively dividing the image into its low- and high-frequency components. The lowest frequency components provide a coarse scale approximation of the image, while the higher frequency components fill in the detail and resolve edges. What we see when we compute a wavelet transform of a typical natural image, as shown in Fig. 1.3, is that most coefficients are very small. Hence, we can obtain a good approximation of the signal by setting the small coefficients to zero, or *thresholding* the coefficients, to obtain a k-sparse representation. When measuring the approximation error using an ℓ_p norm, this procedure yields the *best k-term approximation* of the original signal, i.e., the best approximation of the signal using only k basis elements.[3]

Figure 1.4 shows an example of such an image and its best k-term approximation. This is the heart of nonlinear approximation [77] — nonlinear because the choice of which coefficients to keep in the approximation depends on the signal itself. Similarly, given the knowledge that natural images have approximately sparse wavelet transforms, this same thresholding operation serves as an effec-

[3] Thresholding yields the best k-term approximation of a signal with respect to an orthonormal basis. When redundant frames are used, we must rely on sparse approximation algorithms like those described in Section 1.6 [106, 182].

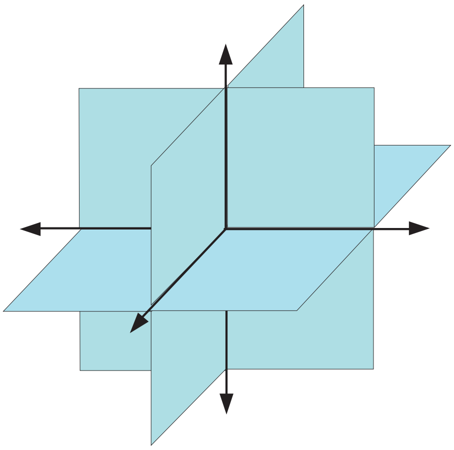


Figure 1.5 Union of subspaces defined by $\Sigma_2 \subset \mathbb{R}^3$, i.e., the set of all 2-sparse signals in $\mathbb{R}^3$.

tive method for rejecting certain common types of noise, which typically do *not* have sparse wavelet transforms [80].

Geometry of sparse signals

Sparsity is a highly nonlinear model, since the choice of which dictionary elements are used can change from signal to signal [77]. This can be seen by observing that given a pair of k-sparse signals, a linear combination of the two signals will in general no longer be k sparse, since their supports may not coincide. That is, for any $x, z \in \Sigma_k$, we do not necessarily have that $x + z \in \Sigma_k$ (although we do have that $x + z \in \Sigma_{2k}$). This is illustrated in Fig. 1.5, which shows Σ_2 embedded in $\mathbb{R}^3$, i.e., the set of all 2-sparse signals in $\mathbb{R}^3$.

The set of sparse signals Σ_k does not form a linear space. Instead it consists of the union of all possible $\binom{n}{k}$ canonical subspaces.[4] In Fig. 1.5 we have only $\binom{3}{2} = 3$ possible subspaces, but for larger values of n and k we must consider a potentially huge number of subspaces. This will have significant algorithmic consequences in the development of the algorithms for sparse approximation and sparse recovery described in Sections 1.5 and 1.6.

Compressible signals

An important point in practice is that few real-world signals are *truly* sparse; rather they are compressible, meaning that they can be well-approximated by a sparse signal. Such signals have been termed compressible, approximately sparse, or relatively sparse in various contexts. Compressible signals are well approximated by sparse signals in the same way that signals living close to a subspace

[4] Union of subspaces

are well approximated by the first few principal components [139]. In fact, we can quantify the compressibility by calculating the error incurred by approximating a signal x by some $\widehat{x} \in \Sigma_k$:

$$\sigma_k(x)_p = \min_{\widehat{x} \in \Sigma_k} \left\| x - \widehat{x} \right\|_p . \tag{1.2}$$

If $x \in \Sigma_k$, then clearly $\sigma_k(x)_p = 0$ for any p. Moreover, one can easily show that the thresholding strategy described above (keeping only the k largest coefficients) results in the optimal approximation as measured by (1.2) for all ℓ_p norms [77].

Another way to think about compressible signals is to consider the rate of decay of their coefficients. For many important classes of signals there exist bases such that the coefficients obey a power law decay, in which case the signals are highly compressible. Specifically, if $x = \Phi c$ and we sort the coefficients c_i such that $|c_1| \geq |c_2| \geq \cdots \geq |c_n|$, then we say that the coefficients obey a power law decay if there exist constants $C_1, q > 0$ such that

$$|c_i| \leq C_1 i^{-q}.$$

The larger q is, the faster the magnitudes decay, and the more compressible a signal is. Because the magnitudes of their coefficients decay so rapidly, compressible signals can be represented accurately by $k \ll n$ coefficients. Specifically, for such signals there exist constants $C_2, r > 0$ depending only on C_1 and q such that

$$\sigma_k(x)_2 \leq C_2 k^{-r}.$$

In fact, one can show that $\sigma_k(x)_2$ will decay as k^{-r} if and only if the sorted coefficients c_i decay as $i^{-r+1/2}$ [77].

1.3.2 Finite unions of subspaces

In certain applications, the signal has a structure that cannot be completely expressed using sparsity alone. For instance, when only certain sparse support patterns are allowable in the signal, it is possible to leverage such constraints to formulate more concise signal models. We give a few representative examples below; see Chapters 2 and 8 for more detail on structured sparsity.

- For piecewise-smooth signals and images, the dominant coefficients in the wavelet transform tend to cluster into a connected rooted subtree inside the wavelet parent-child binary tree [79, 103, 104, 167, 168].
- In applications such as surveillance or neuronal recording, the coefficients might appear clustered together, or spaced apart from each other [49, 50, 147]. See Chapter 11 for more details.
- When multiple sparse signals are recorded simultaneously, their supports might be correlated according to the properties of the sensing environment [7, 63, 76, 114, 121, 185]. One possible structure leads to the multiple measurement vector problem; see Section 1.7 for more details.

- In certain cases the small number of components of a sparse signal correspond not to vectors (columns of a matrix Φ), but rather to points known to lie in particular subspaces. If we construct a frame by concatenating bases for such subspaces, the nonzero coefficients of the signal representations form block structures at known locations [27, 112, 114]. See Chapters 3, 11, and 12 for further description and potential applications of this model.

Such examples of additional structure can be captured in terms of restricting the feasible signal supports to a small subset of the possible $\binom{n}{k}$ selections of nonzero coefficients for a k-sparse signal. These models are often referred to as structured sparsity models [4, 25, 102, 114, 177]. In cases where nonzero coefficients appear in clusters, the structure can be expressed in terms of a sparse union of subspaces [102, 114]. Structured sparse and union of subspace models extend the notion of sparsity to a much broader class of signals that can incorporate both finite-dimensional and infinite-dimensional representations.

In order to define these models, recall that for canonically sparse signals, the union Σ_k is composed of canonical subspaces $\mathcal{U}_i$ that are aligned with k out of the n coordinate axes of $\mathbb{R}^n$. See, for example, Fig. 1.5, which illustrates this for the case where $n = 3$ and $k = 2$. Allowing for more general choices of $\mathcal{U}_i$ leads to powerful representations that accommodate many interesting signal priors. Specifically, given the knowledge that x resides in one of M possible subspaces $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_M$, we have that x lies in the *union* of M subspaces! [114, 177]:

$$x \in \mathcal{U} = \bigcup_{i=1}^{M} \mathcal{U}_i.$$

It is important to note that, as in the generic sparse setting, union models are nonlinear: the sum of two signals from a union $\mathcal{U}$ is generally no longer in $\mathcal{U}$. This nonlinear behavior of the signal set renders any processing that exploits these models more intricate. Therefore, instead of attempting to treat all unions in a unified way, we focus our attention on some specific classes of union models, in order of complexity.

The simplest class of unions arises when the number of subspaces comprising the union is finite, and each subspace has finite dimensions. We call this setup a finite union of subspaces model. Under the finite-dimensional framework, we revisit the two types of models described above:

- *Structured sparse supports*: This class consists of sparse vectors that meet additional restrictions on the support (i.e., the set of indices for the vector's nonzero entries). This corresponds to only certain subspaces $\mathcal{U}_i$ out of the $\binom{n}{k}$ subspaces present in Σ_k being allowed [4].

- *Sparse union of subspaces* where each subspace $\mathcal{U}_i$ comprising the union is a direct sum of k low-dimensional subspaces [114].

$$\mathcal{U}_i = \bigoplus_{j=1}^{k} \mathcal{A}_{i_j}. \tag{1.3}$$

  Here $\{\mathcal{A}_i\}$ are a given set of subspaces with dimensions $\dim(\mathcal{A}_i) = d_i$, and $i_1, i_2, \ldots, i_k$ select k of these subspaces. Thus, each subspace $\mathcal{U}_i$ corresponds to a different choice of k out of M subspaces $\mathcal{A}_i$ that comprise the sum. This framework can model standard sparsity by letting $\mathcal{A}_j$ be the one-dimensional subspace spanned by the j^{th} canonical vector. It can be shown that this model leads to block sparsity in which certain blocks in a vector are zero, and others are not [112].

These two cases can be combined to allow for only certain sums of k subspaces to be part of the union $\mathcal{U}$. Both models can be leveraged to further reduce sampling rate and allow for CS of a broader class of signals.

1.3.3 Unions of subspaces for analog signal models

One of the primary motivations for CS is to design new sensing systems for acquiring continuous-time, analog signals or images. In contrast, the finite-dimensional sparse model described above inherently assumes that the signal x is discrete. It is sometimes possible to extend this model to continuous-time signals using an intermediate discrete representation. For example, a band-limited, periodic signal can be perfectly represented by a finite-length vector consisting of its Nyquist-rate samples. However, it will often be more useful to extend the concept of sparsity to provide union of subspaces models for analog signals [97, 109, 114, 125, 186–188, 239]. Two of the broader frameworks that treat sub-Nyquist sampling of analog signals are Xampling and finite-rate of innovation, which are discussed in Chapters 3 and 4, respectively.

In general, when treating unions of subspaces for analog signals there are three main cases to consider, as elaborated further in Chapter 3 [102]:

- finite unions of infinite dimensional spaces;
- infinite unions of finite dimensional spaces;
- infinite unions of infinite dimensional spaces.

In each of the three settings above there is an element that can take on infinite values, which is a result of the fact that we are considering analog signals: either the underlying subspaces are infinite-dimensional, or the number of subspaces is infinite.

There are many well-known examples of analog signals that can be expressed as a union of subspaces. For example, an important signal class corresponding to a finite union of infinite dimensional spaces is the multiband model [109]. In this model, the analog signal consists of a finite sum of bandlimited signals, where typically the signal components have a relatively small bandwidth but are distributed across a comparatively large frequency range [117, 118, 186, 237, 238]. Sub-Nyquist recovery techniques for this class of signals can be found in [186–188].

Another example of a signal class that can often be expressed as a union of subspaces is the class of signals having a finite rate of innovation [97, 239]. Depending on the specific structure, this model corresponds to an infinite or finite union of finite dimensional subspaces [19, 125, 126], and describes many common signals having a small number of degrees of freedom. In this case, each subspace corresponds to a certain choice of parameter values, with the set of possible values being infinite dimensional, and thus the number of subspaces spanned by the model being infinite as well. The eventual goal is to exploit the available structure in order to reduce the sampling rate; see Chapters 3 and 4 for more details. As we will see in Chapter 3, by relying on the analog union of subspace model we can design efficient hardware that samples analog signals at sub-Nyquist rates, thus moving the analog CS framework from theory to practice.

1.3.4 Low-rank matrix models

Another model closely related to sparsity is the set of low-rank matrices:

$$\mathcal{L} = \{M \in \mathbb{R}^{n_1 \times n_2} : \mathrm{rank}(M) \leq r\}.$$

The set $\mathcal{L}$ consists of matrices M such that $M = \sum_{k=1}^{r} \sigma_k u_k v_k^*$ where $\sigma_1, \sigma_2, \ldots, \sigma_r \geq 0$ are the nonzero singular values, and $u_1, u_2, \ldots, u_r \in \mathbb{R}^{n_1}$, $v_1, v_2, \ldots, v_r \in \mathbb{R}^{n_2}$ are the corresponding singular vectors. Rather than constraining the number of elements used to construct the signal, we are constraining the number of nonzero singular values. One can easily observe that the set $\mathcal{L}$ has $r(n_1 + n_2 - r)$ degrees of freedom by counting the number of free parameters in the singular value decomposition. For small r this is significantly less than the number of entries in the matrix — $n_1 n_2$. Low-rank matrices arise in a variety of practical settings. For example, low-rank (Hankel) matrices correspond to low-order linear, time-invariant systems [198]. In many data-embedding problems, such as sensor geolocation, the matrix of pairwise distances will typically have rank 2 or 3 [172, 212]. Finally, approximately low-rank matrices arise naturally in the context of collaborative filtering systems such as the now-famous Netflix recommendation system [132] and the related problem of *matrix completion*, where a low-rank matrix is recovered from a small sample of its entries [39, 151, 204]. While we do not focus in-depth on matrix completion or the more general problem of low-rank matrix recovery, we note that many of the concepts and tools treated in this book are highly relevant to this emerging field, both from a theoretical and algorithmic perspective [36, 38, 161, 203].

1.3.5 Manifold and parametric models

Parametric or manifold models form another, more general class of low-dimensional signal models. These models arise in cases where (i) a k-dimensional continuously-valued parameter θ can be identified that carries the relevant information about a signal and (ii) the signal $f(\theta) \in \mathbb{R}^n$ changes as a continuous (typically nonlinear) function of these parameters. Typical examples include a one-dimensional (1-D) signal shifted by an unknown time delay (parameterized by the translation variable), a recording of a speech signal (parameterized by the underlying phonemes being spoken), and an image of a 3-D object at an unknown location captured from an unknown viewing angle (parameterized by the 3-D coordinates of the object and its roll, pitch, and yaw) [90, 176, 240]. In these and many other cases, the signal class forms a nonlinear k-dimensional manifold in $\mathbb{R}^n$, i.e.,

$$\mathcal{M} = \{f(\theta) : \theta \in \Theta\},$$

where Θ is the k-dimensional parameter space. Manifold-based methods for image processing have attracted considerable attention, particularly in the machine learning community. They can be applied to diverse applications including data visualization, signal classification and detection, parameter estimation, systems control, clustering, and machine learning [14, 15, 58, 61, 89, 193, 217, 240, 244]. Low-dimensional manifolds have also been proposed as approximate models for a number of nonparametric signal classes such as images of human faces and handwritten digits [30, 150, 229].

Manifold models are closely related to all of the models described above. For example, the set of signals x such that $\|x\|_0 = k$ forms a k-dimensional Riemannian manifold. Similarly, the set of $n_1 \times n_2$ matrices of rank r forms an $r(n_1 + n_2 - r)$-dimensional Riemannian manifold [233].[5] Furthermore, many manifolds can be equivalently described as an infinite union of subspaces.

A number of the signal models used in this book are closely related to manifold models. For example, the union of subspace models in Chapter 3, the finite rate of innovation models considered in Chapter 4, and the continuum models in Chapter 11 can all be viewed from a manifold perspective. For the most part we will not explicitly exploit this structure in the book. However, low-dimensional manifolds have a close connection to many of the key results in CS. In particular, many of the randomized sensing matrices used in CS can also be shown to preserve the structure in low-dimensional manifolds [6]. For details and further applications see [6, 71, 72, 101].

[5] Note that in the case where we allow signals with sparsity less than or equal to k, or matrices of rank less than or equal to r, these sets fail to satisfy certain technical requirements of a topological manifold (due to the behavior where the sparsity/rank changes). However, the manifold viewpoint can still be useful in this context [68].

1.4 Sensing Matrices

In order to make the discussion more concrete, for the remainder of this chapter we will restrict our attention to the standard finite-dimensional CS model. Specifically, given a signal $x \in \mathbb{R}^n$, we consider measurement systems that acquire m linear measurements. We can represent this process mathematically as

$$y = Ax, \tag{1.4}$$

where A is an $m \times n$ matrix and $y \in \mathbb{R}^m$. The matrix A represents a *dimensionality reduction*, i.e., it maps $\mathbb{R}^n$, where n is generally large, into $\mathbb{R}^m$, where m is typically much smaller than n. Note that in the standard CS framework we assume that the measurements are *non-adaptive*, meaning that the rows of A are fixed in advance and do not depend on the previously acquired measurements. In certain settings adaptive measurement schemes can lead to significant performance gains. See Chapter 6 for further details.

As noted earlier, although the standard CS framework assumes that x is a finite-length vector with a discrete-valued index (such as time or space), in practice we will often be interested in designing measurement systems for acquiring continuously-indexed signals such as continuous-time signals or images. It is sometimes possible to extend this model to continuously-indexed signals using an intermediate discrete representation. For a more flexible approach, we refer the reader to Chapters 3 and 4. For now we will simply think of x as a finite-length window of Nyquist-rate samples, and we temporarily ignore the issue of how to directly acquire compressive measurements without first sampling at the Nyquist rate.

There are two main theoretical questions in CS. First, how should we design the sensing matrix A to ensure that it preserves the information in the signal x? Second, how can we recover the original signal x from measurements y? In the case where our data is sparse or compressible, we will see that we can design matrices A with $m \ll n$ that ensure that we will be able to recover the original signal accurately and efficiently using a variety of practical algorithms.

We begin in this section by first addressing the question of how to design the sensing matrix A. Rather than directly proposing a design procedure, we instead consider a number of desirable properties that we might wish A to have. We then provide some important examples of matrix constructions that satisfy these properties.

1.4.1 Null space conditions

A natural place to begin is by considering the null space of A, denoted

$$\mathcal{N}(A) = \{z : Az = 0\}.$$

If we wish to be able to recover *all* sparse signals x from the measurements Ax, then it is immediately clear that for any pair of distinct vectors $x, x' \in \Sigma_k$, we must have $Ax \neq Ax'$, since otherwise it would be impossible to distinguish x from x' based solely on the measurements y. More formally, by observing that if $Ax = Ax'$ then $A(x - x') = 0$ with $x - x' \in \Sigma_{2k}$, we see that A uniquely represents all $x \in \Sigma_k$ if and only if $\mathcal{N}(A)$ contains no vectors in Σ_{2k}. While there are many equivalent ways of characterizing this property, one of the most common is known as the *spark* [86].

Definition 1.1. *The spark of a given matrix A is the smallest number of columns of A that are linearly dependent.*

This definition allows us to pose the following straightforward guarantee.

Theorem 1.1 (Corollary 1 of [86]). *For any vector $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$ if and only if* $\operatorname{spark}(A) > 2k$.

