# Sparse Recovery via Convex Optimization

Thesis by

Paige Alicia Randall

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy


California Institute of Technology

Pasadena, California

2009

(Defended May 22, 2009)

*Science is built upon facts, as a house is built of stones;*
*but an accumulation of facts is no more a science*
*than a heap of stones is a house.*

– *Henri Poincaré*

# Acknowledgments

I would like to thank first of all my advisor, Emmanuel Candès. I am continuously impressed by his wealth of ideas and uncanny ability to detect new and interesting research directions, and his enthusiasm is contagious. Emmanuel always held me to the highest of standards, but also had the faith that I would be able to achieve them, for which I am grateful.

The Candès group members with whom I had the privilege of interacting during my time at Caltech also deserve a special thanks. I can not imagine better friends and colleagues than Stephen Becker, Jose Costa, Laurent Demanet, Hannes Helgason, Yaniv Plan, Justin Romberg, Peter Stobbe, Mike Wakin and Lexing Ying.

I would also like to thank Professors Alexei Kitaev, Bob McEliece, and Houman Owhadi for serving on my candidacy and thesis committees, and for their valuable time and helpful comments.

At Caltech I was fortunate enough to enjoy the company of friends who made being a graduate student at times almost fun and enjoyable. I dare not hope to encounter such a concentration of creative, active, intelligent, and kind people ever again. Thank you Lotty Ackerman, Dísa Elíasdóttir, Mike Armen, Luc Bouten, Cynthia Chiang, Mary Dunlop, Sarah Fogal Sweatlock, Alex Gittens, Dan Grin, Katalin Grubits Othmer, Mark Hesselink, Chris Hirata, Asa Hopkins, Mike Kesden, Chuck Lanski, Elizabeth Lanski, Jennifer Lanski, Zhiyi Li, Stéphane Lintner, Tony Miller, Dónal O'Connell, Jon Othmer, Tracy Northup, Annika Peter, Jonathan Pritchard, Marie-Hélène Rousseau, Ari Stern, John Stockton, Sheri Stoll, Sherry Suyu, Nathalie Vriend, Daniel Wagenaar and Rob Ward.

I also owe a debt of gratitude to my family, all of the Randalls, Hartleroads and Van Handels who have loved and supported me throughout this thesis and throughout my life.

Finally, this thesis would not have been possible without the constant encouragement and unwavering support of Ramon van Handel, who has taught me more about learning, life, and love, than anyone else. Ik hou van je, mijn beer, meer dan alles.

# Abstract

This thesis considers the problem of estimating a sparse signal from a few (possibly noisy) linear measurements. In other words, we have

$$y = Ax + z$$

where $A$ is an $m \times n$ measurement matrix with more columns than rows, $x$ is a sparse signal to be estimated, $z$ is a noise vector, and $y$ is a vector of measurements. This setup arises frequently in many problems ranging from MRI imaging to genomics to compressed sensing.

We begin by relating our setup to an error correction problem over the reals, where a received encoded message is corrupted by a few arbitrary errors, as well as smaller dense errors. We show that under suitable conditions on the encoding matrix and on the number of arbitrary errors, one is able to accurately recover the message.

We next show that we are able to achieve oracle optimality for $x$, up to a log factor and a factor of $\sqrt{s}$, when we require the matrix $A$ to obey an incoherence property. The incoherence property is novel in that it allows the coherence of $A$ to be as large as $O(1/\log n)$ and still allows sparsities as large as $O(m/\log n)$. This is in contrast to other existing results involving coherence where the coherence can only be as large as $O(1/\sqrt{m})$ to allow sparsities as large as $O(\sqrt{m})$. We also do not make the common assumption that the matrix $A$ obeys a restricted eigenvalue condition.

We then show that we can recover a (non-sparse) signal from a few linear measurements when the signal has an exactly sparse representation in an overcomplete dictionary. We again only require that the dictionary obey an incoherence property.

Finally, we introduce the method of $\ell_1$ analysis and show that it is guaranteed to give good recovery of a signal from a few measurements, when the signal can be well represented in a dictionary. We require that the combined measurement/dictionary matrix satisfies a uniform uncertainty principle and we compare our results with the more standard $\ell_1$ synthesis approach.

All our methods involve solving an $\ell_1$ minimization program which can be written as either a linear program or a second-order cone program, and the well-established machinery of convex optimization used to solve it rapidly.

# Contents

# Chapter 1

# Introduction

As scientists and engineers, we are often faced with large, complicated systems. These systems only become tractable if we are able to distill out a few key components that characterize the entire system. Thus sparsity, the notion that only a few components are important out of many, often determines whether a system can be efficiently modeled or is hopelessly complex. In our attempts to understand the world around us we are always trying to simplify and reduce to essentials. In other words, we are always on the lookout for sparsity.

Sparsity is thus one of the themes of this thesis. More specifically, we will be concerned with sparsity as it relates to signal processing and statistics. We begin with a discussion of how sparsity is important in signal compression, approximation and estimation.

## 1.1 Sparsity and compression

As technology improves, the demands of digital data storage and transmission increase. However, because there are only finite resources for storage and transmission, we would like to store and transmit only the data essentials. In other words, we would like to somehow take digital signals and compress them. This is possible because signals in general contain redundant information; the actual content of a signal is often much less than the ambient size of the signal.

For example, if one could find an orthobasis in which many signals of interest are almost sparse, then one could compress them by setting small entries to zero. This strategy underlies transform coders, where a signal is transformed into a basis

where it has quickly decaying coefficients, the small coefficients are set to zero, and only the large coefficients are recorded.

### 1.1.1 A familiar example: JPEG compression of digital images

To be more concrete, photos taken with a digital camera are usually stored as JPEG files on the memory card. Simplifying the inner workings of a digital camera greatly, when a photo is taken a large array of numbers is generated, corresponding to the pixels of the image. (Actually, three arrays of numbers are generated, corresponding usually to YCbCR colorspace, but we neglect these details.) Almost none of these numbers are close to zero because for natural images the light intensity at each pixel is usually not small (or a small variation from an average intensity).

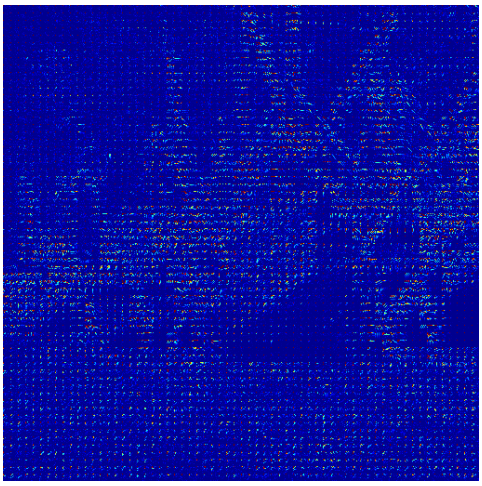
The image is then divided into $8\times 8$ blocks of pixels and each block is transformed by a discrete cosine transformation (DCT)—an orthonormal transform closely related to the Fourier transform. For a one-dimensional signal of length $N$, it is given by

$$X_0 = \sqrt{\frac{1}{N}} \sum_{n=0}^{N-1} x_n$$
$$X_k = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x_n \cos(\pi/N(n+1/2)k) \qquad k = 1, \ldots, N-1$$

where the $X_j$ are the DCT coefficients and the $x_j$ are the signal entries. The two-dimensional transform is just two 1-d transforms, one along each dimension.

The transformed version of the image has only a few large coefficients, and only the largest coefficients are stored on the camera's memory card. See Figure 1.1 for a grayscale image and its block DCT transform. (We note that what we have stated here is is not a strictly correct. The coefficients first undergo quantization and then entropy encoding before being stored. In the quantization stage, the different frequencies of the DCT coefficients are divided by different constants and then rounded. The constants are larger for the high frequency components because the human eye is not as sensitive to high frequency brightness variation. This step causes many of the high frequency coefficients, which are in general smaller than the low frequency components anyway, to be rounded to zero, and is where the compres-

**(a)** Original Boats image



**(b)** Block DCT transform of Boats

**(c)** Reconstruction from 10% of block DCT coefficients

**(d)** Reconstruction from 4% of block DCT coefficients

**Figure 1.1.** 1.1a shows the standard test image Boats [2]. 1.1b shows the block DCT transform of Boats, where dark blue signifies small coefficients and lighter red and yellow signifies larger coefficients. Clearly most of the coefficients are small. 1.1c is a reconstruction of Boats from only the largest 10% of the block DCT coefficients. 1.1d is the reconstruction of Boats from only the largest 4% of the block DCT coefficients. At this compression blocky artifacts become apparent, although the image is still recognizable.

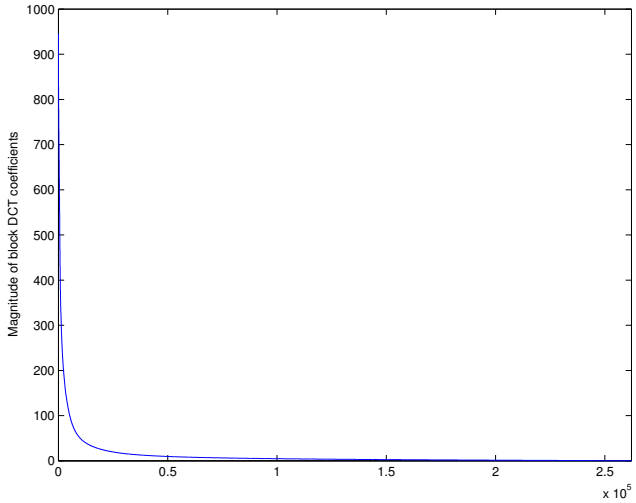



**Figure 1.2.** Block DCT coefficients of the Boats image sorted by magnitude. Note that almost all of the coefficients are very small, while relatively few are significant.

sion occurs. The higher the desired compression, the larger the constants, causing more and more coefficients to be rounded to zero. For a more complete description of the JPEG standard including entropy encoding, which we have neglected entirely here, see [91].)