Proof. We first assume that, for any $y \in \mathbb{R}^m$, there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$. Now suppose for the sake of a contradiction that $\operatorname{spark}(A) \leq 2k$. This means that there exists some set of at most $2k$ columns that are linearly independent, which in turn implies that there exists an $h \in \mathcal{N}(A)$ such that $h \in \Sigma_{2k}$. In this case, since $h \in \Sigma_{2k}$ we can write $h = x - x'$, where $x, x' \in \Sigma_k$. Thus, since $h \in \mathcal{N}(A)$ we have that $A(x - x') = 0$ and hence $Ax = Ax'$. But this contradicts our assumption that there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$. Therefore, we must have that $\operatorname{spark}(A) > 2k$.

Now suppose that $\operatorname{spark}(A) > 2k$. Assume that for some y there exist $x, x' \in \Sigma_k$ such that $y = Ax = Ax'$. We therefore have that $A(x - x') = 0$. Letting $h = x - x'$, we can write this as $Ah = 0$. Since $\operatorname{spark}(A) > 2k$, all sets of up to $2k$ columns of A are linearly independent, and therefore $h = 0$. This in turn implies $x = x'$, proving the theorem. □

It is easy to see that $\operatorname{spark}(A) \in [2, m + 1]$. Therefore, Theorem 1.1 yields the requirement $m \geq 2k$.

When dealing with *exactly* sparse vectors, the spark provides a complete characterization of when sparse recovery is possible. However, when dealing with *approximately* sparse signals we must consider somewhat more restrictive conditions on the null space of A [57]. Roughly speaking, we must also ensure that $\mathcal{N}(A)$ does not contain any vectors that are too compressible in addition to vectors that are sparse. In order to state the formal definition we define the following notation that will prove to be useful throughout much of this book. Suppose that $\Lambda \subset \{1, 2, \ldots, n\}$ is a subset of indices and let $\Lambda^c = \{1, 2, \ldots, n\} \backslash \Lambda$. By x_Λ we typically mean the length n vector obtained by setting the entries of x indexed by Λ^c to zero. Similarly, by A_Λ we typically mean the $m \times n$ matrix obtained by setting the columns of A indexed by Λ^c to zero.[6]

Definition 1.2. *A matrix A satisfies the* null space property *(NSP) of order k if there exists a constant $C > 0$ such that,*

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{k}} \tag{1.5}$$

holds for all $h \in \mathcal{N}(A)$ and for all Λ such that $|\Lambda| \leq k$.

The NSP quantifies the notion that vectors in the null space of A should not be too concentrated on a small subset of indices. For example, if a vector h is exactly k-sparse, then there exists a Λ such that $\|h_{\Lambda^c}\|_1 = 0$ and hence (1.5) implies that $h_\Lambda = 0$ as well. Thus, if a matrix A satisfies the NSP then the only k-sparse vector in $\mathcal{N}(A)$ is $h = 0$.

To fully illustrate the implications of the NSP in the context of sparse recovery, we now briefly discuss how we will measure the performance of sparse recovery algorithms when dealing with general non-sparse x. Towards this end, let $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ represent our specific recovery method. We will focus primarily on guarantees of the form

$$\|\Delta(Ax) - x\|_2 \leq C \frac{\sigma_k(x)_1}{\sqrt{k}} \tag{1.6}$$

for all x, where $\sigma_k(x)_1$ is as defined in (1.2). This guarantees exact recovery of all possible k-sparse signals, but also ensures a degree of robustness to non-sparse signals that directly depends on how well the signals are approximated by k-sparse vectors. Such guarantees are called *instance-optimal* since they guarantee optimal performance for each instance of x [57]. This distinguishes them from guarantees that only hold for some subset of possible signals, such as sparse or compressible signals — the quality of the guarantee adapts to the particular choice of x. These are also commonly referred to as *uniform guarantees* since they hold uniformly for all x.

Our choice of norms in (1.6) is somewhat arbitrary. We could easily measure the reconstruction error using other ℓ_p norms. The choice of p, however, will limit what kinds of guarantees are possible, and will also potentially lead to alternative formulations of the NSP. See, for instance, [57]. Moreover, the form of the right-hand-side of (1.6) might seem somewhat unusual in that we measure the approximation error as $\sigma_k(x)_1/\sqrt{k}$ rather than simply something like $\sigma_k(x)_2$. However, we will see in Section 1.5.3 that such a guarantee is actually not possible

[6] We note that this notation will occasionally be abused to refer to the length $|\Lambda|$ vector obtained by keeping only the entries corresponding to Λ or the $m \times |\Lambda|$ matrix obtained by only keeping the columns corresponding to Λ respectively. The usage should be clear from the context, but in most cases there is no substantive difference between the two.

without taking a prohibitively large number of measurements, and that (1.6) represents the best possible guarantee we can hope to obtain.

We will see in Section 1.5 (Theorem 1.8) that the NSP of order $2k$ is sufficient to establish a guarantee of the form (1.6) for a practical recovery algorithm (ℓ_1 minimization). Moreover, the following adaptation of a theorem in [57] demonstrates that if there exists *any* recovery algorithm satisfying (1.6), then A must necessarily satisfy the NSP of order $2k$.

Theorem 1.2 (Theorem 3.2 of [57])**.** *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ denote a sensing matrix and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ denote an arbitrary recovery algorithm. If the pair (A, Δ) satisfies (1.6) then A satisfies the NSP of order $2k$.*

Proof. Suppose $h \in \mathcal{N}(A)$ and let Λ be the indices corresponding to the $2k$ largest entries of h. We next split Λ into Λ_0 and Λ_1, where $|\Lambda_0| = |\Lambda_1| = k$. Set $x = h_{\Lambda_1} + h_{\Lambda^c}$ and $x' = -h_{\Lambda_0}$, so that $h = x - x'$. Since by construction $x' \in \Sigma_k$, we can apply (1.6) to obtain $x' = \Delta(Ax')$. Moreover, since $h \in \mathcal{N}(A)$, we have

$$Ah = A\left(x - x'\right) = 0$$

so that $Ax' = Ax$. Thus, $x' = \Delta(Ax)$. Finally, we have that

$$\|h_\Lambda\|_2 \leq \|h\|_2 = \|x - x'\|_2 = \|x - \Delta(Ax)\|_2 \leq C\frac{\sigma_k(x)_1}{\sqrt{k}} = \sqrt{2}C\frac{\|h_{\Lambda^c}\|_1}{\sqrt{2k}},$$

where the last inequality follows from (1.6). □

1.4.2 The restricted isometry property

While the NSP is both necessary and sufficient for establishing guarantees of the form (1.6), these guarantees do not account for *noise*. When the measurements are contaminated with noise or have been corrupted by some error such as quantization, it will be useful to consider somewhat stronger conditions. In [43], Candès and Tao introduced the following isometry condition on matrices A and established its important role in CS.

Definition 1.3. *A matrix A satisfies the* restricted isometry property *(RIP) of order k if there exists a $\delta_k \in (0, 1)$ such that*

$$\left(1 - \delta_k\right) \|x\|_2^2 \leq \|Ax\|_2^2 \leq \left(1 + \delta_k\right) \|x\|_2^2, \tag{1.7}$$

holds for all $x \in \Sigma_k$.

If a matrix A satisfies the RIP of order $2k$, then we can interpret (1.7) as saying that A approximately preserves the distance between any pair of k-sparse vectors. This will clearly have fundamental implications concerning robustness to noise. Moreover, the potential applications of such *stable embeddings* range far beyond acquisition for the sole purpose of signal recovery. See Chapter 10 for examples of additional applications.

It is important to note that while in our definition of the RIP we assume bounds that are symmetric about 1, this is merely for notational convenience. In practice, one could instead consider arbitrary bounds

$$\alpha \left\|x\right\|_2^2 \leq \left\|Ax\right\|_2^2 \leq \beta \left\|x\right\|_2^2$$

where $0 < \alpha \leq \beta < \infty$. Given any such bounds, one can always scale A so that it satisfies the symmetric bound about 1 in (1.7). Specifically, multiplying A by $\sqrt{2/(\beta+\alpha)}$ will result in an $\widetilde{A}$ that satisfies (1.7) with constant $\delta_k = (\beta - \alpha)/(\beta+\alpha)$. While we will not explicitly show this, one can check that all of the theorems in this chapter based on the assumption that A satisfies the RIP actually hold as long as there exists some scaling of A that satisfies the RIP. Thus, since we can always scale A to satisfy (1.7), we lose nothing by restricting our attention to this simpler bound.

Note also that if A satisfies the RIP of order k with constant δ_k, then for any $k' < k$ we automatically have that A satisfies the RIP of order k' with constant $\delta_{k'} \leq \delta_k$. Moreover, in [190] it is shown that if A satisfies the RIP of order k with a sufficiently small constant, then it will also automatically satisfy the RIP of order γk for certain γ, albeit with a somewhat worse constant.

Lemma 1.1 (Corollary 3.4 of [190])**.** *Suppose that A satisfies the RIP of order k with constant δ_k. Let γ be a positive integer. Then A satisfies the RIP of order $k' = \gamma \left\lfloor \frac{k}{2} \right\rfloor$ with constant $\delta_{k'} < \gamma \cdot \delta_k$, where $\lfloor \cdot \rfloor$ denotes the floor operator.*

This lemma is trivial for $\gamma = 1, 2$, but for $\gamma \geq 3$ (and $k \geq 4$) this allows us to extend from RIP of order k to higher orders. Note however, that δ_k must be sufficiently small in order for the resulting bound to be useful.

The RIP and stability

We will see in Sections 1.5 and 1.6 that if a matrix A satisfies the RIP, then this is sufficient for a variety of algorithms to be able to successfully recover a sparse signal from noisy measurements. First, however, we will take a closer look at whether the RIP is actually necessary. It should be clear that the lower bound in the RIP is a necessary condition if we wish to be able to recover all sparse signals x from the measurements Ax for the same reasons that the NSP is necessary. We can say even more about the necessity of the RIP by considering the following notion of stability [67].

Definition 1.4. *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ denote a sensing matrix and $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ denote a recovery algorithm. We say that the pair (A, Δ) is C-*stable *if for any $x \in \Sigma_k$ and any $e \in \mathbb{R}^m$ we have that*

$$\left\|\Delta\left(Ax+e\right) - x\right\|_2 \leq C \left\|e\right\|_2 .$$

This definition simply says that if we add a small amount of noise to the measurements, then the impact of this on the recovered signal should not be arbitrarily large. Theorem 1.3 below demonstrates that the existence of any decoding algorithm (potentially impractical) that can stably recover from noisy measurements requires that A satisfy the lower bound of (1.7) with a constant determined by C.

Theorem 1.3 (Theorem 3.1 of [67])**.** *If the pair (A, Δ) is C-stable, then*

$$\frac{1}{C} \|x\|_2 \leq \|Ax\|_2 \tag{1.8}$$

for all $x \in \Sigma_{2k}$.

Proof. Pick any $x, z \in \Sigma_k$. Define

$$e_x = \frac{A(z-x)}{2} \qquad \text{and} \qquad e_z = \frac{A(x-z)}{2},$$

and note that

$$Ax + e_x = Az + e_z = \frac{A(x+z)}{2}.$$

Let $\widehat{x} = \Delta(Ax + e_x) = \Delta(Az + e_z)$. From the triangle inequality and the definition of C-stability, we have that

$$\begin{aligned} \|x - z\|_2 &= \|x - \widehat{x} + \widehat{x} - z\|_2 \\ &\leq \|x - \widehat{x}\|_2 + \|\widehat{x} - z\|_2 \\ &\leq C\|e_x\|_2 + C\|e_z\|_2 \\ &= C\|Ax - Az\|_2 . \end{aligned}$$

Since this holds for any $x, z \in \Sigma_k$, the result follows. □

Note that as $C \to 1$, we have that A must satisfy the lower bound of (1.7) with $\delta_k = 1 - 1/C^2 \to 0$. Thus, if we desire to reduce the impact of noise in our recovered signal then we must adjust A so that it satisfies the lower bound of (1.7) with a tighter constant.

One might respond to this result by arguing that since the upper bound is not necessary, we can avoid redesigning A simply by rescaling A so that as long as A satisfies the RIP with $\delta_{2k} < 1$, the rescaled version αA will satisfy (1.8) for any constant C. In settings where the size of the noise is independent of our choice of A, this is a valid point — by scaling A we are essentially adjusting the gain on the "signal" part of our measurements, and if increasing this gain does not impact the noise, then we can achieve arbitrarily high signal-to-noise ratios, so that eventually the noise is negligible compared to the signal.

However, in practice we will typically not be able to rescale A to be arbitrarily large. Moreover, in many practical settings the noise is not independent of A. For example, consider the case where the noise vector e represents quantization noise produced by a finite dynamic range quantizer with B bits. Suppose the measurements lie in the interval $[-T, T]$, and we have adjusted the quantizer to capture this range. If we rescale A by α, then the measurements now lie between $[-\alpha T, \alpha T]$, and we must scale the dynamic range of our quantizer by α. In this case the resulting quantization error is simply αe, and we have achieved *no reduction* in the reconstruction error.

Measurement bounds

We can also consider how many measurements are necessary to achieve the RIP. If we ignore the impact of δ and focus only on the dimensions of the problem (n, m, and k) then we can establish a simple lower bound, which is proven in Section A.1.

Theorem 1.4 (Theorem 3.5 of [67]). *Let A be an $m \times n$ matrix that satisfies the RIP of order $2k$ with constant $\delta \in (0, \frac{1}{2}]$. Then*

$$m \geq Ck \log\left(\frac{n}{k}\right)$$

where $C = 1/2 \log(\sqrt{24} + 1) \approx 0.28$.

Note that the restriction to $\delta \leq \frac{1}{2}$ is arbitrary and is made merely for convenience — minor modifications to the argument establish bounds for $\delta \leq \delta_{\max}$ for any $\delta_{\max} < 1$. Moreover, although we have made no effort to optimize the constants, it is worth noting that they are already quite reasonable.

While the proof is somewhat less direct, one can establish a similar result (in terms of its dependence on n and k) by examining the *Gelfand width* of the ℓ_1 ball [124]. However, both this result and Theorem 1.4 fail to capture the precise dependence of m on the desired RIP constant δ. In order to quantify this dependence, we can exploit recent results concerning the *Johnson-Lindenstrauss lemma*, which relates to embeddings of finite sets of points in low-dimensional spaces [158]. Specifically, it is shown in [156] that if we are given a point cloud with p points and wish to embed these points in $\mathbb{R}^m$ such that the squared ℓ_2 distance between any pair of points is preserved up to a factor of $1 \pm \epsilon$, then we must have that

$$m \geq \frac{c_0 \log(p)}{\epsilon^2},$$

where $c_0 > 0$ is a constant.

The Johnson-Lindenstrauss lemma is closely related to the RIP. In [5] it is shown that any procedure that can be used for generating a linear, distance-preserving embedding for a point cloud can also be used to construct a matrix that satisfies the RIP. Moreover, in [165] it is shown that if a matrix A satisfies the RIP of order $k = c_1 \log(p)$ with constant δ, then A can be used to construct a distance-preserving embedding for p points with $\epsilon = \delta/4$. Combining these we obtain

$$m \geq \frac{c_0 \log(p)}{\epsilon^2} = \frac{16 c_0 k}{c_1 \delta^2}.$$

Thus, for very small δ the number of measurements required to ensure that A satisfies the RIP of order k will be proportional to k/δ^2, which may be significantly higher than $k \log(n/k)$. See [165] for further details.

The relationship between the RIP and the NSP

Finally, we will now show that if a matrix satisfies the RIP, then it also satisfies the NSP. Thus, the RIP is strictly stronger than the NSP.

Theorem 1.5. *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. Then A satisfies the NSP of order $2k$ with constant*

$$C = \frac{\sqrt{2}\delta_{2k}}{1 - (1 + \sqrt{2})\delta_{2k}}.$$

The proof of this theorem involves two useful lemmas. The first of these follows directly from standard norm inequality by relating a k-sparse vector to a vector in $\mathbb{R}^k$. We include a simple proof for the sake of completeness.

Lemma 1.2. *Suppose $u \in \Sigma_k$. Then*

$$\frac{\|u\|_1}{\sqrt{k}} \leq \|u\|_2 \leq \sqrt{k} \, \|u\|_\infty .$$

Proof. For any u, $\|u\|_1 = |\langle u, \operatorname{sgn}(u) \rangle|$. By applying the Cauchy-Schwarz inequality we obtain $\|u\|_1 \leq \|u\|_2 \, \|\operatorname{sgn}(u)\|_2$. The lower bound follows since $\operatorname{sgn}(u)$ has exactly k nonzero entries all equal to ± 1 (since $u \in \Sigma_k$) and thus $\|\operatorname{sgn}(u)\|_2 = \sqrt{k}$. The upper bound is obtained by observing that each of the k nonzero entries of u can be upper bounded by $\|u\|_\infty$. □

Below we state the second key lemma that we will need in order to prove Theorem 1.5. This result is a general result which holds for arbitrary h, not just vectors $h \in \mathcal{N}(A)$. It should be clear that when we do have $h \in \mathcal{N}(A)$, the argument could be simplified considerably. However, this lemma will prove immensely useful when we turn to the problem of sparse recovery from noisy measurements in Section 1.5, and thus we establish it now in its full generality. The intuition behind this bound will become more clear after reading Section 1.5. We state the lemma here, which is proven in Section A.2.

Lemma 1.3. *Suppose that A satisfies the RIP of order $2k$, and let $h \in \mathbb{R}^n$, $h \neq 0$ be arbitrary. Let Λ_0 be any subset of $\{1, 2, \ldots, n\}$ such that $|\Lambda_0| \leq k$. Define Λ_1 as the index set corresponding to the k entries of $h_{\Lambda_0^c}$ with largest magnitude,* *and set* $\Lambda = \Lambda_0 \cup \Lambda_1$. *Then*

$$\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{k}} + \beta \frac{|\langle Ah_\Lambda, Ah \rangle|}{\|h_\Lambda\|_2},$$

where

$$\alpha = \frac{\sqrt{2}\delta_{2k}}{1-\delta_{2k}}, \quad \beta = \frac{1}{1-\delta_{2k}}.$$

Again, note that Lemma 1.3 holds for arbitrary h. In order to prove Theorem 1.5, we merely need to apply Lemma 1.3 to the case where $h \in \mathcal{N}(A)$.

Proof of Theorem 1.5. Suppose that $h \in \mathcal{N}(A)$. It is sufficient to show that

$$\|h_\Lambda\|_2 \leq C \frac{\|h_{\Lambda^c}\|_1}{\sqrt{k}} \tag{1.9}$$

holds for the case where Λ is the index set corresponding to the $2k$ largest entries of h. Thus, we can take Λ_0 to be the index set corresponding to the k largest entries of h and apply Lemma 1.3.

The second term in Lemma 1.3 vanishes since $Ah = 0$, and thus we have

$$\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{k}}.$$

Using Lemma 1.2,

$$\|h_{\Lambda_0^c}\|_1 = \|h_{\Lambda_1}\|_1 + \|h_{\Lambda^c}\|_1 \leq \sqrt{k}\,\|h_{\Lambda_1}\|_2 + \|h_{\Lambda^c}\|_1$$

resulting in

$$\|h_\Lambda\|_2 \leq \alpha \left(\|h_{\Lambda_1}\|_2 + \frac{\|h_{\Lambda^c}\|_1}{\sqrt{k}} \right).$$

Since $\|h_{\Lambda_1}\|_2 \leq \|h_\Lambda\|_2$, we have that

$$(1-\alpha)\,\|h_\Lambda\|_2 \leq \alpha \frac{\|h_{\Lambda^c}\|_1}{\sqrt{k}}.$$

The assumption $\delta_{2k} < \sqrt{2} - 1$ ensures that $\alpha < 1$, and thus we may divide by $1 - \alpha$ without changing the direction of the inequality to establish (1.9) with constant

$$C = \frac{\alpha}{1-\alpha} = \frac{\sqrt{2}\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}},$$

as desired. □

1.4.3 Coherence

While the spark, NSP, and RIP all provide guarantees for the recovery of k-sparse signals, verifying that a general matrix A satisfies any of these properties has a combinatorial computational complexity, since in each case one must essentially consider $\binom{n}{k}$ submatrices. In many cases it is preferable to use properties of A that are easily computable to provide more concrete recovery guarantees. The *coherence* of a matrix is one such property [86, 222].

Definition 1.5. *The coherence of a matrix A, $\mu(A)$, is the largest absolute inner product between any two columns a_i, a_j of A:*

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

It is possible to show that the coherence of a matrix is always in the range $\mu(A) \in \left[\sqrt{\frac{n-m}{m(n-1)}}, 1\right]$; the lower bound is known as the Welch bound [207, 214, 245]. Note that when $n \gg m$, the lower bound is approximately $\mu(A) \ge 1/\sqrt{m}$. The concept of coherence can also be extended to certain structured sparsity models and specific classes of analog signals [27, 111, 112].

One can sometimes relate coherence to the spark, NSP, and RIP. For example, the coherence and spark properties of a matrix can be related by employing the Gershgorin circle theorem [127, 235].

Theorem 1.6 (Theorem 2 of [127])**.** *The eigenvalues of an $n \times n$ matrix M with entries m_{ij}, $1 \le i, j \le n$, lie in the union of n discs $d_i = d_i(c_i, r_i)$, $1 \le i \le n$, centered at $c_i = m_{ii}$ and with radius $r_i = \sum_{j \ne i} |m_{ij}|$.*

Applying this theorem on the Gram matrix $G = A_\Lambda^T A_\Lambda$ leads to the following straightforward result.

Lemma 1.4. *For any matrix A,*

$$\text{spark}(A) \ge 1 + \frac{1}{\mu(A)}.$$

Proof. Since spark(A) does not depend on the scaling of the columns, we can assume without loss of generality that A has unit-norm columns. Let $\Lambda \subseteq \{1, \dots, n\}$ with $|\Lambda| = p$ determine a set of indices. We consider the restricted Gram matrix $G = A_\Lambda^T A_\Lambda$, which satisfies the following properties:

- $g_{ii} = 1$, $1 \le i \le p$;
- $|g_{ij}| \le \mu(A)$, $1 \le i, j \le p$, $i \ne j$.

From Theorem 1.6, if $\sum_{j \ne i} |g_{ij}| < |g_{ii}|$ then the matrix G is positive definite, so that the columns of A_Λ are linearly independent. Thus, the spark condition implies $(p-1)\mu(A) < 1$ or, equivalently, $p < 1 + 1/\mu(A)$ for all $p < \text{spark}(A)$, yielding spark$(A) \ge 1 + 1/\mu(A)$. □

By merging Theorem 1.1 with Lemma 1.4, we can pose the following condition on A that guarantees uniqueness.

Theorem 1.7 (Theorem 12 of [86]). *If*

$$k < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right),$$

then for each measurement vector $y \in \mathbb{R}^m$ there exists at most one signal $x \in \Sigma_k$ such that $y = Ax$.

Theorem 1.7, together with the Welch bound, provides an upper bound on the level of sparsity k that guarantees uniqueness using coherence: $k = O(\sqrt{m})$. Another straightforward application of the Gershgorin circle theorem (Theorem 1.6) connects the RIP to the coherence property.

Lemma 1.5. *If A has unit-norm columns and coherence $\mu = \mu(A)$, then A satisfies the RIP of order k with $\delta = k\mu$ for all $k < 1/\mu$.*

The proof of this lemma is similar to that of Lemma 1.4.

1.4.4 Sensing matrix constructions

Now that we have defined the relevant properties of a matrix A in the context of CS, we turn to the question of how to construct matrices that satisfy these properties. To begin, it is straightforward to show that an $m \times n$ Vandermonde matrix V constructed from m distinct scalars has $\operatorname{spark}(V) = m + 1$ [57]. Unfortunately, these matrices are poorly conditioned for large values of n, rendering the recovery problem numerically unstable. Similarly, there are known matrices A of size $m \times m^2$ that achieve the coherence lower bound $\mu(A) = 1/\sqrt{m}$, such as the Gabor frame generated from the Alltop sequence [148] and more general equiangular tight frames [214]. These constructions restrict the number of measurements needed to recover a k-sparse signal to be $m = O(k^2 \log n)$. It is also possible to deterministically construct matrices of size $m \times n$ that satisfy the RIP of order k, but such constructions also require m to be relatively large [28, 78, 140, 152]. For example, the construction in [78] requires $m = O(k^2 \log n)$ while the construction in [152] requires $m = O(kn^\alpha)$ for some constant α. In many real-world settings, these results would lead to an unacceptably large requirement on m.

Fortunately, these limitations can be overcome by randomizing the matrix construction. For example, random matrices A of size $m \times n$ whose entries are independent and identically distributed (i.i.d.) with continuous distributions have $\operatorname{spark}(A) = m + 1$ with probability one. More significantly, it can also be shown that random matrices will satisfy the RIP with high probability if the entries are chosen according to a Gaussian, Bernoulli, or more generally any sub-gaussian distribution. See Chapter 5 for details, and in particular, Theorem 5.65. This theorem states that if a matrix A is chosen according to a sub-gaussian distribution with $m = O\left(k \log(n/k)/\delta_{2k}^2\right)$, then A will satisfy the RIP of order $2k$ with probability at least $1 - 2\exp(-c_1\delta^2 m)$. Note that in light of the measurement bounds in Section 1.4.2 we see that this achieves the optimal number of measurements up to a constant. It also follows from Theorem 1.5 that these random constructions provide matrices satisfying the NSP. Furthermore, it can be shown that when the distribution used has zero mean and finite variance, then in the asymptotic regime (as m and n grow) the coherence converges to $\mu(A) = \sqrt{(2\log n)/m}$ [32, 37, 83].

Using random matrices to construct A has a number of additional benefits. To illustrate these, we will focus on the RIP. First, one can show that for random constructions the measurements are *democratic*, meaning that it is possible to recover a signal using any sufficiently large subset of the measurements [73, 169]. Thus, by using random A one can be robust to the loss or corruption of a small fraction of the measurements. Second, and perhaps more significantly, in practice we are often more interested in the setting where x is sparse with respect to some basis Φ. In this case what we actually require is that the product $A\Phi$ satisfies the RIP. If we were to use a deterministic construction then we would need to explicitly take Φ into account in our construction of A, but when A is chosen randomly we can avoid this consideration. For example, if A is chosen according to a Gaussian distribution and Φ is an orthonormal basis then one can easily show that $A\Phi$ will also have a Gaussian distribution, and so provided that m is sufficiently high $A\Phi$ will satisfy the RIP with high probability, just as before. Although less obvious, similar results hold for sub-gaussian distributions as well [5]. This property, sometimes referred to as *universality*, constitutes a significant advantage of using random matrices to construct A. See Chapter 5 for further details on random matrices and their role in CS.

Finally, we note that since the fully random matrix approach is sometimes impractical to build in hardware, several hardware architectures have been implemented and/or proposed that enable random measurements to be acquired in practical settings. Examples include the random demodulator [224], random filtering [225], the modulated wideband converter [187], random convolution [1, 206], and the compressive multiplexer [211]. These architectures typically use a reduced amount of randomness and are modeled via matrices A that have significantly more structure than a fully random matrix. Perhaps somewhat surprisingly, while it is typically not quite as easy as in the fully random case, one can prove that many of these constructions also satisfy the RIP and/or have low coherence. Furthermore, one can analyze the effect of inaccuracies in the matrix A implemented by the system [54, 149]; in the simplest cases, such sensing matrix errors can be addressed through system calibration.

1.5 Signal Recovery via ℓ_1 Minimization

While there now exist a wide variety of approaches to recover a sparse signal x from a small number of linear measurements, as we will see in Section 1.6, we begin by considering a natural first approach to the problem of sparse recovery.

Given measurements y and the knowledge that our original signal x is sparse or compressible, it is natural to attempt to recover x by solving an optimization problem of the form

$$\widehat{x} = \arg\min_{z} \ \|z\|_0 \quad \text{subject to} \quad z \in \mathcal{B}(y), \tag{1.10}$$

where $\mathcal{B}(y)$ ensures that $\widehat{x}$ is consistent with the measurements y. For example, in the case where our measurements are exact and noise-free, we can set $\mathcal{B}(y) = \{z : Az = y\}$. When the measurements have been contaminated with a small amount of bounded noise, we could instead consider $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq \epsilon\}$. In both cases, (1.10) finds the sparsest x that is consistent with the measurements y.

Note that in (1.10) we are inherently assuming that x itself is sparse. In the more common setting where $x = \Phi c$, we can easily modify the approach and instead consider

$$\widehat{c} = \arg\min_{z} \ \|z\|_0 \quad \text{subject to} \quad z \in \mathcal{B}(y) \tag{1.11}$$

where $\mathcal{B}(y) = \{z : A\Phi z = y\}$ or $\mathcal{B}(y) = \{z : \|A\Phi z - y\|_2 \leq \epsilon\}$. By considering $\widetilde{A} = A\Phi$ we see that (1.10) and (1.11) are essentially identical. Moreover, as noted in Section 1.4.4, in many cases the introduction of Φ does not significantly complicate the construction of matrices A such that $\widetilde{A}$ will satisfy the desired properties. Thus, for the remainder of this chapter we will restrict our attention to the case where $\Phi = I$. It is important to note, however, that this restriction does impose certain limits in our analysis when Φ is a general dictionary and not an orthonormal basis. For example, in this case $\|\widehat{x} - x\|_2 = \|\Phi\widehat{c} - \Phi c\|_2 \neq \|\widehat{c} - c\|_2$, and thus a bound on $\|\widehat{c} - c\|_2$ cannot directly be translated into a bound on $\|\widehat{x} - x\|_2$, which is often the metric of interest. For further discussion of these and related issues see [35].

While it is possible to analyze the performance of (1.10) under the appropriate assumptions on A (see [56, 144] for details), we do not pursue this strategy since the objective function $\|\cdot\|_0$ is nonconvex, and hence (1.10) is potentially very difficult to solve. In fact, one can show that for a general matrix A, even finding a solution that approximates the true minimum is NP-hard [189].

One avenue for translating this problem into something more tractable is to replace $\|\cdot\|_0$ with its convex approximation $\|\cdot\|_1$. Specifically, we consider

$$\widehat{x} = \arg\min_{z} \ \|z\|_1 \quad \text{subject to} \quad z \in \mathcal{B}(y). \tag{1.12}$$

Provided that $\mathcal{B}(y)$ is convex, (1.12) is computationally feasible. In fact, when $\mathcal{B}(y) = \{z : Az = y\}$, the resulting problem can be posed as a linear program [53].

While it is clear that replacing (1.10) with (1.12) transforms a computationally intractable problem into a tractable one, it may not be immediately obvious that the solution to (1.12) will be at all similar to the solution to (1.10). However, there are certainly intuitive reasons to expect that the use of ℓ_1 minimization will indeed promote sparsity. As an example, recall that in Fig. 1.2, the solutions to the ℓ_1 minimization problem coincided exactly with the solution to the ℓ_p

minimization problem for any $p < 1$, and notably, was sparse. Moreover, the use of ℓ_1 minimization to promote or exploit sparsity has a long history, dating back at least to the work of Beurling on Fourier transform extrapolation from partial observations [22].

Additionally, in a somewhat different context, in 1965 Logan [91, 174] showed that a bandlimited signal can be perfectly recovered in the presence of *arbitrary* corruptions on a small interval (see also extensions of these conditions in [91]). Again, the recovery method consists of searching for the bandlimited signal that is closest to the observed signal in the ℓ_1 norm. This can be viewed as further validation of the intuition gained from Fig. 1.2 — the ℓ_1 norm is well-suited to sparse errors.

Historically, the use of ℓ_1 minimization on large problems finally became practical with the explosion of computing power in the late 1970's and early 1980's. In one of its first applications, it was demonstrated that geophysical signals consisting of spike trains could be recovered from only the high-frequency components of these signals by exploiting ℓ_1 minimization [171, 216, 242]. Finally, in the 1990's there was renewed interest in these approaches within the signal processing community for the purpose of finding sparse approximations to signals and images when represented in overcomplete dictionaries or unions of bases [53, 182]. Separately, ℓ_1 minimization received significant attention in the statistics literature as a method for variable selection in regression, known as the Lasso [218].

Thus, there are a variety of reasons to suspect that ℓ_1 minimization will provide an accurate method for sparse signal recovery. More importantly, this also constitutes a computationally tractable approach to sparse signal recovery. In this section we provide an overview of ℓ_1 minimization from a theoretical perspective. We discuss algorithms for ℓ_1 minimization in Section 1.6.

1.5.1 Noise-free signal recovery

In order to analyze ℓ_1 minimization algorithms for various specific choices of $\mathcal{B}(y)$, we require the following general result which builds on Lemma 1.3 and is proven in Section A.3.

Lemma 1.6. *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. Let $x, \widehat{x} \in \mathbb{R}^n$ be given, and define $h = \widehat{x} - x$. Let Λ_0 denote the index set corresponding to the k entries of x with largest magnitude and Λ_1 the index set corresponding to the k entries of $h_{\Lambda_0^c}$ with largest magnitude. Set $\Lambda = \Lambda_0 \cup \Lambda_1$. If $\|\widehat{x}\|_1 \leq \|x\|_1$, then*

$$\|h\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_1 \frac{|\langle Ah_\Lambda, Ah \rangle|}{\|h_\Lambda\|_2}.$$

where

$$C_0 = 2\frac{1 - (1 - \sqrt{2})\delta_{2k}}{1 - (1 + \sqrt{2})\delta_{2k}}, \quad C_1 = \frac{2}{1 - (1 + \sqrt{2})\delta_{2k}}.$$

Lemma 1.6 establishes an error bound for the class of ℓ_1 minimization algorithms described by (1.12) when combined with a measurement matrix A satisfying the RIP. In order to obtain specific bounds for concrete examples of $\mathcal{B}(y)$, we must examine how requiring $\widehat{x} \in \mathcal{B}(y)$ affects $|\langle Ah_\Lambda, Ah \rangle|$. As an example, in the case of noise-free measurements we obtain the following theorem.

Theorem 1.8 (Theorem 1.1 of [34]). *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$ and we obtain measurements of the form $y = Ax$. Then when $\mathcal{B}(y) = \{z : Az = y\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|\widehat{x} - x\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}}.$$

Proof. Since $x \in \mathcal{B}(y)$ we can apply Lemma 1.6 to obtain that for $h = \widehat{x} - x$,

$$\|h\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_1 \frac{|\langle Ah_\Lambda, Ah \rangle|}{\|h_\Lambda\|_2}.$$

Furthermore, since $x, \widehat{x} \in \mathcal{B}(y)$ we also have that $y = Ax = A\widehat{x}$ and hence $Ah = 0$. Therefore the second term vanishes, and we obtain the desired result. □

Theorem 1.8 is rather remarkable. By considering the case where $x \in \Sigma_k$ we can see that provided A satisfies the RIP — which as shown in Section 1.4.4 allows for as few as $O(k \log(n/k))$ measurements — we can recover any k-sparse x *exactly.* This result seems improbable on its own, and so one might expect that the procedure would be highly sensitive to noise, but we will see below that Lemma 1.6 can also be used to demonstrate that this approach is actually stable.

Note that Theorem 1.8 assumes that A satisfies the RIP. One could easily modify the argument to replace this with the assumption that A satisfies the NSP instead. Specifically, if we are only interested in the noiseless setting, in which case h lies in the nullspace of A, then Lemma 1.6 simplifies and its proof could essentially be broken into two steps: (i) show that if A satisfies the RIP then it satisfies the NSP (as shown in Theorem 1.5), and (ii) the NSP implies the simplified version of Lemma 1.6. This proof directly mirrors that of Lemma 1.6. Thus, by the same argument as in the proof of Theorem 1.8, it is straightforward to show that if A satisfies the NSP then it will obey the same error bound.

1.5.2 Signal recovery in noise

The ability to perfectly reconstruct a sparse signal from noise-free measurements represents a very promising result. However, in most real-world systems the measurements are likely to be contaminated by some form of noise. For instance, in order to process data in a computer we must be able to represent it using a finite number of bits, and hence the measurements will typically be subject to quantization error. Moreover, systems which are implemented in physical hardware will be subject to a variety of different types of noise depending on the setting. Another important noise source is on the signal itself. In many settings the signal x to be estimated is contaminated by some form of random noise. The implications of this type of noise on the achievable sampling rates has been recently analyzed in [19, 67, 219]. Here we focus on measurement noise, which has received much more attention in the literature.

Perhaps somewhat surprisingly, one can show that it is possible to stably recover sparse signals under a variety of common noise models [18, 42, 87, 88, 144, 169, 170]. As might be expected, both the RIP and coherence are useful in establishing performance guarantees in noise. We begin our discussion below with robustness guarantees for matrices satisfying the RIP. We then turn to results for matrices with low coherence.

Bounded noise

We first provide a bound on the worst-case performance for uniformly bounded noise, as first investigated in [42].