The amount of savings from the compression can be seen in the difference in file sizes from an image stored in RAW format (no compression) and JPEG format. To view the photo on your computer, the inverse process is applied. JPEG is a lossy form of compression, but if the coefficients that are set to zero are very small there is not too much visual loss of quality.

We mention this example of the digital camera not only because it highlights the importance of sparsity in compression, but also because it introduces another theme of this thesis, namely that of fast algorithms. There exist algorithms for rapidly applying the DCT that are only $O(n \log n)$, while matrix-vector multiplication is ordinarily an $O(n^2)$ operation. In the specific case of JPEG compression, in multiplying each 64 pixel image block by a $64 \times 64$ matrix, $O(n \log n)$ versus $O(n^2)$ makes little difference (and depending on what the constants are and how the fast algorithm is implemented, the fast algorithm could actually be slower than the standard method of matrix-vector multiplication for this small of $n$), but later we will encounter scenarios where having a computationally reasonable approach is critical.

### 1.1.2 A second example: the unit step function

However, while the discrete cosine basis seems to do a good job of sparsifying digital natural images, the story does not stop there. For example, consider the step function depicted in Figure 1.3a. The DCT coefficients of this signal are plotted in Figure 1.3b and in Figure 1.3c we show the coefficients of a Haar transformation of the signal. The Haar transform takes neighboring signal values and computes their averages and differences. The differences are recorded and the procedure is repeated on the averages. The differences of the averages are recorded and the procedure is again repeated. Eventually, one is left with the global signal average and the differences of the signal with the average signal computed at different scales. The transform is invertible.

Intuitively, recording the differences of neighboring signal values is a reasonable approach because usually adjacent values show strong correlations—for typical signals, only at a few locations are there large jumps. This is certainly the case for our step function, and in fact its Haar transform contains only eight nonzero coefficients. In contrast, many of the step function DCT coefficients are nonzero. In Figure 1.3d we show the reconstruction of the step function from its eight largest DCT coefficients. The reconstruction from the eight largest Haar coefficients is, of course, exact.

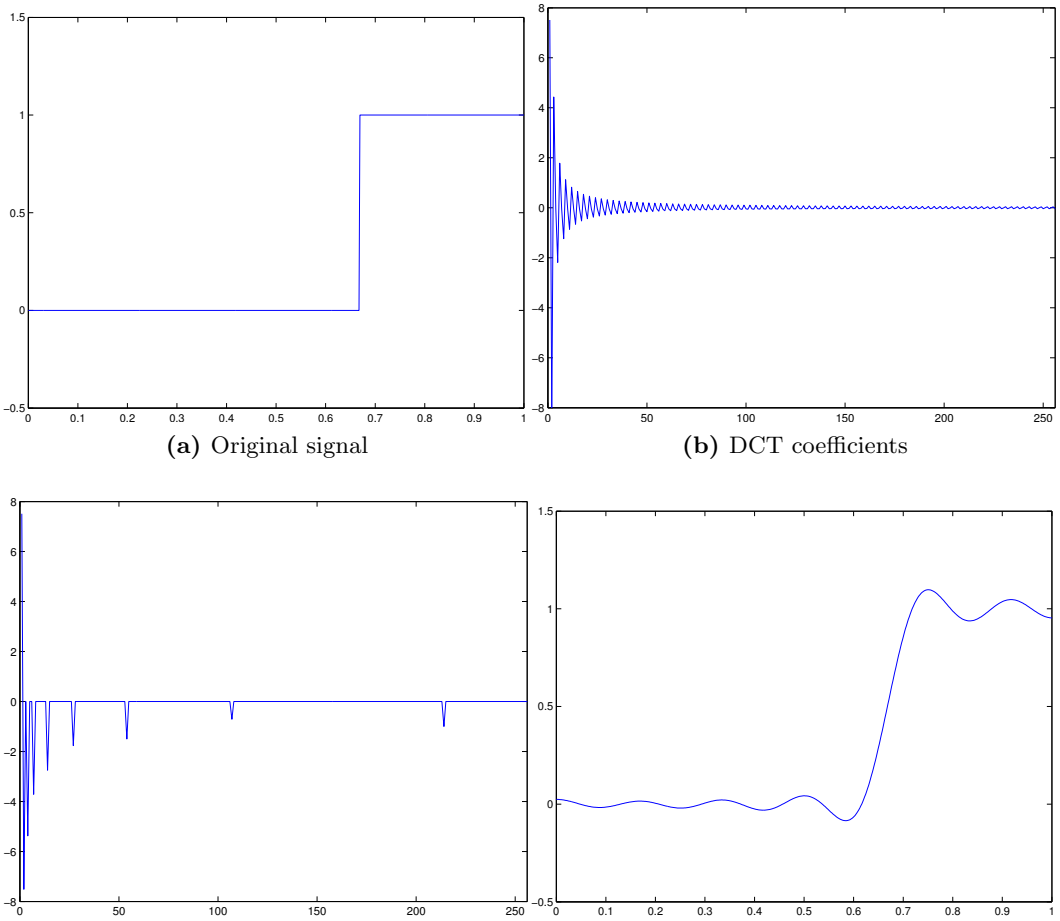
To be better able to quantify this sparsifying property of the Haar transform, we turn to the language of approximation theory.

## 1.2 Sparsity and approximation

In order to make precise our observations about the DCT and Haar transforms, we first introduce some mathematical formalism. Let $\{\psi_i\}_{i=1,2,...}$ be the elements of an orthonormal basis $\Psi$ of a Hilbert space $H$. Then we can write, for any $f \in H$,

$$f = \sum_i \langle f, \psi_i \rangle \psi_i.$$

Let $I_M \subset \{1, 2, ...\}$ be a fixed set of size $M$ and $f_M$ be the orthogonal projection




**(a)** Original signal

**(b)** DCT coefficients

**(c)** Haar transform coefficients

**(d)** Reconstruction of signal from the eight largest DCT coefficients.

**Figure 1.3.** 1.3a shows a unit step function with a jump at $x = 2/3$. 1.3b shows the DCT coefficents of the signal. 1.3c shows the Haar coefficients; note that all but eight are exactly zero. 1.3d shows the reconstruction of the signal from the eight largest DCT coefficients. Because there are only eight non-zero Haar transform coefficients, reconstruction from the eight largest Haar coefficients is exact.

of $f$ over the space generated by $\{\psi_i\}_{i\in I_M}$. This gives

$$f_M = \sum_{i\in I_M} \langle f, \psi_i\rangle \psi_i,$$

and our approximation error measured in the $L^2$ metric is

$$\|f - f_M\|_{L^2}^2 = \sum_{i\notin I_M} |\langle f, \psi_i\rangle|^2.$$

Thus, if a signal concentrates most of its energy in the space spanned by the $\{\psi_i\}_{i\in I_M}$ and $|\langle f, \psi_i\rangle|$ decays quickly outside of $I_M$, the approximation will be good. This is a linear approximation because $I_M$ is fixed in advance.

It is possible to do better if instead a nonlinear approximation is made by letting $I_M$ be the set that contains the $M$ biggest coefficients of $f$ in the basis $\Psi$. Then $f_M$ is what one has by transforming $f$, setting all but the $M$ biggest terms to zero and transforming back. This is typically what is done in compression schemes. If the coefficients of a signal in an orthobasis decay quickly then $f_M$ will approximate $f$ well. If the reordered coefficients of a class of signals decay faster in one basis than the coefficients in another basis, we expect that the first basis will provide better compression.

### 1.2.1 The step function example revisited

We return now to our example of a step function. To examine the coefficient decay we will consider the continuous step function defined on the unit interval as

$$f(t) = \begin{cases} 0 & 0 \le t < \frac{2}{3} \\ 1 & \frac{2}{3} \le t < 1. \end{cases}$$

For simplicity, instead of considering sinusoids we will take the Fourier basis of complex exponentials as our orthonormal basis of $L^2[0,1]$,

$$\psi_k(t) = e^{-i2\pi kt}, \quad k \in \mathbb{Z}.$$

Thus we have

$$\begin{aligned}\langle f, \psi_k \rangle &= \int_0^1 f(t) e^{i2\pi kt} dt \\ &= \int_{2/3}^1 e^{i2\pi kt} dt \\ &= \frac{1}{2\pi ik}(1 - e^{i4\pi k/3})\end{aligned}$$

and

$$|\langle f, \psi_k \rangle|^2 = \begin{cases} 1/3 & k = 0 \\ \frac{1}{2\pi^2 k^2}\left(1 - \cos\left(\frac{4\pi k}{3}\right)\right) & k \neq 0. \end{cases}$$

Keeping the $M = 2n + 1$ lowest frequencies in the Fourier basis (which corresponds to keeping $|k| \leq n$) gives a linear approximation error of

$$\begin{aligned}\|f - f_M\|_{L^2}^2 &= \sum_{|k|>n} \frac{1}{2\pi^2 k^2}\left(1 - \cos\left(\frac{4\pi k}{3}\right)\right) \\ &\leq \sum_{|k|>n} \frac{1}{\pi^2 k^2} \\ &\leq \frac{2}{\pi^2 n}.\end{aligned}$$