Theorem 1.9 (Theorem 1.2 of [34]). *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$ and let $y = Ax + e$ where $\|e\|_2 \leq \epsilon$. Then when $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq \epsilon\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|\widehat{x} - x\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_2 \epsilon,$$

where

$$C_0 = 2\frac{1 - (1 - \sqrt{2})\delta_{2k}}{1 - (1 + \sqrt{2})\delta_{2k}}, \quad C_2 = 4\frac{\sqrt{1 + \delta_{2k}}}{1 - (1 + \sqrt{2})\delta_{2k}}.$$

Proof. We are interested in bounding $\|h\|_2 = \|\widehat{x} - x\|_2$. Since $\|e\|_2 \leq \epsilon$, $x \in \mathcal{B}(y)$, and therefore we know that $\|\widehat{x}\|_1 \leq \|x\|_1$. Thus we may apply Lemma 1.6, and it remains to bound $|\langle Ah_\Lambda, Ah \rangle|$. To do this, we observe that

$$\|Ah\|_2 = \|A(\widehat{x} - x)\|_2 = \|A\widehat{x} - y + y - Ax\|_2 \leq \|A\widehat{x} - y\|_2 + \|y - Ax\|_2 \leq 2\epsilon$$

where the last inequality follows since $x, \widehat{x} \in \mathcal{B}(y)$. Combining this with the RIP and the Cauchy-Schwarz inequality we obtain

$$|\langle Ah_\Lambda, Ah \rangle| \leq \|Ah_\Lambda\|_2 \|Ah\|_2 \leq 2\epsilon\sqrt{1 + \delta_{2k}} \|h_\Lambda\|_2 .$$

Thus,

$$\|h\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_1 2\epsilon\sqrt{1 + \delta_{2k}} = C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_2 \epsilon,$$

completing the proof. □

In order to place this result in context, consider how we would recover a sparse vector x if we happened to already know the k locations of the nonzero coefficients, which we denote by Λ_0. This is referred to as the *oracle estimator*. In this case a natural approach is to reconstruct the signal using a simple pseudoinverse:[7]

$$\begin{aligned} \widehat{x}_{\Lambda_0} &= A^{\dagger}_{\Lambda_0} y = (A^T_{\Lambda_0} A_{\Lambda_0})^{-1} A^T_{\Lambda_0} y \\ \widehat{x}_{\Lambda_0^c} &= 0. \end{aligned} \tag{1.13}$$

The implicit assumption in (1.13) is that A_{Λ_0} has full column-rank (and hence we are considering the case where A_{Λ_0} is the $m \times k$ matrix with the columns indexed by Λ_0^c removed) so that there is a unique solution to the equation $y = A_{\Lambda_0} x_{\Lambda_0}$. With this choice, the recovery error is given by

$$\|\widehat{x} - x\|_2 = \left\|(A^T_{\Lambda_0} A_{\Lambda_0})^{-1} A^T_{\Lambda_0}(Ax + e) - x\right\|_2 = \left\|(A^T_{\Lambda_0} A_{\Lambda_0})^{-1} A^T_{\Lambda_0} e\right\|_2.$$

We now consider the worst-case bound for this error. Using standard properties of the singular value decomposition, it is straightforward to show that if A satisfies the RIP of order $2k$ (with constant δ_{2k}), then the largest singular value of $A^{\dagger}_{\Lambda_0}$ lies in the range $[1/\sqrt{1+\delta_{2k}}, 1/\sqrt{1-\delta_{2k}}]$. Thus, if we consider the worst-case recovery error over all e such that $\|e\|_2 \le \epsilon$, then the recovery error can be bounded by

$$\frac{\epsilon}{\sqrt{1+\delta_{2k}}} \le \|\widehat{x} - x\|_2 \le \frac{\epsilon}{\sqrt{1-\delta_{2k}}}.$$

Therefore, in the case where x is exactly k-sparse, the guarantee for the pseudoinverse recovery method, which is given *perfect knowledge of the true support of* x, cannot improve upon the bound in Theorem 1.9 by more than a constant value.

We now consider a slightly different noise model. Whereas Theorem 1.9 assumed that the noise norm $\|e\|_2$ was small, the theorem below analyzes a different recovery algorithm known as the *Dantzig selector* in the case where $\left\|A^T e\right\|_\infty$ is small [45]. We will see below that this will lead to a simple analysis of the performance of this algorithm in Gaussian noise.

Theorem 1.10. *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$ and we obtain measurements of the form $y = Ax + e$ where $\left\|A^T e\right\|_\infty \le \lambda$. Then when $\mathcal{B}(y) = \{z : \left\|A^T(Az - y)\right\|_\infty \le \lambda\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|\widehat{x} - x\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_3 \sqrt{k} \lambda,$$

where

$$C_0 = 2\frac{1 - (1-\sqrt{2})\delta_{2k}}{1 - (1+\sqrt{2})\delta_{2k}}, \quad C_3 = \frac{4\sqrt{2}}{1 - (1+\sqrt{2})\delta_{2k}}.$$

[7] Note that while the pseudoinverse approach can be improved upon (in terms of ℓ_2 error) by instead considering alternative biased estimators [16, 108, 155, 159, 213], this does not fundamentally change the above conclusions.

Proof. The proof mirrors that of Theorem 1.9. Since $\left\|A^T e\right\|_\infty \leq \lambda$, we again have that $x \in \mathcal{B}(y)$, so $\|\widehat{x}\|_1 \leq \|x\|_1$ and thus Lemma 1.6 applies. We follow a similar approach as in Theorem 1.9 to bound $|\langle Ah_\Lambda, Ah\rangle|$. We first note that

$$\left\|A^T Ah\right\|_\infty \leq \left\|A^T(A\widehat{x} - y)\right\|_\infty + \left\|A^T(y - Ax)\right\|_\infty \leq 2\lambda$$

where the last inequality again follows since $x, \widehat{x} \in \mathcal{B}(y)$. Next, note that $Ah_\Lambda = A_\Lambda h_\Lambda$. Using this we can apply the Cauchy-Schwarz inequality to obtain

$$|\langle Ah_\Lambda, Ah\rangle| = \left|\langle h_\Lambda, A_\Lambda^T Ah\rangle\right| \leq \|h_\Lambda\|_2 \left\|A_\Lambda^T Ah\right\|_2.$$

Finally, since $\left\|A^T Ah\right\|_\infty \leq 2\lambda$, we have that every coefficient of $A^T Ah$ is at most 2λ, and thus $\left\|A_\Lambda^T Ah\right\|_2 \leq \sqrt{2k}(2\lambda)$. Thus,

$$\|h\|_2 \leq C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_1 2\sqrt{2k}\lambda = C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_3\sqrt{k}\lambda,$$

as desired. □

Gaussian noise

Finally, we also consider the performance of these approaches in the presence of Gaussian noise. The case of Gaussian noise was first considered in [144], which examined the performance of ℓ_0 minimization with noisy measurements. We now see that Theorems 1.9 and 1.10 can be leveraged to provide similar guarantees for ℓ_1 minimization. To simplify our discussion we will restrict our attention to the case where $x \in \Sigma_k$, so that $\sigma_k(x)_1 = 0$ and the error bounds in Theorems 1.9 and 1.10 depend only on the noise e.

To begin, suppose that the coefficients of $e \in \mathbb{R}^m$ are i.i.d. according to a Gaussian distribution with mean zero and variance σ^2. By using standard properties of the Gaussian distribution, one can show (see, for example, Corollary 5.17 of Chapter 5) that there exists a constant $c_0 > 0$ such that for any $\epsilon > 0$,

$$\mathbb{P}\left(\|e\|_2 \geq (1+\epsilon)\sqrt{m}\sigma\right) \leq \exp\left(-c_0\epsilon^2 m\right), \tag{1.14}$$

where $\mathbb{P}(E)$ denotes the probability that the event E occurs. Applying this result to Theorem 1.9 with $\epsilon = 1$, we obtain the following result for the special case of Gaussian noise.

Corollary 1.1. *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. Furthermore, suppose that $x \in \Sigma_k$ and that we obtain measurements of the form $y = Ax + e$ where the entries of e are i.i.d. $\mathcal{N}(0, \sigma^2)$. Then when $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq 2\sqrt{m}\sigma\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|\widehat{x} - x\|_2 \leq 8 \frac{\sqrt{1+\delta_{2k}}}{1 - (1+\sqrt{2})\delta_{2k}} \sqrt{m}\sigma$$

with probability at least $1 - \exp(-c_0 m)$.

We can similarly consider Theorem 1.10 in the context of Gaussian noise. If we assume that the columns of A have unit norm, then each coefficient of $A^T e$ is a Gaussian random variable with mean zero and variance σ^2. Using standard tail bounds for the Gaussian distribution (see, for example, (5.5) of Chapter 5), we have that

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

for $i = 1, 2, \ldots, n$. Thus, using the union bound over the bounds for different i, we obtain

$$\mathbb{P}\left(\left\|A^T e\right\|_\infty \geq 2\sqrt{\log n}\sigma\right) \leq n \exp\left(-2\log n\right) = \frac{1}{n}.$$

Applying this to Theorem 1.10, we obtain the following result, which is a simplified version of Theorem 1.1 of [45].

Corollary 1.2. *Suppose that A has unit-norm columns and satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2} - 1$. Furthermore, suppose that $x \in \Sigma_k$ and that we obtain measurements of the form $y = Ax + e$ where the entries of e are i.i.d. $\mathcal{N}(0, \sigma^2)$. Then when $\mathcal{B}(y) = \{z : \left\|A^T(Az - y)\right\|_\infty \leq 2\sqrt{\log n}\sigma\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\left\|\widehat{x} - x\right\|_2 \leq 4\sqrt{2}\frac{\sqrt{1+\delta_{2k}}}{1-(1+\sqrt{2})\delta_{2k}}\sqrt{k\log n}\sigma$$

with probability at least $1 - \frac{1}{n}$.

Ignoring the precise constants and the probabilities with which the stated bounds hold (which we have made no effort to optimize), we observe that in the case when $m = O(k \log n)$ these results appear to be essentially the same. However, there is a subtle difference. Specifically, if m and n are fixed and we consider the effect of varying k, we can see that Corollary 1.2 yields a bound that is adaptive to this change, providing a stronger guarantee when k is small, whereas the bound in Corollary 1.1 does not improve as k is reduced. Thus, while they provide very similar guarantees, there are certain circumstances where the Dantzig selector is preferable. See [45] for further discussion of the comparative advantages of these approaches.

It can also be seen that results such as Corollary 1.2 guarantee that the Dantzig selector achieves an error $\left\|\widehat{x} - x\right\|_2^2$ which is bounded by a constant times $k\sigma^2 \log n$, with high probability. Note that since we typically require $m > k \log n$, this can be substantially lower than the expected noise power $\mathbb{E}\|e\|_2^2 = m\sigma^2$, illustrating the fact that sparsity-based techniques are highly successful in reducing the noise level.

The value $k\sigma^2 \log n$ is nearly optimal in several respects. First, an “oracle” estimator which knows the locations of the nonzero components and uses a least-squares technique to estimate their values achieves an estimation error on the order of $k\sigma^2$. For this reason, guarantees such as Corollary 1.2 are referred to as near-oracle results. The Cramer-Rao bound (CRB) for estimating x is also on the order of $k\sigma^2$ [17]. This is of practical interest since the CRB is achieved by the maximum likelihood estimator at high SNR, implying that for low-noise settings, an error of $k\sigma^2$ is achievable. However, the maximum likelihood estimator is NP-hard to compute, so that near-oracle results are still of interest. Interestingly, the $\log n$ factor is an unavoidable result of the fact that the locations of the nonzero elements are unknown.

Coherence guarantees

Thus far, we have examined performance guarantees based on the RIP. As noted in Section 1.4.3, in practice it is typically impossible to verify that a matrix A satisfies the RIP or calculate the corresponding RIP constant δ. In this respect, results based on coherence are appealing, since they can be used with arbitrary dictionaries.

One quick route to coherence-based performance guarantees is to combine RIP-based results such as Corollaries 1.1 and 1.2 with coherence bounds such as Lemma 1.5. This technique yields guarantees based only on the coherence, but the results are often overly pessimistic. It is typically more enlightening to instead establish guarantees by directly exploiting coherence [18, 37, 87, 88]. In order to illustrate the types of guarantees that this approach can yield, we provide the following representative examples.

Theorem 1.11 (Theorem 3.1 of [88]). *Suppose that A has coherence μ and that $x \in \Sigma_k$ with $k < (1/\mu + 1)/4$. Furthermore, suppose that we obtain measurements of the form $y = Ax + e$. Then when $\mathcal{B}(y) = \{z : \|Az - y\|_2 \le \epsilon\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|x - \widehat{x}\|_2 \le \frac{\|e\|_2 + \epsilon}{\sqrt{1 - \mu(4k-1)}}.$$

Note that this theorem holds for the case where $\epsilon = 0$ as well as where $\|e\|_2 = 0$. Thus, it also applies to the noise-free setting as in Theorem 1.8. Furthermore, there is no requirement that $\|e\|_2 \le \epsilon$. In fact, this theorem is valid even when $\epsilon = 0$ but $\|e\|_2 \ne 0$. This constitutes a significant difference between this result and Theorem 1.9, and might cause us to question whether we actually need to pose alternative algorithms to handle the noisy setting. However, as noted in [88], Theorem 1.11 is the result of a worst-case analysis and will typically overestimate the actual error. In practice, the performance of (1.12) where $\mathcal{B}(y)$ is modified to account for the noise can lead to significant improvements.

In order to describe an additional type of coherence-based guarantee, we must consider an alternative, but equivalent, formulation of (1.12). Specifically, consider the optimization problem

$$\widehat{x} = \arg\min_z \frac{1}{2} \|Az - y\|_2^2 + \lambda \|z\|_1 .$$

This formulation is exploited in the following result, which provides guarantees for (1.5.2) that go beyond what we have seen so far by providing explicit results concerning the recovery of the original support of x.

Theorem 1.12 (Corollary 1 of [18])**.** *Suppose that A has coherence μ and that $x \in \Sigma_k$ with $k \leq 1/(3\mu)$. Furthermore, suppose that we obtain measurements of the form $y = Ax + e$ where the entries of e are i.i.d. $\mathcal{N}(0, \sigma^2)$. Set*

$$\lambda = \sqrt{8\sigma^2(1+\alpha)\log(n-k)}$$

for some fairly small value $\alpha > 0$. Then with probability exceeding

$$\left(1 - \frac{1}{(n-k)^\alpha}\right)(1 - \exp(-k/7)),$$

the solution $\widehat{x}$ to (1.5.2) is unique, $\operatorname{supp}(\widehat{x}) \subset \operatorname{supp}(x)$*, and*

$$\|\widehat{x} - x\|_2^2 \leq \left(\sqrt{3} + 3\sqrt{2(1+\alpha)\log(n-k)}\right)^2 k\sigma^2.$$

In this case we see that we are guaranteed that any nonzero of $\widehat{x}$ corresponds to a true nonzero of x. Note that this analysis allows for the worst-case signal x. It is possible to improve upon this result by instead assuming that the signal x has a limited amount of randomness. Specifically, in [37] it is shown that if $\operatorname{supp}(x)$ is chosen uniformly at random and that the signs of the nonzero entries of x are independent and equally likely to be ± 1, then it is possible to significantly relax the assumption on μ. Moreover, by requiring the nonzeros of x to exceed some minimum magnitude one can also guarantee perfect recovery of the true support.

1.5.3 Instance-optimal guarantees revisited

We now briefly return to the noise-free setting to take a closer look at instance-optimal guarantees for recovering non-sparse signals. To begin, recall that in Theorem 1.8 we bounded the ℓ_2-norm of the reconstruction error $\|\widehat{x} - x\|_2$ by a constant C_0 times $\sigma_k(x)_1/\sqrt{k}$. One can generalize this result to measure the reconstruction error using the ℓ_p-norm for any $p \in [1, 2]$. For example, by a slight modification of these arguments, one can also show that $\|\widehat{x} - x\|_1 \leq C_0\sigma_k(x)_1$ (see [34]). This leads us to ask whether we might replace the bound for the ℓ_2 error with a result of the form $\|\widehat{x} - x\|_2 \leq C\sigma_k(x)_2$. Unfortunately, obtaining such a result requires an unreasonably large number of measurements, as quantified by the following theorem of [57], proven in Section A.4.

Theorem 1.13 (Theorem 5.1 of [57])**.** *Suppose that A is an $m \times n$ matrix and that $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is a recovery algorithm that satisfies*

$$\|x - \Delta(Ax)\|_2 \leq C\sigma_k(x)_2 \tag{1.15}$$

for some $k \geq 1$, then $m > \left(1 - \sqrt{1 - 1/C^2}\right) n$.

Thus, if we want a bound of the form (1.15) that holds for *all* signals x with a constant $C \approx 1$, then regardless of what recovery algorithm we use we will need to take $m \approx n$ measurements. However, in a sense this result is overly pessimistic, and we will now see that the results from Section 1.5.2 can actually allow us to overcome this limitation by essentially treating the approximation error as noise.

Towards this end, notice that all the results concerning ℓ_1 minimization stated thus far are deterministic instance-optimal guarantees that apply simultaneously to all x given any matrix that satisfies the RIP. This is an important theoretical property, but as noted in Section 1.4.4, in practice it is very difficult to obtain a deterministic guarantee that the matrix A satisfies the RIP. In particular, constructions that rely on randomness are only known to satisfy the RIP with high probability. As an example, recall that Theorem 5.65 of Chapter 5 states that if a matrix A is chosen according to a sub-gaussian distribution with $m = O\left(k \log(n/k)/\delta_{2k}^2\right)$, then A will satisfy the RIP of order $2k$ with probability at least $1 - 2\exp(-c_1 \delta^2 m)$. Results of this kind open the door to slightly weaker results that hold only with high probability.

Even within the class of probabilistic results, there are two distinct flavors. The typical approach is to combine a probabilistic construction of a matrix that will satisfy the RIP with high probability with the previous results in this chapter. This yields a procedure that, with high probability, will satisfy a deterministic guarantee applying to all possible signals x. A weaker kind of result is one that states that given a signal x, we can draw a random matrix A and with high probability expect certain performance *for that signal x*. This type of guarantee is sometimes called *instance-optimal in probability*. The distinction is essentially whether or not we need to draw a new random A for each signal x. This may be an important distinction in practice, but if we assume for the moment that it is permissible to draw a new matrix A for each x, then we can see that Theorem 1.13 may be somewhat pessimistic, exhibited by the following result.