Thus

$$\|f - f_M\|_{L^2}^2 = O(M^{-1}).$$

Now we turn to calculating the approximation error in the Haar basis. The orthonormal Haar basis for $L^2[0, 1]$ is [56]

$$\begin{cases} \psi_{j,n}(t) = 2^{j/2}\psi\left(\frac{t - 2^{-j}n}{2^{-j}}\right) & j \geq 0,\ 0 \leq n < 2^j \\ \phi(t) = 1 & 0 \leq t \leq 1 \end{cases}$$

where

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2 \\ -1 & 1/2 \leq t < 1 \\ 0 & \text{else.} \end{cases}$$

The $\phi(t)$ element of the Haar basis is rather special, and we must include it only

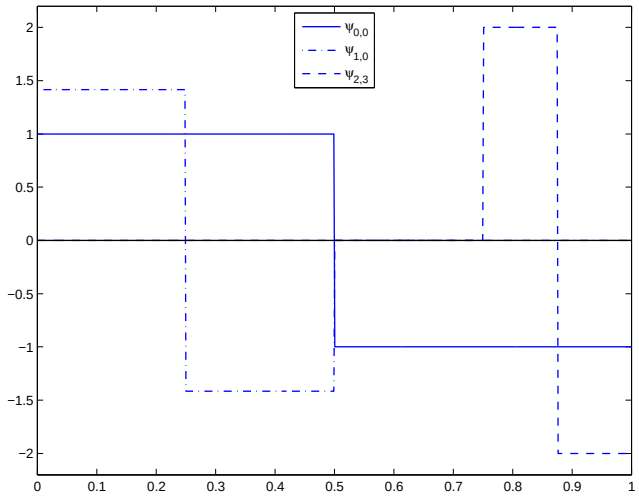



**Figure 1.4.** Haar basis elements at scales $j = 0, 1, 2$. Note that each wavelet has zero average and compact support, the width of which depends on $j$. The height also varies with $j$ to ensure each element has unit norm.

because we want a basis of the interval $[0, 1]$. If we wanted a basis of the real line, for example, we would not need it. The important things to note about the basis elements $\psi_{j,n}(t)$ are that each $\psi_{j,n}$ is nonzero only in $t \in [n2^{-j}, (n+1)2^{-j}]$ and has zero average. See Figure 1.4 for depictions of some elements of the Haar basis at different $j$, or scales, and $n$, or translations.

We now calculate the inner products of our step function $f(t)$ with the elements of the Haar basis. We have $\langle \phi(t), f(t) \rangle = 1/3$. To calculate the inner products of $f(t)$ with the $\psi_{j,n}(t)$ basis elements, we notice that for each $j$, only one $n = n'$ will have a nonzero coefficient, because only one $n$ will have a support that overlaps with the discontinuity of $f(t)$. The other inner products will be zero because on the support of the other basis elements, $f(t)$ will be constant, and the average of each Haar basis element $\psi_{j,n}$ is zero. The nonzero inner product will be bounded by

$$\begin{aligned} |\langle \psi_{j,n'}(t), f(t) \rangle| &\leq (\text{height of } \psi_{j,n'}(t)) \times \frac{1}{2} (\text{support of } \psi_{j,n'}(t)) \\ &= 2^{j/2} \cdot \frac{1}{2} \cdot 2^{-j} = 2^{-j/2-1}, \end{aligned}$$

which decays as $j$ increases.

Thus if we keep the $M$ terms consisting of $\langle f, \phi \rangle$ and the $M - 1$ smallest $j$ (for

whatever $n$ is nonzero) we have

$$\begin{aligned}\|f - f_M\|_{L^2}^2 &= \sum_{j=M-1}^{\infty} |\langle \psi_{j,n}(t), f(t) \rangle|^2 \\ &\leq \frac{1}{4} \sum_{j=M-1}^{\infty} 2^{-j}\end{aligned}$$

which implies

$$\|f - f_M\|_{L^2}^2 = O(2^{-M}).$$

Thus we have shown that the approximation error in the Haar basis decays more quickly than in the Fourier basis for our example step function.

### 1.2.2 The wavelet revolution

Our step function example is, of course, very contrived. The result that the approximation error of the Haar basis decays more quickly than the Fourier basis would only be of interest if it held for all signals in a given class and not just one particular signal. What is somewhat amazing is that similar results do hold for more general classes of functions and even more general transforms than the Haar. In fact, the Haar transform is actually the simplest example of an orthonormal wavelet transform, although Haar studied it in 1910 long before the term wavelet was coined. We will not discuss wavelets in great detail here (this has been the topic of many other Ph.D. theses) but we will mention that, like the Haar, wavelets consist of a function of zero average, $\psi(t)$, that is then scaled and translated. An excellent book on the topic is [65]. We also recommend [81] for a good introduction to filter banks and their relation to wavelets. (This relation connects our description of the discrete Haar transform and the continuous version.)

Wavelets are important for precisely the property that they tend to sparsify signals with discontinuities, while sinusoids, good at approximating uniformly regular functions, do not. Intuitively, this can be seen because smooth functions are very similar to just a few low frequency sinusoids. However, because sinusoids have support over the whole real line, each basis function will see any discontinuity in the signal. As we saw in our example, sinusoid coefficients decay slowly around

discontinuities. Wavelets, however, have a varying support that becomes more concentrated at finer scales. Because they have zero average, on the smooth parts of the signal the fine scale wavelets give small coefficients (here the signal is almost constant over the support of the fine scale wavelet) and only have large coefficients at the discontinuities. Because they have concentrated support, only a few number of wavelets at each scale will see a discontinuity.

More precisely, for $f \in L^2[0,1]$, we say that the approximation error in the Fourier basis obeys

$$\|f - f_M\|_{L^2}^2 = O(M^{-2s})$$

if and only if $f \in W^s[0,1]$ where $W^s[0,1]$ is a Sobelev space [65]. The linear Fourier approximation decays quickly if and only if $f$ has a large regularity exponent in the sense of Sobelev. Moreover, if $f$ is discontinuous then $f \notin W^s[0,1]$ for any $s > 1/2$ and so the approximation error must decay more slowly than $O(M^{-1})$. (Note that we have only been discussing linear approximation error in the Fourier case; we have been projecting onto the $M$ lowest frequencies. It is possible that if one makes a nonlinear approximation, i.e., projects onto the $M$ largest coefficients, one could do better.)

However, wavelet coefficients still decay quickly around discontinuities in one dimension. Using a wavelet basis adapted to $L^2[0,1]$ (so they behave nicely at the boundaries) are compactly supported and are $C^q$ with $q$ vanishing moments, then if $f$ has a finite number of discontinuities on $[0,1]$ and is uniformly Lipschitz $\alpha < q$ between the discontinuities, then the nonlinear approximation error obeys [65]

$$\|f - f_M\|_{L^2}^2 = O(M^{-2\alpha}).$$

Thus if $\alpha > 1/2$ then the wavelet decay is faster than the Fourier decay. The more regular $f$ is between its discontinuities, the more dramatic the improvement over Fourier is.

All of these results lead us to the slogan

*"Wavelets sparsify piecewise smooth functions in one dimension."*

It is precisely this sparsifying property of wavelets that make them so important.

Because discontinuities play such an important role in images (often changes in intensity mark edges), the wavelet transform replaced the discrete cosine as the transform used in JPEG-2000 [75], leading to better compression schemes.

### 1.2.3 The search for a better signal representation

Because of the great success of wavelets, there soon developed an entire industry of people creating new representations with good coefficient decay for certain classes of signals. (We mention here in particular that curvelets [20, 21] have optimal coefficient decay for $C_1$ curves in $\mathbb{R}^2$, which has important consequences in image processing.)

However, it soon became clear that most signals of interest contain combinations of features that are not expressed well in any *one* basis. By considering an over-complete representation, known as a dictionary, one could hope to develop richer and more flexible signal representations. Unfortunately, without a basis, signals no longer have unique representations and so focus turned to finding the sparsest signal representation in a given dictionary.

In this thesis we are almost always concerned with finite, discrete signals. In other words, the signal $f \in \mathbb{R}^n$ is a vector and the dictionary $\Psi \in \mathbb{R}^{n \times N}$ is a matrix with more columns than rows. We would like to find the sparsest $x$ such that $\Psi x = f$. Denoting $\|\cdot\|_{\ell_0}$ as the $\ell_0$ quasi-norm, which counts the number of nonzero elements in a vector (our notation is somewhat misleading as it is not an actual norm), we write this as

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_0} \quad \text{such that} \quad \Psi\tilde{x} = f.$$

This can be viewed as trying to find the sparsest solution to an underdetermined system of linear equations, computationally no easy task [35, 70]. To the best of our knowledge it involves a combinatorial search—checking to see if the signal is one of the elements/columns of the dictionary. If not, extracting all pairs of elements of the dictionary and seeing if the signal lives in the subspace spanned by them. If not, taking all triplets of elements of the dictionary, etc. We note that in the very least

the signal will be able to be represented by $n$ elements, as the dictionary is always assumed to be full rank.

Alternatively, one might be satisfied with a sparse representation that is merely close to $f$. In other words, we would like to find the sparsest $x$ such that $\|\Psi x - f\|_{\ell_2}$ is small, i.e., find the solution to

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_0} \quad \text{such that} \quad \|\Psi\tilde{x} - f\|_{\ell_2} < \epsilon.$$

Solving either of these problems seems very difficult, and so we pause here to discuss some closely related problems that arise in statistics.

## 1.3 Sparsity and statistical estimation

Not only is sparsity useful in compression and approximation, but also in statistical estimation and nonparametric regression. We have observations $y \in \mathbb{R}^n$ of an unknown signal $s \in \mathbb{R}^n$ in noise. In other words, we have the setup

$$y = s + z.$$

We assume that the noise is i.i.d. white noise, $z \sim N(0, \sigma^2 I)$. We would like to estimate $s$. If nothing at all is known about $s$, a reasonable estimate of $s$, $\hat{s}$, computed via maximum likelihood, is $\hat{s} = y$. This gives an expected mean squared error of

$$\mathbf{E}\|\hat{s} - s\|_{\ell_2}^2 = n\sigma^2.$$

However, if $s$ is sparse, by which we mean most components of $s$ are zero or close to zero, this information can be used to construct an estimator that exploits this as much as possible, possibly lowering the expected mean squared error.

More specifically, by using a *thresholding rule*, we can obtain an estimate that is almost as good as what we would obtain with the assistance of an oracle. An oracle, in the spirit of the Greeks, is something that provides information that is not normally available.

Say the oracle tells us which coefficients of $s$ are above the noise level, i.e.,

the oracle gives the set $J^\star \subset \{1, \ldots, n\}$ such that $|s_j| > \sigma \ \forall j \in J^\star$. Using this information we obtain the ideal estimate $\hat{s}_{\text{Ideal}}$ (ideal because in reality we do not have access to the oracle information), which is just the minimized empirical risk,

$$\hat{s}_{\text{Ideal}} = \arg \min_{\tilde{s}: supp(\tilde{s}) = J^\star} \|y - \tilde{s}\|_{\ell_2}^2.$$

This gives

$$\hat{s}_{\text{Ideal}} = \begin{cases} y_j & |s_j| > \sigma \\ 0 & |s_j| \leq \sigma, \end{cases}$$

and a simple calculation shows

$$\mathbf{E}\|\hat{s}_{\text{Ideal}} - s\|_{\ell_2}^2 = \sigma^2 |J^\star| + \sum_{j \notin J^\star} s_j^2.$$

Now obviously this is not a real estimator as it depends on the unknown $s$ (which is what we are trying to estimate!), but we note that it has an MSE close to $|J^\star|\sigma^2$. (Off of $J^\star$ the signal is below the noise level.) If the sparsity is small, this MSE could be considerably smaller than $n\sigma^2$. Thus we would be very happy if we could find a real estimate such that

$$\mathbf{E}\|\hat{s} - s\|_{\ell_2}^2 \approx \mathbf{E}\|\hat{s}_{\text{Ideal}} - s\|_{\ell_2}^2.$$

In fact, in [46], they essentially do just that when the estimate comes from hard or soft thresholding. Hard thresholding is implemented with

$$\hat{s}_{\text{hard}} = \begin{cases} y_j & |y_j| > \lambda \\ 0 & |y_j| \leq \lambda, \end{cases}$$

while soft thresholding is implemented with

$$\hat{s}_{\text{soft}} = \begin{cases} y_j - \lambda & y_j > \lambda \\ y_j + \lambda & y_j < -\lambda \\ 0 & |y_j| \leq \lambda. \end{cases}$$

The threshold $\lambda$ is generally chosen so that there is a high probability that it is just above the maximum level of the noise coefficients. For our Gaussian noise vector $z$, this implies that $\lambda = \sigma\sqrt{2\log n}$.

What is proven in [46] is that

$$\mathbf{E}\|\hat{s} - s\|_{\ell_2}^2 \leq (2\log n + 1)\left(\sigma^2 + \mathbf{E}\|\hat{s}_{\text{Ideal}} - s\|_{\ell_2}^2\right),$$

where $\hat{s}$ is either the hard or soft thresholding estimate. This is what is known as an *oracle inequality*. The expected mean squared error of the estimate achieves the ideal expected mean-squared error up to constants and a log factor, all without the use of an oracle. In other words, neglecting the log factor and constants, the hard or soft thesholding estimate MSE is about $s\sigma^2$ instead of $n\sigma^2$, which is what we would get from the naive maximum likelihood estimate. We have managed to exploit the sparsity of the signal to find an estimate with a lower MSE.

It is important to note that this method works not only if the signal is sparse, but also if it has a sparse representation in an orthonormal basis. To see this, take $W$ to be the orthonormal basis, and write $f = W^* s$, where $f$ is the (non-sparse) signal and $s$ is the sparse representation of it in the basis $W$. Then we have

$$y = f + z \quad \iff \quad \tilde{y} = s + \tilde{z}$$

where $\tilde{y} = Wy$ and $\tilde{z} = Wz$. Because $\tilde{z} \sim N(0, \sigma^2 I)$, and $s$ is sparse, we can apply thresholding to obtain $\hat{s}$. Transforming back gives $\hat{f} = W^*\hat{s}$. Because $\|\hat{s} - s\|_{\ell_2} = \|\hat{f} - f\|_{\ell_2}$ via Parseval, we also have oracle optimality for $f$.

We have already noted that certain signals have coefficients that decay very rapidly in a wavelet basis. Thus thresholding can be an effective way to denoise these signals; this is the famous wavelet thresholding result of Donoho and Johnstone [46].

(See also [58] for a nice introduction and description of the results.)

### 1.3.1 Connection with model selection and linear regression

What we have just been discussing has close connections with model selection. Model selection is the task of selecting a model from a list of potential models, balancing goodness of fit and complexity (bias and variance). When dealing with a sparse signal, it makes sense to consider a model to consist of all exactly sparse signals with a given support $J$ and the collection of models to contain all possible subsets $J$. In other words, a given model $S(J)$ can be written

$$S(J) = \{s \in \mathbb{R}^n : s_j = 0 \ \forall j \notin J\}.$$

For a given model $S(J)$, we calculate the best estimate of $s$ as

$$\begin{aligned} \hat{s}(J) &= \arg \min_{\tilde{s} \in S(J)} \|\tilde{s} - y\|_{\ell_2}^2 \\ &= \begin{cases} y_j & j \in J \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Ideally, an oracle would then select among all the models the one which minimizes the expected error between $s$ and $\hat{s}(J)$, i.e.,

$$\begin{aligned} J^\star &= \arg \min_{J \subset \{1,\dots,n\}} \mathbf{E}\|\hat{s}(J) - s\|_{\ell_2}^2 \\ &= \arg \min_{J \subset \{1,\dots,n\}} |J|\sigma^2 + \sum_{j \notin J} s_j^2 \\ &= \{j : \ |s_j| > \sigma\}. \end{aligned}$$

This gives the ideal estimate $\hat{s}(J^\star)$ as

$$\hat{s}(J^\star) = \begin{cases} y_j & |s_j| > \sigma \\ 0 & \text{else} \end{cases}$$

and

$$\mathbf{E}\|\hat{s}(J^\star) - s\|_{\ell_2}^2 = \sum_{j=1}^{n} \min(\sigma^2, s_j^2).$$

Again, we would like to find a real estimator $\hat{s}$ so that

$$\mathbf{E}\|\hat{s} - s\|_{\ell_2}^2 \approx \mathbf{E}\|\hat{s}(J^\star) - s\|_{\ell_2}^2.$$

Now, it is not hard to show that the hard thresholding estimate is the solution to the following optimization program

$$\min_{\tilde{s}} \|y - \tilde{s}\|_{\ell_2}^2 + \lambda^2 \|\tilde{s}\|_{\ell_0},$$

and that soft thresholding estimate is the solution to

$$\min_{\tilde{s}} \|y - \tilde{s}\|_{\ell_2}^2 + 2\lambda \|\tilde{s}\|_{\ell_1},$$

where the $\ell_1$ norm is $\|x\|_{\ell_1} = \sum_i |x_i|$. We point out that these optimizations involve minimizing a tradeoff between goodness of fit, the squared error between $y$ and the estimate, and a penalty that measures the complexity in either the $\ell_0$ or $\ell_1$ norms, weighted by a parameter $\lambda$. This tradeoff characterizes model selection problems.

There are some very nice results in the model selection literature that achieve oracle inequality results for more general setups than what we have discussed (more general classes of models, non Gaussian noise, etc.) and explore under what conditions an oracle inequality is obtainable and what the value of $\lambda$ should be, see [6, 9] and the references therein. See also [88] for a very readable discussion of thresholding and oracle inequalities.

One common theme of all these model selection results is that the complexity term in the minimization is always measured by $\ell_0$. This makes a lot of intuitive sense, as $\ell_0$ is the natural measure of the size of the model—the number of nonzero elements it contains. In the simple case of hard thresholding that we discussed above, the minimization can be performed and the estimate explicitly calculated. In more general setups, however, this is no longer the case.

More specifically, in the case of linear regression where there are more regressors

than parameters, one is faced with the problem of solving

$$\min_{\tilde{x}} \|A\tilde{x} - y\|_{\ell_2}^2 + \lambda \|\tilde{x}\|_{\ell_0},$$

which seems to have no computationally feasible way to determine $\hat{x}$. Again, we are stymied by the computational intractability of $\ell_0$.

## 1.4 Problem setups at the focus of this thesis

Thus far we have been motivating why sparsity is interesting through examples in signal processing, approximation theory and statistics, emphasizing the importance of fast algorithms and finding sparse signal representations, where sparsity can be thought of as having only a few nonzero coefficients, or coefficients that decay quickly. In searching for sparse signal representations for a general class of signals, the problems became much more difficult when the dictionary became overcomplete. Similarly, in model selection/sparse linear regression, when the number of regressors became more than the number of response variables, easily computable solutions ceased to exist. These types of problems are exactly the focus of this thesis. (We pause here to briefly mention that the difficulty really arises when the dictionary elements or regressors are no longer orthogonal, which can also happen in the undercomplete case. In fact, many of the results we will discuss also apply to undercomplete setups. However, we will mostly focus on the overcomplete case because until recently it was not as well-studied, while a growing number of important problems fit into this setup.)