Theorem 1.14. *Let $x \in \mathbb{R}^n$ be fixed. Set $\delta_{2k} < \sqrt{2} - 1$ Suppose that A is an $m \times n$ sub-gaussian random matrix with $m = O\left(k \log(n/k)/\delta_{2k}^2\right)$. Suppose we obtain measurements of the form $y = Ax$. Set $\epsilon = 2\sigma_k(x)_2$. Then with probability exceeding $1 - 2\exp(-c_1 \delta^2 m) - \exp(-c_0 m)$, when $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq \epsilon\}$, the solution $\widehat{x}$ to (1.12) obeys*

$$\|\hat{x} - x\|_2 \leq \frac{8\sqrt{1+\delta_{2k}} - (1+\sqrt{2})\delta_{2k}}{1 - (1+\sqrt{2})\delta_{2k}} \sigma_k(x)_2.$$

Proof. First we recall that, as noted above, from Theorem 5.65 of Chapter 5 we have that A will satisfy the RIP of order $2k$ with probability at least $1 - 2\exp(-c_1 \delta^2 m)$. Next, let Λ denote the index set corresponding to the k entries of x with largest magnitude and write $x = x_\Lambda + x_{\Lambda^c}$. Since $x_\Lambda \in \Sigma_k$, we can write $Ax = Ax_\Lambda + Ax_{\Lambda^c} = Ax_\Lambda + e$. If A is sub-gaussian then Ax_{Λ^c} is also sub-gaussian (see Chapter 5 for details), and one can apply a similar result to (1.14) to obtain that with probability at least $1 - \exp(-c_0 m)$, $\|Ax_{\Lambda^c}\|_2 \leq 2\|x_{\Lambda^c}\|_2 = 2\sigma_k(x)_2$. Thus, applying the union bound we have that with probability exceeding $1 - 2\exp(-c_1\delta^2 m) - \exp(-c_0 m)$, we satisfy the necessary conditions to apply Theorem 1.9 to x_Λ, in which case $\sigma_k(x_\Lambda)_1 = 0$ and hence

$$\|\widehat{x} - x_\Lambda\|_2 \leq 2C_2\sigma_k(x)_2.$$

From the triangle inequality we thus obtain

$$\|\widehat{x} - x\|_2 = \|\widehat{x} - x_\Lambda + x_\Lambda - x\|_2 \leq \|\widehat{x} - x_\Lambda\|_2 + \|x_\Lambda - x\|_2 \leq (2C_2 + 1)\,\sigma_k(x)_2$$

which establishes the theorem. □

Thus, while it is not possible to achieve a deterministic guarantee of the form in (1.15) without taking a prohibitively large number of measurements, it *is* possible to show that such performance guarantees can hold with high probability while simultaneously taking far fewer measurements than would be suggested by Theorem 1.13. Note that the above result applies only to the case where the parameter is selected correctly, which requires some limited knowledge of x, namely $\sigma_k(x)_2$. In practice this limitation can easily be overcome through a parameter selection technique such as cross-validation [243], but there also exist more intricate analyses of ℓ_1 minimization that show it is possible to obtain similar performance without requiring an oracle for parameter selection [248]. Note that Theorem 1.14 can also be generalized to handle other measurement matrices and to the case where x is compressible rather than sparse. Moreover, this proof technique is applicable to a variety of the greedy algorithms described in Chapter 8 that do not require knowledge of the noise level to establish similar results [56, 190].

1.5.4 The cross-polytope and phase transitions

While the RIP-based analysis of ℓ_1 minimization allows us to establish a variety of guarantees under different noise settings, one drawback is that the analysis of how many measurements are actually required for a matrix to satisfy the RIP is relatively loose. An alternative approach to analyzing ℓ_1 minimization algorithms is to examine them from a more geometric perspective. Towards this end, we define the closed ℓ_1 ball, also known as the *cross-polytope*:

$$C^n = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}.$$

Note that C^n is the convex hull of $2n$ points $\{p_i\}_{i=1}^{2n}$. Let $AC^n \subseteq \mathbb{R}^m$ denote the convex polytope defined as either the convex hull of $\{Ap_i\}_{i=1}^{2n}$ or equivalently as

$$AC^n = \{y \in \mathbb{R}^m : y = Ax, x \in C^n\}.$$

For any $x \in \Sigma_k$, we can associate a k-face of C^n with the support and sign pattern of x. One can show that the number of k-faces of AC^n is precisely the number of index sets of size k for which signals supported on them can be recovered by (1.12) with $\mathcal{B}(y) = \{z : Az = y\}$. Thus, ℓ_1 minimization yields the same solution as ℓ_0 minimization for all $x \in \Sigma_k$ if and only if the number of k-faces of AC^n is identical to the number of k-faces of C^n. Moreover, by counting the number of k-faces of AC^n, we can quantify exactly what fraction of sparse vectors can be recovered using ℓ_1 minimization with A as our sensing matrix. See [81, 84, 92–94] for more details and [95] for an overview of the implications of this body of work. Note also that by replacing the cross-polytope with certain other polytopes (the simplex and the hypercube), one can apply the same technique to obtain results concerning the recovery of more limited signal classes, such as sparse signals with nonnegative or bounded entries [95].

Given this result, one can then study random matrix constructions from this perspective to obtain probabilistic bounds on the number of k-faces of AC^n with A is generated at random, such as from a Gaussian distribution. Under the assumption that $k = \rho m$ and $m = \gamma n$, one can obtain asymptotic results as $n \to \infty$. This analysis leads to the *phase transition* phenomenon, where for very large problem sizes there are sharp thresholds dictating that the fraction of k-faces preserved will tend to either one or zero with very high probability, depending on ρ and γ [95]. For the precise values of ρ and γ which will enable successful recovery and for further discussion of similar results, see Chapters 7 and 9.

These results provide sharp bounds on the minimum number of measurements required in the noiseless case. In general, these bounds are significantly stronger than the corresponding measurement bounds obtained within the RIP-based framework, which tend to be extremely loose in terms of the constants involved. However, these sharper bounds also require somewhat more intricate analysis and typically more restrictive assumptions on A (such as it being Gaussian). Thus, one of the main strengths of the RIP-based analysis presented in this chapter is that it gives results for a very broad class of matrices that can also be extended to noisy settings.

1.6 Signal Recovery Algorithms

We now discuss a number of algorithmic approaches to the problem of signal recovery from CS measurements. While this problem has received significant attention in recent years in the context of CS, many of these techniques predate the field of CS. There are a variety of algorithms that have been used in applications such as sparse approximation, statistics, geophysics, and theoretical computer science that were developed to exploit sparsity in other contexts and can be brought to bear on the CS recovery problem. We briefly review some of these, and refer the reader to later chapters as well as the overview in [226] for further details.

Note that we restrict our attention here to algorithms that actually reconstruct the original signal x. In some settings the end goal is to solve some kind of inference problem such as detection, classification, or parameter estimation, in which case a full reconstruction may not be necessary [69–71, 74, 100, 101, 143, 145].

ℓ_1 minimization algorithms

The ℓ_1 minimization approach analyzed in Section 1.5 provides a powerful framework for recovering sparse signals. The power of ℓ_1 minimization is that not only will it lead to a provably accurate recovery, but the formulations described in Section 1.5 are also convex optimization problems for which there exist efficient and accurate numerical solvers [194]. For example, (1.12) with $\mathcal{B}(y) = \{z : Az = y\}$ can be posed as a linear program. In the cases where $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq \epsilon\}$ or $\mathcal{B}(y) = \{z : \left\|A^T(Az - y)\right\|_\infty \leq \lambda\}$, the minimization problem (1.12) becomes a convex program with a conic constraint.

While these optimization problems could all be solved using general-purpose convex optimization software, there now also exist a tremendous variety of algorithms designed to explicitly solve these problems in the context of CS. This body of literature has primarily focussed on the case where $\mathcal{B}(y) = \{z : \|Az - y\|_2 \leq \epsilon\}$. However, there exist multiple equivalent formulations of this program. For instance, the majority of ℓ_1 minimization algorithms in the literature have actually considered the unconstrained version of this problem, i.e.,

$$\widehat{x} = \arg\min_z \frac{1}{2} \|Az - y\|_2^2 + \lambda \|z\|_1 .$$

See, for example, [11, 120, 122, 138, 175, 197, 246, 249–251]. Note that for some choice of the parameter λ this optimization problem will yield the same result as the constrained version of the problem given by

$$\widehat{x} = \arg\min_z \|z\|_1 \quad \text{subject to} \quad \|Az - y\|_2 \leq \epsilon.$$

However, in general the value of λ which makes these problems equivalent is unknown a priori. Several approaches for choosing λ are discussed in [110, 123, 133]. Since in many settings ϵ is a more natural parameterization (being determined by the noise or quantization level), it is also useful to have algorithms that directly solve the latter formulation. While there are fewer efforts in this direction, there also exist some excellent solvers for this problem [12, 13, 231]. Note that [13] also provides solvers for a variety of other ℓ_1 minimization problems, such as for the Dantzig selector.

Greedy algorithms

While convex optimization techniques are powerful methods for computing sparse representations, there are also a variety of greedy/iterative methods for

Algorithm 1.1 Orthogonal Matching Pursuit

```
**Inputs:** CS matrix/dictionary $A$, measurement vector $y$
**Initialize:** $\widehat{x}_0 = 0$, $r_0 = y$, $\Lambda_0 = \emptyset$.
**for** $i = 1$; $i := i + 1$ until stopping criterion is met **do**
    $g_i \leftarrow A^T r_{i-1}$ {form signal estimate from residual}
    $\Lambda_i \leftarrow \Lambda_{i-1} \cup \mathrm{supp}(H_1(g_i))$ {add largest residual entry to support}
    $\widehat{x}_i|_{\Lambda_i} \leftarrow A^{\dagger}_{\Lambda_i} y$, $\widehat{x}_i|_{\Lambda_i^c} \leftarrow 0$ {update signal estimate}
    $r_i \leftarrow y - A\widehat{x}_i$ {update measurement residual}
**end for**
**Output:** Sparse representation $\widehat{x}$
```

solving such problems [21, 23, 24, 56, 64, 66, 75, 85, 96, 153, 182, 183, 190–192, 220, 222, 223]. Greedy algorithms rely on iterative approximation of the signal coefficients and support, either by iteratively identifying the support of the signal until a convergence criterion is met, or alternatively by obtaining an improved estimate of the sparse signal at each iteration that attempts to account for the mismatch to the measured data. Some greedy methods can actually be shown to have performance guarantees that match those obtained for convex optimization approaches. In fact, some of the more sophisticated greedy algorithms are remarkably similar to those used for ℓ_1 minimization described above. However, the techniques required to prove performance guarantees are substantially different.

We refer the reader to Chapter 8 for a more detailed overview of greedy algorithms and their performance. Here we briefly highlight some of the most common methods and their theoretical guarantees. Two of the oldest and simplest greedy approaches are *Orthogonal Matching Pursuit* (OMP) and *iterative thresholding.* We first consider OMP [183], which begins by finding the column of A most correlated with the measurements. The algorithm then repeats this step by correlating the columns with the signal residual, which is obtained by subtracting the contribution of a partial estimate of the signal from the original measurement vector. The algorithm is formally defined as Algorithm 1.1, where $H_k(x)$ denotes the *hard thresholding* operator on x that sets all entries to zero except for the k entries of x with largest magnitude. The stopping criterion can consist of either a limit on the number of iterations, which also limits the number of nonzeros in $\widehat{x}$, or a requirement that $y \approx A\widehat{x}$ in some sense. Note that in either case, if OMP runs for m iterations then it will always produce an estimate $\widehat{x}$ such that $y = A\widehat{x}$. Iterative thresholding algorithms are often even more straightforward. For an overview see [107]. As an example, we consider *iterative hard thresholding* (IHT) [24], which is described in Algorithm 1.2. Starting from an initial signal estimate $\widehat{x}_0 = 0$, the algorithm iterates a gradient descent step followed by hard thresholding until a convergence criterion is met.

OMP and IHT both satisfy many of the same guarantees as ℓ_1 minimization. For example, under a slightly stronger assumption on the RIP constant, iterative

Algorithm 1.2 Iterative Hard Thresholding

```
Inputs: CS matrix/dictionary $A$, measurement vector $y$, sparsity level $k$
Initialize: $\widehat{x}_0 = 0$.
for $i = 1$; $i := i + 1$ until stopping criterion is met do
  $\widehat{x}_i = H_k\left(\widehat{x}_{i-1} + A^T(y - A\widehat{x}_{i-1})\right)$
end for
Output: Sparse representation $\widehat{x}$
```

hard thresholding satisfies a very similar guarantee to that of Theorem 1.9. We refer the reader to Chapter 8 for further details on the theoretical properties of thresholding algorithms, and focus here on OMP.

The simplest guarantees for OMP state that for exactly k-sparse x with noise-free measurements $y = Ax$, OMP will recover x exactly in k iterations. This analysis has been performed for both matrices satisfying the RIP [75] and matrices with bounded coherence [220]. In both results, however, the required constants are relatively small, so that the results only apply when $m = O(k^2 \log(n))$.

There have been many efforts to improve upon these basic results. As one example, in [173] the required number of measurements is reduced to $m = O(k^{1.6} \log(n))$ by allowing OMP to run for more than k iterations. More recently, it has been shown that this can be even further relaxed to the more familiar $m = O(k \log(n))$ and that OMP is stable with respect to bounded noise, yielding a guarantee along the lines of Theorem 1.9 but only for exactly sparse signals [254]. Both of these analyses have exploited the RIP. There has also been recent progress in using the RIP to analyze the performance of OMP on non-sparse signals [10]. At present, however, RIP-based analysis of OMP remains a topic of ongoing work.

Note that all of the above efforts have aimed at establishing uniform guarantees (although often restricted to exactly sparse signals). In light of our discussion of probabilistic guarantees in Section 1.5.3, one might expect to see improvements by considering less restrictive guarantees. As an example, it has been shown that by considering random matrices for A OMP can recover k-sparse signals in k iterations with high probability using only $m = O(k \log(n))$ measurements [222]. Similar improvements are also possible by placing restrictions on the smallest nonzero value of the signal, as in [88]. Furthermore, such restrictions also enable near-optimal recovery guarantees when the measurements are corrupted by Gaussian noise [18].

Combinatorial algorithms

In addition to ℓ_1 minimization and greedy algorithms, there is another important class of sparse recovery algorithms that we will refer to as *combinatorial algorithms*. These algorithms, mostly developed by the theoretical computer science community, in many cases pre-date the compressive sensing literature but are highly relevant to the sparse signal recovery problem.

The historically oldest of these algorithms were developed in the context of *combinatorial group testing* [98, 116, 160, 210]. In this problem we suppose that there are n total items and k anomalous elements that we wish to find. For example, we might wish to identify defective products in an industrial setting, or identify a subset of diseased tissue samples in a medical context. In both of these cases the vector x indicates which elements are anomalous, i.e., $x_i \neq 0$ for the k anomalous elements and $x_i = 0$ otherwise. Our goal is to design a collection of tests that allow us to identify the support (and possibly the values of the nonzeros) of x while also minimizing the number of tests performed. In the simplest practical setting these tests are represented by a binary matrix A whose entries a_{ij} are equal to 1 if and only if the j^{th} item is used in the i^{th} test. If the output of the test is linear with respect to the inputs, then the problem of recovering the vector x is essentially the same as the standard sparse recovery problem in CS.

Another application area in which combinatorial algorithms have proven useful is computation on *data streams* [59, 189]. As an example of a typical data streaming problem, suppose that x_i represents the number of packets passing through a network router with destination i. Simply storing the vector x is typically infeasible since the total number of possible destinations (represented by a 32-bit IP address) is $n = 2^{32}$. Thus, instead of attempting to store x directly, one can store $y = Ax$ where A is an $m \times n$ matrix with $m \ll n$. In this context the vector y is often called a *sketch*. Note that in this problem y is computed in a different manner than in the compressive sensing context. Specifically, in the network traffic example we do not ever observe x_i directly, rather we observe increments to x_i (when a packet with destination i passes through the router). Thus we construct y iteratively by adding the i^{th} column to y each time we observe an increment to x_i, which we can do since $y = Ax$ is linear. When the network traffic is dominated by traffic to a small number of destinations, the vector x is compressible, and thus the problem of recovering x from the sketch Ax is again essentially the same as the sparse recovery problem in CS.

Despite the fact that in both of these settings we ultimately wish to recover a sparse signal from a small number of linear measurements, there are also some important differences between these settings and CS. First, in these settings it is natural to assume that the designer of the reconstruction algorithm also has full control over A, and is thus free to choose A in a manner that reduces the amount of computation required to perform recovery. For example, it is often useful to design A so that it has very few nonzeros, i.e., the sensing matrix itself is also sparse [8, 128, 154]. In general, most methods involve careful construction of the sampling matrix A (although some schemes do involve "generic" sparse matrices, for example, see [20]). This is in contrast with the optimization and greedy methods that work with any matrix satisfying the conditions described in Section 1.4. Of course, this additional optimization can often lead to significantly faster algorithms [51, 60, 129, 130].

Second, note that the computational complexity of all the convex methods and greedy algorithms described above is always at least linear in terms of n, since in order to recover x we must at least incur the computational cost of reading out all n entries of x. While this may be acceptable in most typical CS applications, this becomes impractical when n is extremely large, as in the network monitoring example. In this context, one may seek to develop algorithms whose complexity is linear only in the *length of the representation* of the signal, i.e., its sparsity k. In this case the algorithm does not return a complete reconstruction of x but instead returns only its k largest elements (and their indices). As surprising as it may seem, such algorithms are indeed possible. See [60, 129, 130] for examples.

1.7 Multiple Measurement Vectors

Many applications that match the properties of CS involve distributed acquisition of multiple correlated signals. The multiple signal case where all l signals involved are sparse and exhibit the same indices for their nonzero coefficients is well known in sparse approximation literature, where it has been termed the multiple measurement vector (MMV) problem [52, 63, 134, 185, 221, 223, 232]. In the MMV setting, rather than trying to recover each single sparse vector x_i independently, $1 \leq i \leq l$, the goal is to jointly recover the set of vectors by exploiting their common sparse support. Stacking these vectors into the columns of a matrix X, there will be at most k non-zero rows in X. That is, not only is each vector k-sparse, but the non-zero values occur on a common location set. We therefore say that X is *row-sparse* and use the notation $\Lambda = \text{supp}(X)$ to denote the index set corresponding to non-zero rows.[8]

MMV problems appear quite naturally in many different application areas. Early work on MMV algorithms focused on magnetoencephalography, which is a modality for imaging the brain [134, 135, 200]. Similar ideas were also developed in the context of array processing [135, 157, 181], equalization of sparse communication channels [2, 62, 119, 142], and more recently cognitive radio and multiband communications [9, 114, 186–188, 252].

Conditions on measurement matrices
As in standard CS, we assume that we are given measurements $\{y_i\}_{i=1}^{l}$ where each vector is of length $m < n$. Letting Y be the $m \times l$ matrix with columns y_i, our problem is to recover X assuming a known measurement matrix A so that $Y = AX$. Clearly, we can apply any CS method to recover x_i from y_i as before. However, since the vectors $\{x_i\}$ all have a common support, we expect intuitively to improve the recovery ability by exploiting this joint information. In

[8] The MMV problem can be converted into a block-sparse recovery problem through appropriate rasterizing of the matrix X and the construction of a single matrix $A' \in \mathbb{R}^{lm \times ln}$ dependent on the matrix $A \in \mathbb{R}^{m \times n}$ used for each of the signals.

other words, we should in general be able to reduce the number of measurements ml needed to represent X below sl, where s is the number of measurements required to recover one vector x_i for a given matrix A.

Since $|\Lambda| = k$, the rank of X satisfies $\mathrm{rank}(X) \leq k$. When $\mathrm{rank}(X) = 1$, all the sparse vectors x_i are multiples of each other, so that there is no advantage to their joint processing. However, when $\mathrm{rank}(X)$ is large, we expect to be able to exploit the diversity in its columns in order to benefit from joint recovery. This essential result is captured nicely by the following necessary and sufficient uniqueness condition:

Theorem 1.15 (Theorem 2 of [76])**.** *A necessary and sufficient condition for the measurements $Y = AX$ to uniquely determine the row sparse matrix X is that*

$$|\mathrm{supp}(X)| < \frac{\mathrm{spark}(A) - 1 + \mathrm{rank}(X)}{2}. \tag{1.16}$$

As shown in [76], we can replace $\mathrm{rank}(X)$ by $\mathrm{rank}(Y)$ in (1.16). The sufficient direction of this condition was shown in [185] to hold even in the case where there are infinitely many vectors x_i. A direct consequence of Theorem 1.15 is that matrices X with larger rank can be recovered from fewer measurements. Alternatively, matrices X with larger support can be recovered from the same number of measurements. When $\mathrm{rank}(X) = k$ and $\mathrm{spark}(A)$ takes on its largest possible value equal to $m + 1$, condition (1.16) becomes $m \geq k + 1$. Therefore, in this best-case scenario, only $k + 1$ measurements per signal are needed to ensure uniqueness. This is much lower than the value of $2k$ obtained in standard CS via the spark (cf. Theorem 1.7), which we refer to here as the single measurement vector (SMV) setting. Furthermore, when X is full rank, it can be recovered by a simple algorithm, in contrast to the combinatorial complexity needed to solve the SMV problem from $2k$ measurements for general matrices A. See Chapter 8 for more details.

Recovery Algorithms

A variety of algorithms have been proposed that exploit the joint sparsity in different ways when X is not full rank. As in the SMV setting, two main approaches to solving MMV problems are based on convex optimization and greedy methods. The analogue of (1.10) in the MMV case is

$$\widehat{X} = \arg \min_{X \in \mathbb{R}^{n \times l}} \|X\|_{p,0} \text{ subject to } Y = AX, \tag{1.17}$$

where we define the $\ell_{p,q}$ norms for matrices as

$$\|X\|_{p,q} = \left(\sum_i \|x^i\|_p^q \right)^{1/q}$$

with x^i denoting the i^{th} row of X. With a slight abuse of notation, we also consider the $q = 0$ case where $\|X\|_{p,0} = |\text{supp}(X)|$ for any p. Optimization based algorithms relax the ℓ_0 norm in (1.17) and attempt to recover X by mixed norm minimization:

$$\widehat{X} = \arg \min_{X \in \mathbb{R}^{n \times l}} \|X\|_{p,q} \text{ subject to } Y = AX$$

for some $p, q \geq 1$; values for p and q of 1, 2, and ∞ have been advocated [52, 63, 114, 121, 221, 223]. The standard greedy approaches in the SMV setting have also been extended to the MMV case; see Chapter 8 for more details. Furthermore, one can also reduce the MMV problem into an SMV problem and solve using standard CS recovery algorithms [185]. This reduction can be particularly beneficial in large scale problems, such as those resulting from analog sampling.

MMV models can also be used to perform blind CS, in which the sparsifying basis is learned together with the representation coefficients [131]. While all standard CS algorithms assume that the sparsity basis is known in the recovery process, blind CS does not require this knowledge. When multiple measurements are available it can be shown that under certain conditions on the sparsity basis, blind CS is possible thus avoiding the need to know the sparsity basis in both the sampling and the recovery process.

In terms of theoretical guarantees, it can be shown that MMV extensions of SMV algorithms will recover X under similar conditions to the SMV setting in the worst-case scenario [4, 52, 114, 115] so that theoretical equivalence results for arbitrary values of X do not predict any performance gain with joint sparsity. In practice, however, multichannel reconstruction techniques perform much better than recovering each channel individually. The reason for this discrepancy is that these results apply to all possible input signals, and are therefore worst-case measures. Clearly, if we input the same signal to each channel, namely when $\text{rank}(X) = 1$, no additional information on the joint support is provided from multiple measurements. However, as we have seen in Theorem 1.15, higher ranks of the input X improve the recovery ability.

Another way to improve performance guarantees is by considering random values of X and developing conditions under which X is recovered with high probability [7, 115, 137, 208]. Average case analysis can be used to show that fewer measurements are needed in order to recover X exactly [115]. In addition, under a mild condition on the sparsity and on the matrix A, the failure probability decays exponentially in the number of channels l [115].

Finally, we note that algorithms similar to those used for MMV recovery can also be adapted to block-sparse reconstruction [112, 114, 253].

1.8 Summary

CS is an exciting, rapidly growing, field that has attracted considerable attention in signal processing, statistics, and computer science, as well as the broader scientific community. Since its initial development, only a few years ago, thousands of papers have appeared in this area, and hundreds of conferences, workshops, and special sessions have been dedicated to this growing research field. In this chapter, we have reviewed some of the basics of the theory underlying CS. We have also aimed, throughout our summary, to highlight new directions and application areas that are at the frontier of CS research. This chapter should serve as a review to practitioners wanting to join this emerging field, and as a reference for researchers. Our hope is that this presentation will attract the interest of both mathematicians and engineers in the desire to encourage further research into this new frontier as well as promote the use of CS in practical applications. In subsequent chapters of the book, we will see how the fundamentals presented in this chapter are expanded and extended in many exciting directions, including new models for describing structure in both analog and discrete-time signals, new sensing design techniques, more advanced recovery results and powerful new recovery algorithms, and emerging applications of the basic theory and its extensions.

Acknowledgements

The authors would like to thank Ewout van den Berg, Piotr Indyk, Yaniv Plan, and the authors contributing to this book for their valuable feedback on a preliminary version of this manuscript.

A Appendix: Proofs for Chapter 1

A.1 Proof of Theorem 1.4

To prove Theorem 1.4 we first provide a preliminary lemma. The proof of this result is based on techniques from [166].

Lemma A.1. *Let k and n satisfying $k < n/2$ be given. There exists a set $X \subset \Sigma_k$ such that for any $x \in X$ we have $\|x\|_2 \le \sqrt{k}$ and for any $x, z \in X$ with $x \neq z$*

$$\|x - z\|_2 \ge \sqrt{k/2} \tag{A.1}$$

and

$$\log |X| \ge \frac{k}{2} \log \left(\frac{n}{k}\right).$$

Proof. We will begin by considering the set

$$U = \left\{ x \in \{0, +1, -1\}^n : \|x\|_0 = k \right\}.$$

By construction, $\|x\|_2^2 = k$ for all $x \in U$. Thus if we construct X by picking elements from U then we automatically have $\|x\|_2 \le \sqrt{k}$.

Next, observe that $|U| = \binom{n}{k} 2^k$. Note also that $\|x - z\|_0 \le \|x - z\|_2^2$, and thus if $\|x - z\|_2^2 \le k/2$ then $\|x - z\|_0 \le k/2$. From this we observe that for any fixed $x \in U$,

$$\left| \left\{ z \in U : \|x - z\|_2^2 \le k/2 \right\} \right| \le \left| \{ z \in U : \|x - z\|_0 \le k/2 \} \right| \le \binom{n}{k/2} 3^{k/2}.$$

Thus, suppose we construct the set X by iteratively choosing points that satisfy (A.1). After adding j points to the set, there are at least

$$\binom{n}{k} 2^k - j \binom{n}{k/2} 3^{k/2}$$

points left to pick from. Thus, we can construct a set of size $|X|$ provided that

$$|X| \binom{n}{k/2} 3^{k/2} \le \binom{n}{k} 2^k \tag{A.2}$$

Next, observe that

$$\frac{\binom{n}{k}}{\binom{n}{k/2}} = \frac{(k/2)!(n-k/2)!}{k!(n-k)!} = \prod_{i=1}^{k/2} \frac{n-k+i}{k/2+i} \geq \left(\frac{n}{k} - \frac{1}{2}\right)^{k/2},$$

where the inequality follows from the fact that $(n-k+i)/(k/2+i)$ is decreasing as a function of i. Thus, if we set $|X| = (n/k)^{k/2}$ then we have

$$|X| \left(\frac{3}{4}\right)^{k/2} = \left(\frac{3n}{4k}\right)^{k/2} = \left(\frac{n}{k} - \frac{n}{4k}\right)^{k/2} \leq \left(\frac{n}{k} - \frac{1}{2}\right)^{k/2} \leq \frac{\binom{n}{k}}{\binom{n}{k/2}}.$$

Hence, (A.2) holds for $|X| = (n/k)^{k/2}$, which establishes the lemma. □

Using this lemma, we can establish Theorem 1.4.

Theorem 1.4 (Theorem 3.5 of [67])**.** *Let A be an $m \times n$ matrix that satisfies the RIP of order $2k$ with constant $\delta \in (0, \frac{1}{2}]$. Then*

$$m \geq Ck \log\left(\frac{n}{k}\right)$$

where $C = 1/2\log(\sqrt{24}+1) \approx 0.28$.

Proof. We first note that since A satisfies the RIP, then for the set of points X in Lemma A.1 we have,

$$\|Ax - Az\|_2 \geq \sqrt{1-\delta}\, \|x - z\|_2 \geq \sqrt{k/4}$$

for all $x, z \in X$, since $x - z \in \Sigma_{2k}$ and $\delta \leq \frac{1}{2}$. Similarly, we also have

$$\|Ax\|_2 \leq \sqrt{1+\delta}\, \|x\|_2 \leq \sqrt{3k/2}$$

for all $x \in X$.

From the lower bound we can say that for any pair of points $x, z \in X$, if we center balls of radius $\sqrt{k/4}/2 = \sqrt{k/16}$ at Ax and Az, then these balls will be disjoint. In turn, the upper bound tells us that the entire set of balls is itself contained within a larger ball of radius $\sqrt{3k/2} + \sqrt{k/16}$. If we let $B^m(r) = \{x \in \mathbb{R}^m : \|x\|_2 \leq r\}$, then this implies that

$$\begin{aligned}
& & \mathrm{Vol}\left(B^m\left(\sqrt{3k/2} + \sqrt{k/16}\right)\right) &\geq |X| \cdot \mathrm{Vol}\left(B^m\left(\sqrt{k/16}\right)\right), \\
&\Leftrightarrow & \left(\sqrt{3k/2} + \sqrt{k/16}\right)^m &\geq |X| \cdot \left(\sqrt{k/16}\right)^m, \\
&\Leftrightarrow & \left(\sqrt{24} + 1\right)^m &\geq |X|, \\
&\Leftrightarrow & m &\geq \frac{\log |X|}{\log\left(\sqrt{24}+1\right)}.
\end{aligned}$$

The theorem follows by applying the bound for $|X|$ from Lemma A.1. □

A.2 Proof of Lemma 1.3

To begin, we establish the following preliminary lemmas.

Lemma A.2. *Suppose u, v are orthogonal vectors. Then*

$$\|u\|_2 + \|v\|_2 \le \sqrt{2}\,\|u+v\|_2 .$$

Proof. We begin by defining the 2×1 vector $w = [\|u\|_2 , \|v\|_2]^T$. By applying Lemma 1.2 with $k = 2$, we have $\|w\|_1 \le \sqrt{2}\,\|w\|_2$. From this we obtain

$$\|u\|_2 + \|v\|_2 \le \sqrt{2}\sqrt{\|u\|_2^2 + \|v\|_2^2}.$$

Since u and v are orthogonal, $\|u\|_2^2 + \|v\|_2^2 = \|u+v\|_2^2$, which yields the desired result. □

Lemma A.3. *If A satisfies the RIP of order $2k$, then for any pair of vectors $u, v \in \Sigma_k$ with disjoint support,*

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

Proof. Suppose $u, v \in \Sigma_k$ with disjoint support and that $\|u\|_2 = \|v\|_2 = 1$. Then, $u \pm v \in \Sigma_{2k}$ and $\|u \pm v\|_2^2 = 2$. Using the RIP we have

$$2(1-\delta_{2k}) \le \|Au \pm Av\|_2^2 \le 2(1+\delta_{2k}).$$

Finally, applying the parallelogram identity

$$|\langle Au, Av\rangle| \le \frac{1}{4}\left|\|Au+Av\|_2^2 - \|Au-Av\|_2^2\right| \le \delta_{2k}$$

establishes the lemma. □

Lemma A.4. *Let Λ_0 be an arbitrary subset of $\{1, 2, \ldots, n\}$ such that $|\Lambda_0| \le k$. For any vector $u \in \mathbb{R}^n$, define Λ_1 as the index set corresponding to the k largest entries of $u_{\Lambda_0^c}$ (in absolute value), Λ_2 as the index set corresponding to the next k largest entries, and so on. Then*

$$\sum_{j\ge 2} \|u_{\Lambda_j}\|_2 \le \frac{\|u_{\Lambda_0^c}\|_1}{\sqrt{k}}.$$

Proof. We begin by observing that for $j \ge 2$,

$$\|u_{\Lambda_j}\|_\infty \le \frac{\|u_{\Lambda_{j-1}}\|_1}{k}$$

since the Λ_j sort u to have decreasing magnitude. Applying Lemma 1.2 we have

$$\sum_{j\ge 2} \|u_{\Lambda_j}\|_2 \le \sqrt{k}\sum_{j\ge 2} \|u_{\Lambda_j}\|_\infty \le \frac{1}{\sqrt{k}}\sum_{j\ge 1} \|u_{\Lambda_j}\|_1 = \frac{\|u_{\Lambda_0^c}\|_1}{\sqrt{k}},$$

proving the lemma. □

We are now in a position to prove Lemma 1.3. The key ideas in this proof follow from [34].

Lemma 1.3. *Suppose that A satisfies the RIP of order $2k$. Let Λ_0 be an arbitrary subset of $\{1,2,\ldots,n\}$ such that $|\Lambda_0| \le k$, and let $h \in \mathbb{R}^n$ be given. Define Λ_1 as the index set corresponding to the k entries of $h_{\Lambda_0^c}$ with largest magnitude, and set $\Lambda = \Lambda_0 \cup \Lambda_1$. Then*

$$\|h_\Lambda\|_2 \le \alpha \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{k}} + \beta \frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2},$$

where

$$\alpha = \frac{\sqrt{2}\delta_{2k}}{1-\delta_{2k}}, \quad \beta = \frac{1}{1-\delta_{2k}}.$$

Proof. Since $h_\Lambda \in \Sigma_{2k}$, the lower bound on the RIP immediately yields

$$(1-\delta_{2k})\|h_\Lambda\|_2^2 \le \|Ah_\Lambda\|_2^2. \tag{A.3}$$

Define Λ_j as in Lemma A.4, then since $Ah_\Lambda = Ah - \sum_{j\ge2} Ah_{\Lambda_j}$, we can rewrite (A.3) as

$$(1-\delta_{2k})\|h_\Lambda\|_2^2 \le \langle Ah_\Lambda, Ah\rangle - \left\langle Ah_\Lambda, \sum_{j\ge2} Ah_{\Lambda_j}\right\rangle. \tag{A.4}$$

In order to bound the second term of (A.4), we use Lemma A.3, which implies that

$$\left|\left\langle Ah_{\Lambda_i}, Ah_{\Lambda_j}\right\rangle\right| \le \delta_{2k}\|h_{\Lambda_i}\|_2 \left\|h_{\Lambda_j}\right\|_2, \tag{A.5}$$

for any i,j. Furthermore, Lemma A.2 yields $\|h_{\Lambda_0}\|_2 + \|h_{\Lambda_1}\|_2 \le \sqrt{2}\|h_\Lambda\|_2$. Substituting into (A.5) we obtain

$$\begin{aligned}\left|\left\langle Ah_\Lambda, \sum_{j\ge2} Ah_{\Lambda_j}\right\rangle\right| &= \left|\sum_{j\ge2}\left\langle Ah_{\Lambda_0}, Ah_{\Lambda_j}\right\rangle + \sum_{j\ge2}\left\langle Ah_{\Lambda_1}, Ah_{\Lambda_j}\right\rangle\right| \\ &\le \sum_{j\ge2}\left|\left\langle Ah_{\Lambda_0}, Ah_{\Lambda_j}\right\rangle\right| + \sum_{j\ge2}\left|\left\langle Ah_{\Lambda_1}, Ah_{\Lambda_j}\right\rangle\right| \\ &\le \delta_{2k}\|h_{\Lambda_0}\|_2 \sum_{j\ge2}\left\|h_{\Lambda_j}\right\|_2 + \delta_{2k}\|h_{\Lambda_1}\|_2 \sum_{j\ge2}\left\|h_{\Lambda_j}\right\|_2 \\ &\le \sqrt{2}\delta_{2k}\|h_\Lambda\|_2 \sum_{j\ge2}\left\|h_{\Lambda_j}\right\|_2.\end{aligned}$$

From Lemma A.4, this reduces to

$$\left|\left\langle Ah_\Lambda, \sum_{j\ge2} Ah_{\Lambda_j}\right\rangle\right| \le \sqrt{2}\delta_{2k}\|h_\Lambda\|_2 \frac{\|h_{\Lambda_0^c}\|_1}{\sqrt{k}}. \tag{A.6}$$

Combining (A.6) with (A.4) we obtain

$$\begin{aligned}(1-\delta_{2k})\left\|h_\Lambda\right\|_2^2 &\le \left|\langle Ah_\Lambda, Ah\rangle - \left\langle Ah_\Lambda, \sum_{j\ge 2} Ah_{\Lambda_j}\right\rangle\right| \\ &\le \left|\langle Ah_\Lambda, Ah\rangle\right| + \left|\left\langle Ah_\Lambda, \sum_{j\ge 2} Ah_{\Lambda_j}\right\rangle\right| \\ &\le \left|\langle Ah_\Lambda, Ah\rangle\right| + \sqrt{2}\delta_{2k}\left\|h_\Lambda\right\|_2 \frac{\left\|h_{\Lambda_0^c}\right\|_1}{\sqrt{k}},\end{aligned}$$

which yields the desired result upon rearranging. □

A.3 Proof of Lemma 1.6

We now return to the proof of Lemma 1.6. The key ideas in this proof follow from [34].