We will almost exclusively be interested in real-valued, finite dimensional, discrete signals, hence our signals are vectors $x \in \mathbb{R}^N$ and $A \in \mathbb{R}^{n \times N}$ is a matrix with more columns than rows. $A$ is always assumed to be full rank and, unless otherwise stated, has unit normalized columns. We are interested in both noisy and noiseless setups, where the noise term $z \in \mathbb{R}^n$ is stochastic (in which case we will always assume that it is Gaussian) or deterministic (in which case we will always assume that it is bounded, i.e., $\|z\|_{\ell_2} \leq \epsilon$ for some $\epsilon > 0$). The signal $x$ is of course sparse, meaning that it has only a few nonzero terms. Sometimes we will say that a signal

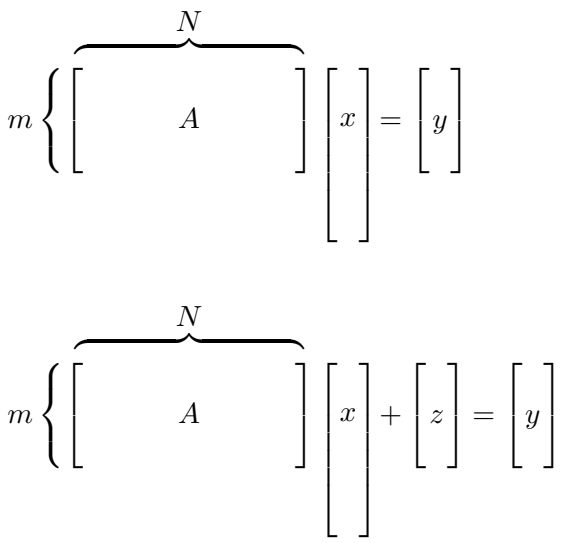



**Figure 1.5.** Depiction of problem setups at the heart of this thesis. Note that the matrix $A$ has more rows than columns. From $y$ we would like to find a good estimate of $x$, or possibly $Ax$ in the case of noise, using a computationally reasonable method.

is sparse when it has only a few large terms and the rest are small but maybe not zero. We will also refer to signals with decaying coefficients as sparse. Which precise definition of sparsity we mean should be clear from context.

In this setup, we are interested in estimating $x$ or $Ax$ in a computationally tractable way. We would like to know under what conditions on $A$ and the sparsity of $x$ can we achieve good estimates? In other words, when can we guarantee that $\|\hat{x} - x\|_{\ell_2}$ or $\|A\hat{x} - Ax\|_{\ell_2}$ is small? How close can we come to what can be achieved with an oracle?

Sometimes to emphasize that we are thinking of $A$ as a dictionary we will write it as $\Psi$. Similarly, to emphasize that we are thinking of $A$ as a measurement matrix, we will write it as $\Phi$. This idea of taking linear measurements of sparse signals arises in the field of compressed sensing which we describe further in Section 1.7.2.

## 1.5 Computationally tractable algorithms

We turn now to the problem of finding computationally tractable methods to attack our problems. There have been several different approaches, and in the sequel we discuss two that have met with some success, namely greedy methods in the form

matching pursuits and its cousins, and methods that replace the $\ell_0$ quasi-norm with the convex $\ell_1$ norm. We by no means want to suggest that these are the only techniques available, but they are of interest to us because they have met with the most provable successes.

### 1.5.1 Greedy algorithms

In the context of finding a sparse representation of a signal in a dictionary $\Psi$ containing $N > n$ vectors $\{\psi_i\}$ in $\mathbb{R}^n$, Mallat and Zhang introduced a greedy algorithm they called Matching Pursuit [66]. In what follows we will assume the elements of the dictionary have unit norm and the dictionary has full rank. Matching Pursuit selects the element of the dictionary that is most correlated with the signal $f$, projects the signal onto that element and then selects the element of the dictionary that is most correlated with the residual. This process is then repeated.

In other words, letting $r_0 = f$, we have

$$\psi_{n+1} = \arg\max_{\psi_i \in \Psi} |\langle r_n, \psi_i \rangle|$$

$$r_{n+1} = r_n - \langle r_n, \psi_{n+1} \rangle \psi_{n+1}.$$

We note that $r_{n+1}$ is orthogonal to $\psi_{n+1}$.

By summing the second equation from $n = 0$ to $n = M - 1$ we have

$$f = \sum_{n=0}^{M-1} \langle r_n, \psi_{n+1} \rangle \psi_{n+1} + r_M.$$

It can be shown that for finite dimensional signals the residual converges exponentially to zero as the number of iterations goes to infinity [35], and so we have

$$f = \sum_{n=0}^{\infty} \langle r_n, \psi_{n+1} \rangle \psi_{n+1}.$$

However, the rate of convergence of the residual decreases as the dimension of the signal increases, and for infinite dimensional signals the convergence is no longer exponential. Also, even in the case of finite dimensional signals, an infinite number of iterations is necessary to completely reduce the residual. This is because the

element of the dictionary selected on the $n$th iteration, $\psi_n$, is not guaranteed to be orthogonal to the previously selected dictionary elements $\{\psi_i\}_{1\leq i<n}$. Thus even though the new residual is orthogonal to $\psi_n$, when subtracting the projection of $r_{n-1}$ over $\psi_n$ the algorithm reintroduces new components in the directions of the $\{\psi_i\}_{0\leq i<n}$.

This procedure can be improved by implementing a version called Orthogonal Matching Pursuit (OMP) [73, 36] where the residual is orthogonally projected onto the space spanned by all the previously selected elements of the dictionary. In other words we have

$$\begin{aligned}\psi_{n+1} &= \arg\max_{\psi_i\in\Psi} |\langle r_n, \psi_i\rangle| \\ r_{n+1} &= r_n - \arg\min_{\tilde{f}\in \operatorname{span}\{\psi_i\}_{1\leq i\leq n+1}} \|r_n - \tilde{f}\|_{\ell_2}.\end{aligned}$$

This procedure is guaranteed to give a zero residual after at most $n$ steps because $f \in \mathbb{R}^n$ and after $n$ steps the $n$ selected elements of the dictionary span $\mathbb{R}^n$. One of the most attractive features of OMP is that it is computationally feasible and admits simple, fast implementations.

We conclude this section by noting that greedy methods have different names in different fields. In signal processing, as we have been discussing, they are known as pursuits, in statistics they are known as forward stepwise regression, and in approximation theory they are called greedy algorithms.

### 1.5.2 Convex relaxation algorithms

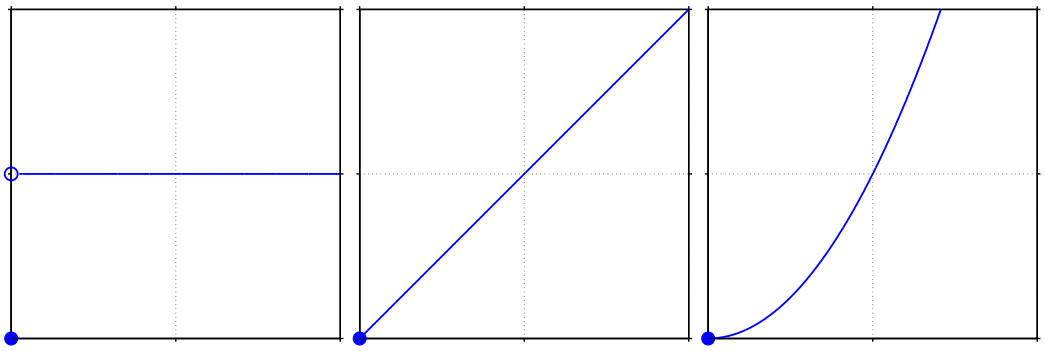
Besides greedy methods, another popular computationally tractable approach to our problems is to replace $\ell_0$ with the convex norm $\ell_1$. For example, in the case of finding a sparse representation of a signal in a dictionary, instead of looking for the solution of

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_0} \quad \text{such that} \quad A\tilde{x} = y$$

we would solve

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad A\tilde{x} = y. \tag{1.1}$$

This is often called Basis Pursuit [31].



**(a)** $\ell_0$ penalty **(b)** $\ell_1$ penalty **(c)** $\ell_2$ penalty

**Figure 1.6.** Plots of the $\ell_0$, $\ell_1$ and $\ell_2$ penalty functions. Note that $\ell_1$ maximally penalizes small but nonzero terms, while still maintaining convexity.

In the context of linear regression and model selection, instead of looking for the solution of

$$\min_{\tilde{x}} \|A\tilde{x} - y\|_{\ell_2}^2 + \lambda \|\tilde{x}\|_{\ell_0}$$

we would solve

$$\min_{\tilde{x}} \|A\tilde{x} - y\|_{\ell_2}^2 + \lambda \|\tilde{x}\|_{\ell_1}. \tag{1.2}$$

This is called the lasso [83].

Another popular setup for noisy measurements $y = Ax + z$, where $z$ satisfies $\|z\|_{\ell_2} < \epsilon$, is

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|A\tilde{x} - y\|_{\ell_2} < \epsilon. \tag{1.3}$$

The advantage of using the $\ell_1$ norm is that it is convex (a function is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

for any $x, y$ in the domain of $f$ and any $0 \leq t \leq 1$) and so (1.1), (1.2) and (1.3) can be written as convex programs and all the well established machinery of convex optimization can be used. In particular, there exist fast interior point and log barrier algorithms [10] that can be employed.

The use of $\ell_0$ always involves sparsity, whether looking for a sparse dictionary representation of a signal or a simple model that well represents given data. But why do we expect intuitively that by minimizing an $\ell_1$ norm we will arrive at a

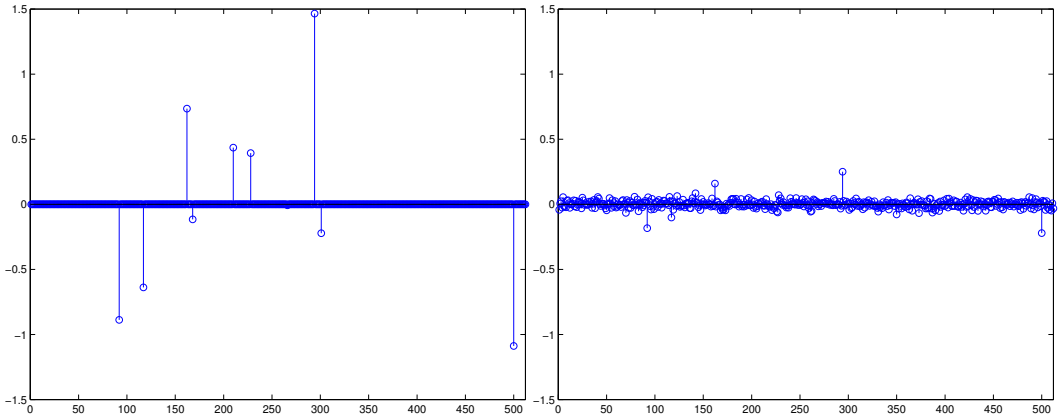



**Figure 1.7.** 1.7a shows a sparse signal with ten nonzero terms. We form $y = Ax$ where $A$ is a $256 \times 512$ matrix with i.i.d. Gaussian entries. 1.7b is the minimum energy solution, i.e., the solution to $\min_{\tilde{x}} \|\tilde{x}\|_{\ell_2}$ such that $y = A\tilde{x}$. Minimizing $\ell_1$ instead of $\ell_2$, however, recovers the sparse signal exactly.

sparse solution? In some sense this is because $\ell_1$ penalizes small but nonzero terms as much as possible while maintaining convexity, and so instead of favoring solutions with many small terms, it favors solutions with terms that are either zero or not—in other words, sparse solutions. This should be compared with the $\ell_2$ penalty, for example, which is also convex, but which penalizes small but nonzero terms less and tends to produce dense solutions, see Figures 1.6 and 1.7.

We mention here that the sparsity promoting properties of $\ell_1$ had been empirically observed for decades in geophysics. In reflection seismology, $\ell_1$ minimization had been proposed and sucessfully used to determine the boundaries between subsurface layers of the earth [82, 32, 78], while results quantifying the ability of $\ell_1$ to recover sparse reflection traces began to appear in [48, 47].

## 1.6 Brief survey of known results

We give now a brief survey of known results in the field. We make no pretense of being exhaustive, but instead try to give a sense of what types of results have been proven, with an emphasis on results that have inspired or are closely related to results in this thesis. A good review article on some of the early work is [11].

### 1.6.1 Coherence results

All of the earliest sparse recovery results rely on a property known as coherence. The coherence of a matrix $A$ is

$$\mu(A) = \max_{i \neq j} \frac{|\langle a_i, a_j \rangle|}{\|a_i\|_{\ell_2} \|a_j\|_{\ell_2}},$$

where the $a_i$ are columns of the matrix $A \in \mathbb{R}^{m \times N}$. Coherence is basically a measure of how similar the columns of a matrix are. Notice that $\mu \leq 1$, and for an orthogonal matrix $\mu(A) = 0$. In fact, one can show that for a general matrix $A$

$$\sqrt{\frac{N-m}{m(N-1)}} \leq \mu(A) \leq 1.$$

Some of the earliest sparse recovery results [43, 55] showed that in the noiseless case, $y = Ax$, if the sparsity $s$ of $x$ was smaller than the inverse of the coherence of $A$, $s < c/\mu(A)$, for a given constant $c$, then exact recovery of $x$ using the convex relaxation (1.1) was possible. As for greedy methods, for awhile it was thought that greedy algorithms were unpromising as tools for sparse recovery because researchers found examples of specific $A$ and sparse $x$ when a greedy approach returned a fully dense $\hat{x}$ [37]. However, progress was made in [54] and refined in [84] that also showed that if $s < O(1/\mu(A))$, then exact recovery of $x$ using OMP was achieved.

Slightly later, [44] showed stability results in the noisy setup $y = Ax + z$ for $\|z\|_{\ell_2} < \epsilon$, for both OMP and $\ell_1$ methods. In this case, exact recovery of $x$ is no longer possible, but it is shown that $\hat{x}$ is still close to $x$. For example, in the case of $\ell_1$, they showed that if $s < 1/4(1/\mu + 1)$ then

$$\|x - \hat{x}\|_{\ell_2} \leq \frac{\epsilon}{\sqrt{1 - \mu(s-1)}}.$$

A result of a similar flavor for $\ell_1$ is given in [85]. In the case of stochastic noise, [12, 13] develop oracle inequality bounds on $\|Ax - A\hat{x}\|_{\ell_2}$ and $\|x - \hat{x}\|_{\ell_1}$ using the lasso, again assuming $s < O(1/\mu)$.

However, these results did not seem to be capturing the whole story. For example, a dictionary made up of $n$ spikes and $n$ sines, $\Psi = [I \ F^*]$, has coherence

$\mu = 1/\sqrt{n}$. Restricting $s < O(1/\mu)$ gives $s < O(\sqrt{n})$ and examples of signals could be constructed so that if $s = \sqrt{n}$ then recovery of the signal was not possible. However, numerical experiments [48, 45] seemed to show both $\ell_1$ and greedy methods working well for much larger sparsities, and that the the example signals where recovery was not possible were somehow pathological. Focus then turned to showing that the methods could work well for larger sparsities.

### 1.6.2 Uniform uncertainty results

The first paper to show good recovery for less restrictive sparsities was [26]. They show that if $A$ is a Fourier matrix with randomly selected rows and $x$ is a sparse vector with random support of size $s$, then by solving the $\ell_1$ minimization program (1.1), if $s < Cm/\log(N)$, one recovers $x$ with high probability. The key to their proof was showing that submatrices of $A$ formed by selecting columns of $A$ had well behaved singular values.

Similar results were obtained in [42] and [87] for $A$ a Gaussian matrix with i.i.d. entries where one solves Basis Pursuit or OMP, repectively. Again, the results hinged on the singular values of submatrices of $A$. In [29] and [28] these singular value conditions were formalized as the *restricted isometry property* (RIP), also refered to as the *uniform uncertainty principle* (UUP), a condition on the restricted isometry constants of the matrix $A$, which are defined as

**Definition 1.6.1.** *For each integer $s = 1, 2, \ldots$, define the isometry constant $\delta_s$ of a matrix $A$ as the smallest number such that*

$$(1 - \delta_s)\|x\|_{\ell_2}^2 \leq \|Ax\|_{\ell_2}^2 \leq (1 + \delta_s)\|x\|_{\ell_2}^2$$

*for all $s$-sparse vectors $x$. A vector is said to be $s$-sparse if it has at most $s$ nonzero entries.*

This basically says that $A$ acts almost like an isometry on sparse vectors. An equivalent way to state it is as a condition on the singular values of $A_T$,

$$1 - \delta_s \leq \lambda(A_T^* A_T) \leq 1 + \delta_s,$$

where $T \subset \{1, \ldots, N\}$, $|T| \leq s$.

In the noiseless case, it was shown in [28] that if $A$ had sufficiently small restricted isometry constants, then for all $s$-sparse $x$, solving (1.1) would determine $x$ exactly. Moreover, it was shown that for various ensembles of matrices, the isometry constants $\delta_s$ are small. For example, consider $A$ drawn from the following ensembles of matrices:

- *The Gaussian ensemble.* The entries of $A$ are i.i.d. Gaussian $N \sim (0, 1/m)$.
- *The binary ensemble.* The entries of $A$ are i.i.d. entries from the Bernoulli distribution, $\mathbf{P}(A_{ij} = \pm 1/\sqrt{m}) = 1/2$, or some other subguassian distribution.
- *The Fourier ensemble.* $A$ is a $N \times N$ Fourier matrix with $m$ rows sampled uniformly at random.

In the case of the first two ensembles, the restricted isometry property is satisfied with high probability if [28, 69]

$$s \leq Cm/\log(N/m),$$

while in the third it is satisfied if [29]

$$s \leq Cm/\log(N)^6.$$

Attention then shifted to showing stable recovery in the presence of noise for larger sparsities. Results include [41, 27] for the case of deterministic noise, $\|z\|_{\ell_2} < \epsilon$, when one solves (1.3). The results in [27] and refined in [19] required that the isometry constant $\delta_{2s}$ be sufficiently small, and were particularly nice in that even if $x$ was not exactly sparse one could still show

$$\|\hat{x} - x\|_{\ell_2} < C_{\delta_1} \frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + C_{\delta_2} \epsilon$$

where $C_{\delta_1}$ and $C_{\delta_2}$ are constants that can be explicitly stated in terms of the restricted isometry constant $\delta_{2s}$, and $x_s$ is the best $s$-term approximation of $x$.

In the case of stochastic noise, [30] again required that the restricted isometry constants be sufficiently small, but instead of solving (1.2) or (1.3), it introduced

the following linear program, called the Dantzig selector

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|A^*(A\tilde{x} - y)\|_{\ell_\infty} < \lambda. \tag{1.4}$$

Nice oracle inequality bounds were achieved for $\|x - \hat{x}\|_{\ell_2}$. The Dantzig selector is a close cousin of the lasso, and it is no surprise that there are many similar results for the solution of (1.2); we mention in particular [8, 94, 68, 93]. These results all require that a UUP-like requirement on the singular values of $A$ be met in order to achieve good recovery.