Lemma 1.6. *Suppose that A satisfies the RIP of order $2k$ with $\delta_{2k} < \sqrt{2}-1$. Let $x, \widehat{x} \in \mathbb{R}^n$ be given, and define $h = \widehat{x} - x$. Let Λ_0 denote the index set corresponding to the k entries of x with largest magnitude and Λ_1 the index set corresponding to the k entries of $h_{\Lambda_0^c}$ with largest magnitude. Set $\Lambda = \Lambda_0 \cup \Lambda_1$. If $\left\|\widehat{x}\right\|_1 \le \left\|x\right\|_1$, then*

$$\left\|h\right\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_1 \frac{\left|\langle Ah_\Lambda, Ah\rangle\right|}{\left\|h_\Lambda\right\|_2}.$$

where

$$C_0 = 2\frac{1-(1-\sqrt{2})\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}, \quad C_1 = \frac{2}{1-(1+\sqrt{2})\delta_{2k}}.$$

Proof. We begin by observing that $h = h_\Lambda + h_{\Lambda^c}$, so that from the triangle inequality

$$\left\|h\right\|_2 \le \left\|h_\Lambda\right\|_2 + \left\|h_{\Lambda^c}\right\|_2. \tag{A.7}$$

We first aim to bound $\left\|h_{\Lambda^c}\right\|_2$. From Lemma A.4 we have

$$\left\|h_{\Lambda^c}\right\|_2 = \left\|\sum_{j\ge 2} h_{\Lambda_j}\right\|_2 \le \sum_{j\ge 2}\left\|h_{\Lambda_j}\right\|_2 \le \frac{\left\|h_{\Lambda_0^c}\right\|_1}{\sqrt{k}}, \tag{A.8}$$

where the Λ_j are defined as in Lemma A.4, i.e., Λ_1 is the index set corresponding to the k largest entries of $h_{\Lambda_0^c}$ (in absolute value), Λ_2 as the index set corresponding to the next k largest entries, and so on.

We now wish to bound $\left\|h_{\Lambda_0^c}\right\|_1$. Since $\|x\|_1 \geq \|\widehat{x}\|_1$, by applying the triangle inequality we obtain

$$\begin{aligned}\|x\|_1 \geq \|x+h\|_1 &= \|x_{\Lambda_0} + h_{\Lambda_0}\|_1 + \left\|x_{\Lambda_0^c} + h_{\Lambda_0^c}\right\|_1 \\ &\geq \|x_{\Lambda_0}\|_1 - \|h_{\Lambda_0}\|_1 + \left\|h_{\Lambda_0^c}\right\|_1 - \left\|x_{\Lambda_0^c}\right\|_1 .\end{aligned}$$

Rearranging and again applying the triangle inequality,

$$\begin{aligned}\left\|h_{\Lambda_0^c}\right\|_1 &\leq \|x\|_1 - \|x_{\Lambda_0}\|_1 + \|h_{\Lambda_0}\|_1 + \left\|x_{\Lambda_0^c}\right\|_1 \\ &\leq \|x - x_{\Lambda_0}\|_1 + \|h_{\Lambda_0}\|_1 + \left\|x_{\Lambda_0^c}\right\|_1 .\end{aligned}$$

Recalling that $\sigma_k(x)_1 = \left\|x_{\Lambda_0^c}\right\|_1 = \|x - x_{\Lambda_0}\|_1$,

$$\left\|h_{\Lambda_0^c}\right\|_1 \leq \|h_{\Lambda_0}\|_1 + 2\sigma_k(x)_1. \tag{A.9}$$

Combining this with (A.8) we obtain

$$\|h_{\Lambda^c}\|_2 \leq \frac{\|h_{\Lambda_0}\|_1 + 2\sigma_k(x)_1}{\sqrt{k}} \leq \|h_{\Lambda_0}\|_2 + 2\frac{\sigma_k(x)_1}{\sqrt{k}}$$

where the last inequality follows from Lemma 1.2. By observing that $\|h_{\Lambda_0}\|_2 \leq \|h_\Lambda\|_2$ this combines with (A.7) to yield

$$\|h\|_2 \leq 2\left\|h_\Lambda\right\|_2 + 2\frac{\sigma_k(x)_1}{\sqrt{k}}. \tag{A.10}$$

We now turn to establishing a bound for $\|h_\Lambda\|_2$. Combining Lemma 1.3 with (A.9) and applying Lemma 1.2 we obtain

$$\begin{aligned}\|h_\Lambda\|_2 &\leq \alpha\frac{\left\|h_{\Lambda_0^c}\right\|_1}{\sqrt{k}} + \beta\frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2} \\ &\leq \alpha\frac{\|h_{\Lambda_0}\|_1 + 2\sigma_k(x)_1}{\sqrt{k}} + \beta\frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2} \\ &\leq \alpha\left\|h_{\Lambda_0}\right\|_2 + 2\alpha\frac{\sigma_k(x)_1}{\sqrt{k}} + \beta\frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2}.\end{aligned}$$

Since $\|h_{\Lambda_0}\|_2 \leq \|h_\Lambda\|_2$,

$$(1-\alpha)\left\|h_\Lambda\right\|_2 \leq 2\alpha\frac{\sigma_k(x)_1}{\sqrt{k}} + \beta\frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2}.$$

The assumption that $\delta_{2k} < \sqrt{2} - 1$ ensures that $\alpha < 1$. Dividing by $(1-\alpha)$ and combining with (A.10) results in

$$\|h\|_2 \leq \left(\frac{4\alpha}{1-\alpha} + 2\right)\frac{\sigma_k(x)_1}{\sqrt{k}} + \frac{2\beta}{1-\alpha}\frac{|\langle Ah_\Lambda, Ah\rangle|}{\|h_\Lambda\|_2}.$$

Plugging in for α and β yields the desired constants. □

A.4 Proof of Theorem 1.13

Theorem 1.13 (Theorem 5.1 of [57]). *Suppose that A is an $m \times n$ matrix and that $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is a recovery algorithm that satisfies*

$$\|x - \Delta(Ax)\|_2 \leq C\sigma_k(x)_2 \tag{A.11}$$

for some $k \geq 1$, then $m > \left(1 - \sqrt{1 - 1/C^2}\right) n$.

Proof. We begin by letting $h \in \mathbb{R}^n$ denote any vector in $\mathcal{N}(A)$. We write $h = h_\Lambda + h_{\Lambda^c}$ where Λ is an arbitrary set of indices satisfying $|\Lambda| \leq k$. Set $x = h_{\Lambda^c}$, and note that $Ax = Ah_{\Lambda^c} = Ah - Ah_\Lambda = -Ah_\Lambda$ since $h \in \mathcal{N}(A)$. Since $h_\Lambda \in \Sigma_k$, (A.11) implies that $\Delta(Ax) = \Delta(-Ah_\Lambda) = -h_\Lambda$. Hence, $\|x - \Delta(Ax)\|_2 = \|h_{\Lambda^c} - (-h_\Lambda)\|_2 = \|h\|_2$. Furthermore, we observe that $\sigma_k(x)_2 \leq \|x\|_2$, since by definition $\sigma_k(x)_2 \leq \|x - \widetilde{x}\|_2$ for all $\widetilde{x} \in \Sigma_k$, including $\widetilde{x} = 0$. Thus $\|h\|_2 \leq C\|h_{\Lambda^c}\|_2$. Since $\|h\|_2^2 = \|h_\Lambda\|_2^2 + \|h_{\Lambda^c}\|_2^2$, this yields

$$\|h_\Lambda\|_2^2 = \|h\|_2^2 - \|h_{\Lambda^c}\|_2^2 \leq \|h\|_2^2 - \frac{1}{C^2}\|h\|_2^2 = \left(1 - \frac{1}{C^2}\right)\|h\|_2^2 .$$

This must hold for any vector $h \in \mathcal{N}(A)$ and for any set of indices Λ such that $|\Lambda| \leq k$. In particular, let $\{v_i\}_{i=1}^{n-m}$ be an orthonormal basis for $\mathcal{N}(A)$, and define the vectors $\{h_i\}_{i=1}^n$ as follows:

$$h_j = \sum_{i=1}^{n-m} v_i(j)v_i. \tag{A.12}$$

We note that $h_j = \sum_{i=1}^{n-m}\langle e_j, v_i\rangle v_i$ where e_j denotes the vector of all zeros except for a 1 in the j-th entry. Thus we see that $h_j = P_\mathcal{N} e_j$ where $P_\mathcal{N}$ denotes an orthogonal projection onto $\mathcal{N}(A)$. Since $\|P_\mathcal{N} e_j\|_2^2 + \|P_{\mathcal{N}^\perp} e_j\|_2^2 = \|e_j\|_2^2 = 1$, we have that $\|h_j\|_2 \leq 1$. Thus, by setting $\Lambda = \{j\}$ for h_j we observe that

$$\left|\sum_{i=1}^{n-m} |v_i(j)|^2\right|^2 = |h_j(j)|^2 \leq \left(1 - \frac{1}{C^2}\right)\|h_j\|_2^2 \leq 1 - \frac{1}{C^2}.$$

Summing over $j = 1, 2, \ldots, n$, we obtain

$$n\sqrt{1 - 1/C^2} \geq \sum_{j=1}^{n}\sum_{i=1}^{n-m} |v_i(j)|^2 = \sum_{i=1}^{n-m}\sum_{j=1}^{n} |v_i(j)|^2 = \sum_{i=1}^{n-m} \|v_i\|_2^2 = n - m,$$

and thus $m \geq \left(1 - \sqrt{1 - 1/C^2}\right) n$ as desired. □

References

[1] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak. Toeplitz-structured compressed sensing matrices. In *Proc. IEEE Work. Stat. Signal Processing*, Madison, WI, Aug. 2007.

[2] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak. Compressed channel sensing: A new approach to estimating sparse multipath channels. *Proc. IEEE*, 98(6):1058–1076, 2010.

[3] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Mag.*, 24(4):118–120, 124, 2007.

[4] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Const. Approx.*, 28(3):253–263, 2008.

[6] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Found. Comput. Math.*, 9(1):51–77, 2009.

[7] D. Baron, M. Duarte, S. Sarvotham, M. Wakin, and R. Baraniuk. Distributed compressed sensing of jointly sparse signals. In *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2005.

[8] D. Baron, S. Sarvotham, and R. Baraniuk. Sudocodes - fast measurement and reconstruction of sparse signals. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Seattle, WA, Jul. 2006.

[9] J. Bazerque and G. Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Trans. Signal Processing*, 58(3):1847–1862, 2010.

[10] P. Bechler and P. Wojtaszczyk. Error estimates for orthogonal matching pursuit and random dictionaries. Preprint, Aug. 2009.

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

[12] S. Becker, J. Bobin, and E. Candès. NESTA: A fast and accurate first-order method for sparse recovery. Submitted. Available from `arXiv:0904.3367`, Apr. 2009.

[13] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. Preprint, 2010.

[14] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[15] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56:209–239, 2004.

[16] Z. Ben-Haim and Y. C. Eldar. Blind minimax estimation. *IEEE Trans. Inform. Theory*, 53(9):3145–3157, 2007.

[17] Z. Ben-Haim and Y. C. Eldar. The Cramer-Rao bound for estimating a sparse parameter vector. *IEEE Trans. Signal Processing*, 58(6):3384–3389, 2010.

[18] Z. Ben-Haim, Y. C. Eldar, and M. Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Trans. Signal Processing*, 58(10):5030–5043, 2010.

[19] Z. Ben-Haim, T. Michaeli, and Y. C. Eldar. Performance bounds and design criteria for estimating finite rate of innovation signals. Preprint, 2010.

[20] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. In *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2008.

[21] R. Berinde, P. Indyk, and M. Ruzic. Practical near-optimal sparse recovery in the ℓ_1 norm. In *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2008.

[22] A. Beurling. Sur les intégrales de Fourier absolument convergentes et leur application à une transformation fonctionelle. In *Proc. Scandinavian Math. Congress*, Helsinki, Finland, 1938.

[23] T. Blumensath and M. Davies. Gradient pursuits. *IEEE Trans. Signal Processing*, 56(6):2370–2382, 2008.

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

[25] T. Blumensath and M. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inform. Theory*, 55(4):1872–1882, 2009.

[26] B. Bodmann, P. Cassaza, and G. Kutyniok. A quantitative notion of redundancy for finite frames. To appear in *Appl. Comput. Harmon. Anal.*, 2011.

[27] P. Boufounos, H. Rauhut, and G. Kutyniok. Sparse recovery from combined fusion frame measurements. To appear in *IEEE Trans. Inform. Theory*, 2011.

[28] J. Bourgain, S. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova. Explicit constructions of rip matrices and related problems. *To appear in* Duke Math. J., 2011.

[29] Y. Bresler and P. Feng. Spectrum-blind minimum-rate sampling and reconstruction of 2-D multiband signals. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Zurich, Switzerland, Sept. 1996.

[30] D. Broomhead and M. Kirby. The Whitney reduction network: A method for computing autoassociative graphs. *Neural Comput.*, 13:2595–2616, 2001.

[31] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51(1):34–81, Feb. 2009.

[32] T. Cai and T. Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. Preprint, 2010.

[33] E. Candès. Compressive sampling. In *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

[34] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes rendus de l'Académie des Sciences, Série I*, 346(9-10):589–592, 2008.

[35] E. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. To appear in *Appl. Comput. Harmon. Anal.*, 2011.

[36] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? Preprint, 2009.

[37] E. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Stat.*, 37(5A):2145–2177, 2009.

[38] E. Candès and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010.

[39] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[40] E. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.*, 6(2):227–254, 2006.

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

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

[43] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

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

[45] E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.*, 35(6):2313–2351, 2007.

[46] C. Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Math. Ann.*, 64:95–115, 1907.

[47] C. Carathéodory. Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen. *Rend. Circ. Mat. Palermo*, 32:193–217, 1911.

[48] P. Casazza and G. Kutyniok. *Finite Frames*. Birkhäuser, Boston, MA, 2012.

[49] V. Cevher, M. Duarte, C. Hegde, and R. Baraniuk. Sparse signal recovery using Markov random fields. In *Proc. Adv. in Neural Processing Systems (NIPS)*, Vancouver, BC, Dec. 2008.

[50] V. Cevher, P. Indyk, C. Hegde, and R. G. Baraniuk. Recovery of clustered sparse signals from compressive measurements. In *Proc. Sampling Theory and Applications (SampTA)*, Marseille, France, May 2009.

[51] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Proc. Int. Coll. Autom. Lang. Programm.*, Málaga, Spain, Jul. 2002.

[52] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans. Signal Processing*, 54(12):4634–4643, 2006.

[53] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1998.

[54] Y. Chi, L. Scharf, A. Pezeshki, and R. Calderbank. Sensitivity to basis mismatch in compressed sensing. Preprint, May 2010.

[55] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, Boston, MA, 2003.

[56] A. Cohen, W. Dahmen, and R. DeVore. Instance optimal decoding by thresholding in compressed sensing. In *Int. Conf. Harmonic Analysis and Partial Differential Equations*, Madrid, Spain, Jun. 2008.

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

[58] R. Coifman and M. Maggioni. Diffusion wavelets. *Appl. Comput. Harmon. Anal.*, 21(1):53–94, 2006.

[59] G. Cormode and M. Hadjieleftheriou. Finding the frequent items in streams of data. *Comm. ACM*, 52(10):97–105, 2009.

[60] G. Cormode and S. Muthukrishnan. Improved data stream summaries: The count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[61] J. Costa and A. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Processing*, 52(8):2210–2221, 2004.

[62] S. Cotter and B. Rao. Sparse channel estimation via matching pursuit with application to equalization. *IEEE Trans. Communications*, 50(3):374–377, 2002.

[63] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 53(7):2477–2488, 2005.

[64] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, 55(5):2230–2249, 2009.

[65] I. Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, PA, 1992.

[66] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[67] M. Davenport. *Random observations on random observations: Sparse signal acquisition and processing*. PhD thesis, Rice University, Aug. 2010.

[68] M. Davenport and R. Baraniuk. Sparse geodesic paths. In *Proc. AAAI Fall Symp. on Manifold Learning*, Arlington, VA, Nov. 2009.

[69] M. Davenport, P. Boufounos, and R. Baraniuk. Compressive domain interference cancellation. In *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Saint-Malo, France, Apr. 2009.

[70] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk. Signal processing with compressive measurements. *IEEE J. Select. Top. Signal Processing*, 4(2):445–460, 2010.

[71] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk. The smashed filter for compressive classification and target recognition. In *Proc. IS&T/SPIE Symp. Elec. Imag.: Comp. Imag.*, San Jose, CA, Jan. 2007.

[72] M. Davenport, C. Hegde, M. Duarte, and R. Baraniuk. Joint manifolds for data fusion. *IEEE Trans. Image Processing*, 19(10):2580–2594, 2010.

[73] M. Davenport, J. Laska, P. Boufouons, and R. Baraniuk. A simple proof that random matrices are democratic. Technical Report TREE 0906, Rice Univ., ECE Dept., Nov. 2009.

[74] M. Davenport, S. Schnelle, J.P. Slavinsky, R. Baraniuk, M. Wakin, , and P. Boufounos. A wideband compressive radio receiver. In *Proc. IEEE Conf. Mil. Comm. (MILCOM)*, San Jose, CA, Oct. 2010.

[75] M. Davenport and M. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *IEEE Trans. Inform. Theory*, 56(9):4395–4401, 2010.

[76] M. Davies and Y. C. Eldar. Rank awareness in joint sparse recovery. Preprint, Apr. 2010.

[77] R. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[78] R. DeVore. Deterministic constructions of compressed sensing matrices. *J. Complex.*, 23(4):918–925, 2007.

[79] M. Do and C. La. Tree-based majorize-minimize algorithm for compressed sensing with sparse-tree prior. In *Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Saint Thomas, US Virgin Islands, Dec. 2007.

[80] D. Donoho. Denoising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, 1995.

[81] D. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical Report 2005-04, Stanford Univ., Stat. Dept., Jan. 2005.

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

[83] D. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.

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

[85] D. Donoho, I. Drori, Y. Tsaig, and J.-L. Stark. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Preprint, 2006.

[86] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100(5):2197–2202, 2003.

[87] D. Donoho and M. Elad. On the stability of basis pursuit in the presence of noise. *EURASIP Signal Processing J.*, 86(3):511–532, 2006.

[88] D. Donoho, M. Elad, and V. Temlyahov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[89] D. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci.*, 100(10):5591–5596, 2003.

[90] D. Donoho and C. Grimes. Image manifolds which are isometric to Euclidean space. *J. Math. Imag. and Vision*, 23(1):5–24, 2005.

[91] D. Donoho and B. Logan. Signal recovery and the large sieve. *SIAM J. Appl. Math.*, 52(6):577–591, 1992.

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

[93] D. Donoho and J. Tanner. Sparse nonnegative solutions of undetermined linear equations by linear programming. *Proc. Natl. Acad. Sci.*, 102(27):9446–9451, 2005.

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

[95] D. Donoho and J. Tanner. Precise undersampling theorems. *Proc. IEEE*, 98(6):913–924, 2010.

[96] D. Donoho and Y. Tsaig. Fast solution of ℓ_1 norm minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54(11):4789–4812, 2008.

[97] P. Dragotti, M. Vetterli, and T. Blu. Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang-fix. *IEEE Trans. Signal Processing*, 55(5):1741–1757, 2007.