On the greedy algorithm front, in [72] it was shown that a greedy algorithm called ROMP was able to achieve similar results as in [27], again requiring that the restricted isometry constants of $A$ be sufficiently small, with some spurious log factors. In [71] these log factors were removed using a greedy algorithm called CoSaMP.

### 1.6.3 And back to coherence

While the uniform uncertainty results are impressive because they hold for all sufficiently sparse $x$, they also have drawbacks. For example, while it is possible to show that a matrix drawn from one of the ensembles discussed earlier obeys the UUP with high probability, for a particular matrix of interest it can be a difficult condition to check, likely involving calculating the singular values of all submatrices of the matrix—a combinatorial calculation. Also, because of its uniformity property, it is a rather strong requirement on *all* of the submatrices of $A$. Perhaps we would be satisfied if, instead of for any $x$, we had a signal model for $x$, and would be happy instead of being guaranteed good reconstruction, we got good reconstruction with high probability. Finally, there are matrices such as $[I \ F^*]$ that can be shown to violate the UUP for fairly small sparsities, but numerically still seem to perform well.

There are results that show for a specific matrix of interest, one can still achieve good reconstruction if $x$ is taken from a random signal model. (See [26, 15], for example.) It would be nice to find a condition that is easily verifiable that would capture some of this behavior for sparsities greater than $O(1/\mu)$. A result along

these lines is [22] where the concept of coherence again comes into play. They show that if the coherence of $A$ is sufficiently small

$$\mu(A) \leq c_0 \cdot (\log N)^{-1},$$

and $x$ comes from a suitable statistical model, then the solution to (1.2) achieves an oracle inequality for $Ax$ with high probability if $s \leq c_1 N/(\|A\|^2 \log N)$. Chapter 3 of this thesis extends these results to the Dantzig selector and oracle inequalities involving $x$.

## 1.7 Applications

We have already mentioned that our problems of interest are motivated by signal compression, approximation, and linear regression problems. However, we pause here to mention a few more example applications in greater detail.

### 1.7.1 Error correction

At first glance, it may appear that error correction has very little to do with the types of problems and setups we have been discussing. After all, in a typical error correction setup, redundancy is added to the message signal to be sent in hopes that the redundancy will help remove errors. Thus linear coding matrices have more rows than columns, the exact opposite situation of our matrix $A$. In addition, the messages to be sent typically are not at all sparse. However, it *is* reasonable to assume that the errors are sparse, and this is the connection with our problems.

In an error correction setup over the reals with a linear encoding matrix one has

$$y = Mf + e$$

where $y \in \mathbb{R}^N$ is the received, corrupted message, $M \in \mathbb{R}^{N \times n}$ is the rull rank encoding matrix with $N > n$, $f \in \mathbb{R}^n$ is the message to be sent, and $e \in \mathbb{R}^N$ is a vector of sparse errors. If one multiplies through by an annihilation matrix $A$ such that $A \in \mathbb{R}^{(N-n)\times N}$ and $AM = 0$ (take, for example, $A$ to be the orthogonal

projection onto the null space of $M$) then we have

$$\tilde{y} = Ae$$

where $\tilde{y} = Ay$ and $A$ has more columns than rows. If we can solve for (or well-approximate) the sparse $e$ then we can reconstruct $f$ because $M$ is known. In [28] the problem is explored when $e$ is exactly sparse, and in Chapter 2 of this thesis, the problem is explored when $e$ is a combination of sparse arbitrary errors and smaller dense errors.

### 1.7.2 Compressed sensing

A much discussed example of our setup is known as compressed sensing or compressive sampling [40, 18]. In the traditional method of signal compression which we have discussed earlier, a signal is sensed or sampled, transformed into a basis where it is sparse, and then most of the transformed data thrown out. This raises the question, why go through all the effort of acquiring so much data when in the end so much is discarded? Is it possible to somehow take clever measurements of a signal so that the number of measurements is close to the amount of important information contained in the signal?

In the language of our setup, we view $A$ as a measurement matrix, which traditionally is called $\Phi$. In other words, we take linear measurements of a sparse signal $x$ (or a signal $f$ that can be sparsely represented in a basis) and look for answers to questions like: How few measurements can we take and still be able to reconstruct $x$ or estimate it well? What should these measurements be?

This type of setup arises in real-world applications when measurement devices naturally acquire encoded samples rather than direct signal samples, and taking many samples is undesirable or impractical. This is the case, for example, in MRI imaging, where the scanner measures Fourier coefficients of the object being scanned. (See [92] for an introdution to MRI imaging and [64] for a discussion of its relation to compressed sensing.)

One shortcoming of all the known results discussed earlier when applied to the compressed sensing setup is that it is required that the signal to be measured is

sparse or sparse in a basis. However, as noted before, sometimes it is preferable to consider signals that can be well represented in an overcomplete dictionary instead of a basis. In Chapters 4 and 5 of this thesis, we will describe compressed sensing results when the measured signal is sparse in a dictionary. In other words, we have $\Phi f = \Phi\Psi x$ where $x$ is sparse.

## 1.8 Organization of thesis

In Chapter 2 we discuss an error correction problem over the reals where a received encoded message is corrupted by a few arbitrary gross errors, as well as smaller errors affecting all the entries. We show that under suitable conditions on the encoding matrix and on the number of gross errors, one is able to accurately recover the message by solving either of two convex optimization programs. We note that Chapter 2 has appeared in [24], and also in a condensed form in [23].

In Chapter 3 we examine the statistical estimation problem of recovering a signal from noisy measurements, $y = Ax + z$. We would like to estimate $x$ or $Ax$. We explore how close one can get to oracle optimality where the estimate $x$ is required to be computed in a computationally tractable way.

In Chapter 4 we consider the problem of reconstructing a signal from a limited number of random linear measurements, where the signal can be sparsely represented by the elements in a dictionary and the dictionary obeys an incoherence property. We show that by solving an $\ell_1$ minimization program one can recover the signal exactly from the measurements.

Finally, in Chapter 5 we again explore the problem of estimating a signal from a limited number of linear measurements. However, we no longer require that the signal be exactly sparsely represented by the dictionary. We show that if the combined measurement/dictionary matrix obeys certain requirements then by solving a convex optimization program known as analysis, the signal can be accurately recovered. This nicely complements known results for the more standard synthesis approach.

We conclude this introduction by noting that all work in this thesis is joint with Emmanuel Candès. Also, we are grateful to Peter Stobbe for sharing his Gabor

dictionary code used in Chapter 5. Finally, we have done our best to make each chapter self-contained. This has led to some redundancy in definitions and proofs; we hope this will not be distracting but will instead facilitate the readability of the thesis.

# Chapter 2

# Highly robust error correction by convex programming

## 2.1 Abstract

This chapter discusses a stylized communications problem where one wishes to transmit a real-valued signal $x \in \mathbb{R}^n$ (a block of $n$ pieces of information) to a remote receiver. We ask whether it is possible to transmit this information reliably when a fraction of the transmitted codeword is corrupted by arbitrary gross errors, and when in addition, all the entries of the codeword are contaminated by smaller errors (e.g., quantization errors).

We show that if one encodes the information as $Ax$ where $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ is a suitable coding matrix, there are two decoding schemes that allow the recovery of the block of $n$ pieces of information $x$ with nearly the same accuracy as if no gross errors occur upon transmission (or equivalently as if one has an oracle supplying perfect information about the sites and amplitudes of the gross errors). Moreover, both decoding strategies are very concrete and only involve solving simple convex optimization programs, either a linear program or a second-order cone program. We complement our study with numerical simulations showing that the encoder/decoder pair performs remarkably well.

## 2.2 Introduction

This chapter discusses a coding problem over the reals. We wish to transmit a block of $n$ real values—a vector $x \in \mathbb{R}^n$—to a remote receiver. A possible way to address

this problem is to communicate the codeword $Ax$ where $A$ is an $m$ by $n$ coding matrix with $m \geq n$. Now a recurrent problem with real communication or storage devices is that some portions of the transmitted codeword may become corrupted; when this occurs, parts of the received codeword are unreliable and may have nothing to do with their original values. We represent this as receiving a distorted codeword $y = Ax + z_0$. The question is whether one can recover the signal $x$ from the received data $y$.

It has recently been shown [28, 16] that one could recover the information $x$ exactly—under suitable conditions on the coding matrix $A$—provided that the fraction of corrupted entries of $Ax$ is not too large. In greater details, [28] proved that if the corruption $z_0$ contains at most a fixed fraction of nonzero entries, then the signal $x \in \mathbb{R}^n$ is the unique solution of the minimum-$\ell_1$ approximation problem

$$\min_{\tilde{x} \in \mathbb{R}^n} \|y - A\tilde{x}\|_{\ell_1}. \tag{2.1}$$

What may appear as a surprise is the fact that this requires no assumption whatsoever about the corruption pattern $z_0$ except that it must be sparse. In particular, the decoding algorithm is provably exact even though the entries of $z_0$—and thus of $y$ as well—may be arbitrary large, for example.

While this is interesting, it may not be realistic to assume that except for some gross errors, one is able to receive the values of $Ax$ with infinite precision. A better model would assume instead that the receiver gets

$$y = Ax + z_0, \qquad z_0 = e + z, \tag{2.2}$$

where $e$ is a possibly sparse vector of gross errors and $z$ is a vector of small errors affecting all the entries. In other words, one is willing to assume that there are malicious errors affecting a fraction of the entries of the transmitted codeword *and* in addition, smaller errors affecting all the entries. For instance, one could think of $z$ as some sort of quantization error which limits the precision/resolution of the transmitted information. In this more practical scenario, we ask whether it is still possible to recover the signal $x$ accurately. The subject of this chapter is to show

that it is in fact possible to recover the original signal with nearly the same accuracy as if one had a perfect communication system in which no gross errors occur upon transmission. Further, the recovery algorithms are very concrete and practical; they involve solving very convenient convex optimization problems.