[98] D. Du and F. Hwang. *Combinatorial group testing and its applications.* World Scientific, Singapore, 2000.

[99] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Mag.*, 25(2):83–91, 2008.

[100] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk. Sparse signal detection from incoherent projections. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[101] M. Duarte, M. Davenport, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk. Multiscale random projections for compressive classification. In *Proc. IEEE Int. Conf.*

Image Processing (ICIP), San Antonio, TX, Sept. 2007.

[102] M. Duarte and Y. C. Eldar. Structured compressed sensing: Theory and applications. Preprint, 2010.

[103] M. Duarte, M. Wakin, and R. Baraniuk. Fast reconstruction of piecewise smooth signals from random projections. In *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Rennes, France, Nov. 2005.

[104] M. Duarte, M. Wakin, and R. Baraniuk. Wavelet-domain compressive signal reconstruction using a hidden Markov tree model. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008.

[105] T. Dvorkind, Y. C. Eldar, and E. Matusiak. Nonlinear and non-ideal sampling: Theory and methods. *IEEE Trans. Signal Processing*, 56(12):471–481, 2009.

[106] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer, New York, NY, 2010.

[107] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *Proc. SPIE Optics Photonics: Wavelets*, San Diego, CA, Aug. 2007.

[108] Y. C. Eldar. *Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramer-Rao bound.* Foundation and Trends in Signal Processing, 2008.

[109] Y. C. Eldar. Compressed sensing of analog signals in shift-invariant spaces. *IEEE Trans. Signal Processing*, 57(8):2986–2997, 2009.

[110] Y. C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Signal Processing*, 57(2):471–481, 2009.

[111] Y. C. Eldar. Uncertainty relations for shift-invariant analog signals. *IEEE Trans. Inform. Theory*, 55(12):5742–5757, 2009.

[112] Y. C. Eldar, P. Kuppinger, and H. Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Processing*, 58(6):3042–3054, 2010.

[113] Y. C. Eldar and T. Michaeli. Beyond bandlimited sampling. *IEEE Signal Processing Mag.*, 26(3):48–68, 2009.

[114] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inform. Theory*, 55(11):5302–5316, 2009.

[115] Y. C. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Inform. Theory*, 6(1):505–519, 2010.

[116] Y. Erlich, N. Shental, A. Amir, and O. Zuk. Compressed sensing approach for high throughput carrier screen. In *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sept. 2009.

[117] P. Feng. *Universal spectrum blind minimum rate sampling and reconstruction of multiband signals.* PhD thesis, University of Illinois at Urbana-Champaign, Mar. 1997.

[118] P. Feng and Y. Bresler. Spectrum-blind minimum-rate sampling and reconstruction of multiband signals. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Atlanta, GA, May 1996.

[119] I. Fevrier, S. Gelfand, and M. Fitz. Reduced complexity decision feedback equalization for multipath channels with large delay spreads. *IEEE Trans. Communications*, 47(6):927–937, 1999.

[120] M. Figueiredo, R. Nowak, and S. Wright. Gradient projections for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Select. Top. Signal Processing*, 1(4):586–597, 2007.

[121] M. Fornassier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.

[122] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stats. Software*, 33(1):1–22, 2010.

[123] N. Galatsanos and A. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Image Processing*, 1(3):322–336, 1992.

[124] A. Garnaev and E. Gluskin. The widths of Euclidean balls. *Dokl. An. SSSR*, 277:1048–1052, 1984.

[125] K. Gedalyahu and Y. C. Eldar. Time-delay estimation from low-rate samples: A union of subspaces approach. *IEEE Trans. Signal Processing*, 58(6):3017–3031, 2010.

[126] K. Gedalyahu, R. Tur, and Y. C. Eldar. Multichannel sampling of pulse streams at the rate of innovation. *To appear in* IEEE Trans. Signal Processing, 2011.

[127] S. Geršgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk SSSR Ser. Fiz.-Mat.*, 6:749–754, 1931.

[128] A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proc. IEEE*, 98(6):937–947, 2010.

[129] A. Gilbert, Y. Li, E. Porat, and M. Strauss. Approximate sparse recovery: Optimizaing time and measurements. In *Proc. ACM Symp. Theory of Comput.*, Cambridge, MA, Jun. 2010.

[130] A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin. One sketch for all: Fast algorithms for compressed sensing. In *Proc. ACM Symp. Theory of Comput.*, San Diego, CA, Jun. 2007.

[131] S. Gleichman and Y. C. Eldar. Blind compressed sensing. Preprint, 2010.

[132] D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Comm. of the ACM*, 35(12):61–70, 1992.

[133] G. Golub and M. Heath. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1970.

[134] I. Gorodnitsky, J. George, and B. Rao. Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm. *Electroencephalography and Clinical Neurophysiology*, 95(4):231–251, 1995.

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

[136] I. Gorodnitsky, B. Rao, and J. George. Source localization in magnetoencephalagraphy using an iterative weighted minimum norm algorithm. In *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 1992.

[137] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008.

[138] E. Hale, W. Yin, and Y. Zhang. A fixed-point continuation method for ℓ_1-regularized minimization with applications to compressed sensing. Technical Report TR07-07, Rice Univ., CAAM Dept., 2007.

[139] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, New York, NY, 2001.

[140] J. Haupt, L. Applebaum, and R. Nowak. On the restricted isometry of deterministically subsampled Fourier matrices. In *Conf. Information Sciences and Systems (CISS)*,

Princeton, NJ, Mar. 2010.

[141] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak. Compressed sensing for networked data. *IEEE Signal Processing Mag.*, 25(2):92–101, 2008.

[142] J. Haupt, W. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Trans. Inform. Theory*, 56(11):5862–5875, 2010.

[143] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2006.

[144] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Trans. Inform. Theory*, 52(9):4036–4048, 2006.

[145] J. Haupt and R. Nowak. Compressive sampling for signal detection. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Honolulu, HI, Apr. 2007.

[146] D. Healy. Analog-to-information: Baa #05-35, 2005. Available online at `http://www.darpa.mil/mto/solicitations/baa05-35/s/index.html`.

[147] C. Hegde, M. Duarte, and V. Cevher. Compressive sensing recovery of spike trains using a structured sparsity model. In *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Saint-Malo, France, Apr. 2009.

[148] M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Processing*, 57(6):2275–2284, 2009.

[149] M. Herman and T. Strohmer. General deviants: An analysis of perturbations in compressed sensing. *IEEE J. Select. Top. Signal Processing*, 4(2):342–349, 2010.

[150] G. Hinton, P. Dayan, and M. Revow. Modelling the manifolds of images of handwritten digits. *IEEE Trans. Neural Networks*, 8(1):65–74, 1997.

[151] L. Hogben, editor. *Handbook of Linear Algebra*. Discrete Mathematics and its Applications. Chapman & Hall / CRC, Boca Raton, FL, 2007.

[152] P. Indyk. Explicit constructions for compressed sensing of sparse signals. In *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 30–33, Jan. 2008.

[153] P. Indyk and M. Ruzic. Near-optimal sparse recovery in the ℓ_1 norm. In *Proc. IEEE Symp. Found. Comp. Science (FOCS)*, Philadelphia, PA, Oct. 2008.

[154] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inform. Theory*, 55(9):4299–4308, 2009.

[155] W. James and C. Stein. Estimation of quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–379. University of California Press, Berkeley, 1961.

[156] T. Jayram and D. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with sub-constant error. In *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, San Francisco, CA, Jan. 2011.

[157] B. Jeffs. Sparse inverse solution methods for signal and image processing applications. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, volume 3, pages 1885–1888, May 1998.

[158] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Proc. Conf. Modern Anal. and Prob.*, New Haven, CT, Jun. 1982.

[159] G. Judge and M. Bock. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North-Holland, Amsterdam, 1978.

[160] R. Kainkaryam, A. Breux, A. Gilbert, P. Woolf, and J. Schiefelbein. poolMC: Smart pooling of mRNA samples in microarray experiments. *BMC Bioinformatics*, 11(1):299, 2010.

[161] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, 2010.

[162] V. Kotelnikov. On the carrying capacity of the ether and wire in telecommunications. In *Izd. Red. Upr. Svyazi RKKA*, Moscow, Russia, 1933.

[163] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part I). *IEEE Signal Processing Mag.*, 24(4):86–104, 2007.

[164] J. Kovačević and A. Chebira. Life beyond bases: The advent of frames (Part II). *IEEE Signal Processing Mag.*, 24(4):115–125, 2007.

[165] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. Preprint, Sept. 2010.

[166] T. Kühn. A lower estimate for entropy numbers. *J. Approx. Theory*, 110(1):120–124, 2001.

[167] C. La and M. Do. Signal reconstruction using sparse tree representation. In *Proc. SPIE Optics Photonics: Wavelets*, San Diego, CA, Aug. 2005.

[168] C. La and M. Do. Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In *IEEE Int. Conf. Image Processing (ICIP)*, Atlanta, GA, Oct. 2006.

[169] J. Laska, P. Boufounos, M. Davenport, and R. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. Preprint, 2009.

[170] J. Laska, M. Davenport, and R. Baraniuk. Exact signal recovery from corrupted measurements through the pursuit of justice. In *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2009.

[171] S. Levy and P. Fullagar. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, 46(9):1235–1243, 1981.

[172] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[173] E. Livshitz. On efficiency of Orthogonal Matching Pursuit. Preprint, Apr. 2010.

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

[175] I. Loris. On the performance of algorithms for the minimization of ℓ_1-penalized functions. *Inverse Problems*, 25(3):035008, 2009.

[176] H. Lu. *Geometric Theory of Images.* PhD thesis, University of California, San Diego, 1998.

[177] Y. Lu and M. Do. Sampling signals from a union of subspaces. *IEEE Signal Processing Mag.*, 25(2):41–47, Mar. 2008.

[178] M. Lustig, D. Donoho, and J. Pauly. Rapid MR imaging with compressed sensing and randomly under-sampled 3DFT trajectories. In *Proc. Annual Meeting of ISMRM*, Seattle, WA, May 2006.

[179] M. Lustig, J. Lee, D. Donoho, and J. Pauly. Faster imaging with randomly perturbed, under-sampled spirals and ℓ_1 reconstruction. In *Proc. Annual Meeting of ISMRM*, Miami, FL, May 2005.

[180] M. Lustig, J. Santos, J. Lee, D. Donoho, and J. Pauly. Application of compressed sensing for rapid MR imaging. In *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Rennes, France, Nov. 2005.

[181] D. Malioutov, M. Cetin, and A. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Processing*, 53(8):3010–3022, Aug. 2005.

[182] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, CA, 1999.

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

[184] R. Marcia, Z. Harmany, and R. Willett. Compressive coded aperture imaging. In *Proc. IS&T/SPIE Symp. Elec. Imag.: Comp. Imag.*, San Jose, CA, Jan. 2009.

[185] M. Mishali and Y. C. Eldar. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Trans. Signal Processing*, 56(10):4692–4702, 2008.

[186] M. Mishali and Y. C. Eldar. Blind multi-band signal reconstruction: Compressed sensing for analog signals. *IEEE Trans. Signal Processing*, 57(3):993–1009, 2009.

[187] M. Mishali and Y. C. Eldar. From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals. *IEEE J. Select. Top. Signal Processing*, 4(2):375–391, 2010.

[188] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan. Xampling: Analog to digital at sub-Nyquist rates. *IET Circuits, Devices, & Systems*, 5(1):8–20, 2011.

[189] S. Muthukrishnan. *Data Streams: Algorithms and Applications*, volume 1 of *Found. Trends in Theoretical Comput. Science*. Now Publishers, Boston, MA, 2005.

[190] D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2009.

[191] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.*, 9(3):317–334, 2009.

[192] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Select. Top. Signal Processing*, 4(2):310–316, 2010.

[193] P. Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. Technical Report TR-2008-01, Univ. of Chicago, Comput. Science Dept., 2008.

[194] J. Nocedal and S. Wright. *Numerical Optimization.* Springer-Verlag, 1999.

[195] H. Nyquist. Certain topics in telegraph transmission theory. *Trans. AIEE*, 47:617–644, 1928.

[196] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse representation. *Nature*, 381:607–609, 1996.

[197] S. Osher, Y. Mao, B. Dong, and W. Yin. Fast linearized Bregman iterations for compressive sensing and sparse denoising. *Comm. in Math. Sciences*, 8(1):93–111, 2010.

[198] J. Partington. *An Introduction to Hankel Operators.* Cambridge University Press, Cambridge, England, 1988.

[199] W. Pennebaker and J. Mitchell. *JPEG Still Image Data Compression Standard.* Van Nostrand Reinhold, 1993.

[200] J. Phillips, R. Leahy, and J. Mosher. MEG-based imaging of focal neuronal current sources. *IEEE Trans. Medical Imaging*, 16(3):338–348, 1997.

[201] R. Prony. Essai expérimental et analytique sur les lois de la Dilatabilité des fluides élastiques et sur celles de la Force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *J. de l'École Polytechnique,* Floréal et Prairial III, 1(2):24–76, 1795. R. Prony is Gaspard Riche, baron de Prony.

[202] B. Rao. Signal processing with the sparseness constraint. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Seattle, WA, May 1998.

[203] B. Recht. A simpler approach to matrix completion. To appear in *J. Machine Learning Research*, 2009.

[204] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[205] R. Robucci, L. Chiu, J. Gray, J. Romberg, P. Hasler, and D. Anderson. Compressive sensing on a CMOS separable transform image sensor. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008.

[206] J. Romberg. Compressive sensing by random convolution. *SIAM J. Imag. Sci.*, 2(4):1098–1128, 2009.

[207] M. Rosenfeld. *The Mathematics of Paul Erdős II*, chapter In praise of the Gram matrix, pages 318–323. Springer, Berlin, Germany, 1996.

[208] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Lett.*, 14(11):828–831, 2007.

[209] C. Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949.

[210] N. Shental, A. Amir, and O. Zuk. Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Research*, 38(19):e179, 2009.

[211] J. P. Slavinsky, J. Laska, M. Davenport, and R. Baraniuk. The compressive mutliplexer for multi-channel compressive sensing. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[212] A. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Math. Programming, Series A and B*, 109(2):367–384, 2007.

[213] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 197–206. University of California Press, Berkeley, 1956.

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

[215] D. Taubman and M. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer, 2001.

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

[217] J. Tenenbaum, V.de Silva, and J. Landford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[218] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc B*, 58(1):267–288, 1996.

[219] J. Treichler, M. Davenport, and R. Baraniuk. Application of compressive sensing to the design of wideband signal acquisition receivers. In *Proc. U.S./Australia Joint Work. Defense Apps. of Signal Processing (DASP)*, Lihue, Hawaii, Sept. 2009.

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

[221] J. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86(3):589–602, 2006.

[222] J. Tropp and A. Gilbert. Signal recovery from partial information via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, 2007.

[223] J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006.

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

[225] J. Tropp, M. Wakin, M. Duarte, D. Baron, and R. Baraniuk. Random filters for compressive sampling and reconstruction. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[226] J. Tropp and S. Wright. Computational methods for sparse solution of linear inverse problems. *Proc. IEEE*, 98(6):948–958, 2010.

[227] J. Trzasko and A. Manduca. Highly undersampled magnetic resonance image reconstruction via homotopic ℓ_0-minimization. *IEEE Trans. Med. Imaging*, 28(1):106–121, 2009.

[228] R. Tur, Y. C. Eldar, and Z. Friedman. Innovation rate sampling of pulse streams with application to ultrasound imaging. *To appear in* IEEE Trans. Signal Processing, 2011.

[229] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[230] M. Unser. Sampling—50 years after Shannon. *Proc. IEEE*, 88(4):569–587, 2000.

[231] E. van den Berg and M. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. on Sci. Comp.*, 31(2):890–912, 2008.

[232] E. van den Berg and M. Friedlander. Theoretical and empirical results for recovery from multiple measurements. *IEEE Trans. Inform. Theory*, 56(5):2516–2527, 2010.

[233] B. Vandereycken and S. Vandewalle. Riemannian optimization approach for computing low-rank solutions of Lyapunov equations. In *Proc. SIAM Conf. on Optimization*, Boston, MA, May 2008.

[234] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, NY, 1999.

[235] R. Varga. *Geršgorin and His Circles.* Springer, Berlin, Germany, 2004.

[236] S. Vasanawala, M. Alley, R. Barth, B. Hargreaves, J. Pauly, and M. Lustig. Faster pediatric MRI via compressed sensing. In *Proc. Annual Meeting Soc. Pediatric Radiology (SPR)*, Carlsbad, CA, Apr. 2009.

[237] R. Venkataramani and Y. Bresler. Further results on spectrum blind sampling of 2-D signals. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Chicago, IL, Oct. 1998.

[238] R. Venkataramani and Y. Bresler. Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals. *IEEE Trans. Inform. Theory*, 46(6):2173–2183, 2000.

[239] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Processing*, 50(6):1417–1428, 2002.

[240] M. Wakin, D. Donoho, H. Choi, and R. Baraniuk. The multiscale structure of non-differentiable image manifolds. In *Proc. SPIE Optics Photonics: Wavelets*, San Diego, CA, Aug. 2005.

[241] R. Walden. Analog-to-digital converter survey and analysis. *IEEE J. Selected Areas Comm.*, 17(4):539–550, 1999.

[242] C. Walker and T. Ulrych. Autoregressive recovery of the acoustic impedance. *Geophysics*, 48(10):1338–1350, 1983.

[243] R. Ward. Compressive sensing with cross validation. *IEEE Trans. Inform. Theory*, 55(12):5773–5782, 2009.

[244] K. Weinberger and L. Saul. Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Computer Vision*, 70(1):77–90, 2006.

[245] L. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Inform. Theory*, 20(3):397–399, 1974.

[246] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation. *SIAM J. on Sci. Comp.*, 32(4):1832–1857, 2010.

[247] E. Whittaker. On the functions which are represented by the expansions of the interpolation theory. *Proc. Royal Soc. Edinburgh, Sec. A*, 35:181–194, 1915.

[248] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *Found. Comput. Math.*, 10(1):1–13, 2010.

[249] S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Processing*, 57(7):2479–2493, 2009.

[250] A. Yang, S. Sastray, A. Ganesh, and Y. Ma. Fast ℓ_1-minimization algorithms and an application in robust face recognition: A review. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Hong Kong, Sept. 2010.

[251] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM J. Imag. Sci.*, 1(1):143–168, 2008.

[252] Z. Yu, S. Hoyos, and B. Sadler. Mixed-signal parallel compressed sensing and reception for cognitive radio. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, pages 3861–3864, Las Vegas, NV, Apr. 2008.

[253] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. Ser. B*, 68(1):49–67, 2006.

[254] T. Zhang. Sparse recovery with Orthogonal Matching Pursuit under RIP. Preprint, May 2010.