Before expanding on our results, we would like to comment on the practical relevance of our model. Coding theory generally assumes that data take on values in a finite field, but there are a number of applications where encoding over the reals is of direct interest. We give two examples. The first example concerns Orthogonal Frequency-Division Multiplexing for wireless and wideband digital communication. Here, one can experience deep fades at certain frequencies (because of mulipathing for instance) and/or frequency jamming because of strong interferers so that large parts of the data are unreliable. The second example is in the area of digital computations. Here, researchers are currently interested in error correction over the reals to protect real-valued results of onboard computations which are executed by circuits that are subject to faults due, for example, to radiation. As we will see, our work introduces an encoding strategy which is robust to such errors, which runs in polynomial time, and which provably obeys optimal bounds.

To understand the claims of this chapter in a more quantitative fashion, suppose that we had a perfect channel in which no gross errors ever occur; that is, we assume $e = 0$ in (2.2). Then we would receive $y = Ax + z$ and would reconstruct $x$ by the method of least-squares which, assuming that $A$ has full rank, takes the form

$$x^{\mathrm{Ideal}} = (A^*A)^{-1}A^*y. \tag{2.3}$$

In this ideal situation, the reconstruction error would then obey

$$\|x^{\mathrm{Ideal}} - x\|_{\ell_2} = \|(A^*A)^{-1}A^*z\|_{\ell_2}. \tag{2.4}$$

Suppose we design the coding matrix $A$ with orthonormal columns so that $A^*A = I$. Then we would obtain a reconstruction error whose maximum size is just about that of $z$. If the smaller errors $z_i$ are i.i.d. $N(0, \sigma^2)$, then the mean-squared error (MSE)

would obey

$$\mathbf{E}\|x^{\text{Ideal}} - x\|_{\ell_2}^2 = \sigma^2 \text{Tr}((A^*A)^{-1}).$$

If $A^*A = I$, then the MSE is equal to $n\sigma^2$.

The question then is, can one hope to do almost as well as this optimal mean squared error without knowing $e$ or even the support of $e$ in advance? This chapter shows that one can in fact do almost as well by solving very simple convex programs. This holds for *all* signals $x \in \mathbb{R}^n$ and *all* sparse gross errors no matter how adversary.

Two concrete decoding strategies are introduced: one based on second-order cone programming (SOCP) in Section 2.3, and another based on linear programming (LP) in Section 2.4. We introduce two different decoding strategies because in certain situations it may be preferable to solve an LP over an SOCP or vice-versa. Also, we show theoretically that the two methods scale differently, so in a particular setup one method could outperform the other. For instance, it is an open question whether or not the SOCP decoder can achieve the adaptive bounds of the LP decoder. In Section 2.5 we compare the empirical performances of the two decoders in a series of numerical experiments before proving our results in Section 2.6, followed by a discussion in Section 2.7.

We conclude the introduction by noting that this chapter is part of a larger body of work. In particular, besides the obvious connections with [28, 16], it draws on recent results [27, 30] showing that the theory and practice of compressed sensing (also known as compressive sampling) is robust vis a vis noise. The connection with this work should become clear in our proofs.

## 2.3 Decoding by second-order cone programming

To recover the signal $x$ from the corrupted vector $y$ (2.2) we propose solving the following optimization program:

$$(P_2) \quad \min \|y - A\tilde{x} - \tilde{z}\|_{\ell_1} \text{ subject to } \|\tilde{z}\|_{\ell_2} \le \varepsilon, \qquad (2.5)$$
$$A^*\tilde{z} = 0,$$

with variables $\tilde{x} \in \mathbb{R}^n$ and $\tilde{z} \in \mathbb{R}^m$. The parameter $\varepsilon$ above depends on the magnitude of the small errors and shall be specified later. The program $(P_2)$ is equivalent to

$$\begin{aligned} \min \mathbf{1}^* \tilde{u}, \quad \text{subject to} \quad & -\tilde{u} \le y - A\tilde{x} - \tilde{z} \le \tilde{u}, \\ & \|\tilde{z}\|_{\ell_2} \le \varepsilon, \\ & A^*\tilde{z} = 0, \end{aligned} \tag{2.6}$$

where we added the slack optimization variable $\tilde{u} \in \mathbb{R}^m$. In the above formulation, $\mathbf{1}$ is a vector of ones and the vector inequality $u \le v$ means componentwise, i.e., $u_i \le v_i$ for all $i$. The program (2.6) is a second-order cone program and as a result, $(P_2)$ can be solved efficiently using standard optimization algorithms, see [10].

The first key point of this chapter is that the SOCP decoder is highly robust against imperfections in communication channels. Here and below, $V$ denotes the subspace spanned by the columns of $A$, and $Q \in \mathbb{R}^{m \times (m-n)}$ is a matrix whose columns form an orthobasis of $V^\perp$, the orthogonal complement to $V$. Such a matrix $Q$ is a kind of parity-check matrix since $Q^*A = 0$. Applying $Q^*$ on both sides of (2.2) gives

$$Q^*y = Q^*e + Q^*z. \tag{2.7}$$

Now if we could somehow get an accurate estimate $\hat{e}$ of $e$ from $Q^*y$, we could reconstruct $x$ by applying the method of Least Squares to the vector $y$ corrected for the gross errors:

$$\hat{x} = (A^*A)^{-1}A^*(y - \hat{e}). \tag{2.8}$$

If $\hat{e}$ were very accurate, we would probably do very well.

The point is that under suitable conditions, $(P_2)$ provides such accurate esti-

mates. Introduce $\tilde{e} = y - A\tilde{x} - \tilde{z}$, and observe the following equivalence:

$$\begin{aligned} (P_2) \quad \Leftrightarrow \quad & \min \quad && \|\tilde{e}\|_{\ell_1} \\ & \text{subject to} \quad && \tilde{e} = y - A\tilde{x} - \tilde{z}, \\ & && A^*\tilde{z} = 0, \;\; \|\tilde{z}\|_{\ell_2} \leq \varepsilon, \\ \Leftrightarrow \quad (P_2') \quad & \min \quad && \|\tilde{e}\|_{\ell_1} \\ & \text{subject to} \quad && \|Q^*(y - \tilde{e})\|_{\ell_2} \leq \varepsilon. \end{aligned} \tag{2.9}$$

We only need to argue about the second equivalence since the first is immediate. Observe that the condition $A^*\tilde{z} = 0$ decomposes $y - \tilde{e}$ as the superposition of an arbitrary element in $V$ (the vector $A\tilde{x}$) and of an element in $V^\perp$ (the vector $\tilde{z}$) whose Euclidean length is less than $\varepsilon$. In other words, $\tilde{z} = P_{V^\perp}(y - \tilde{e})$ where $P_{V^\perp} = QQ^*$ is the orthonormal projector onto $V^\perp$ so that the problem is that of minimizing the $\ell_1$ norm of $\tilde{e}$ under the constraint $\|P_{V^\perp}(y - \tilde{e})\|_{\ell_2} \leq \varepsilon$. The claim follows from the identity $\|P_{V^\perp} v\|_{\ell_2} = \|Q^* v\|_{\ell_2}$ which holds for all $v \in \mathbb{R}^m$.

The equivalence between $(P_2)$ and $(P_2')$ asserts that if $(\hat{x}, \hat{z})$ is solution to $(P_2)$, then $\hat{e} = y - A\hat{x} - \hat{z}$ is solution to $(P_2')$ and vice versa; if $\hat{e}$ is solution to $(P_2')$, then there is a unique way to write $y - \hat{e}$ as the sum $A\hat{x} + \hat{z}$ with $z \in V^\perp$, and the pair $(\hat{x}, \hat{z})$ is solution to $(P_2)$. We note, and this is important, that the solution $\hat{x}$ to $(P_2)$ is also given by the corrected least squares formula (2.8). Equally important is to note that even though we use the matrix $Q$ to explain the rationale behind the methodology, one should keep in mind that $Q$ does not play any special role in $(P_2)$.

The issue here is that if $\|P_{V^\perp} v\|_{\ell_2}$ is approximately proportional to $\|v\|_{\ell_2}$ for all sparse vectors $v \in \mathbb{R}^m$, then the solution $\hat{e}$ to $(P_2')$ is close to $e$, provided that $e$ is sufficiently sparse [27]. Quantitatively speaking, if $\varepsilon$ is chosen so that $\|P_{V^\perp} z\|_{\ell_2} \leq \varepsilon$, then $\|e - \hat{e}\|$ is less than a numerical constant times $\varepsilon$; that is, the reconstruction error is within the noise level. The key concept underlying this theory is the so-called *restricted isometry property.*

**Definition 2.3.1.** *Define the isometry constant $\delta_k$ of a matrix $\Phi$ as the smallest number such that*

$$(1 - \delta_k)\|x\|_{\ell_2}^2 \leq \|\Phi x\|_{\ell_2}^2 \leq (1 + \delta_k)\|x\|_{\ell_2}^2 \tag{2.10}$$

*holds for all k-sparse vectors x (a k-sparse vector has at most k nonzero entries).*

In the sequel, we shall be concerned with the isometry constants of $A^*$ times a scalar. Since $AA^*$ is the orthogonal projection $P_V$ onto $V$, we will be thus interested in subspaces $V$ such that $P_V$ nearly acts as an isometry on sparse vectors. Our first result states that the SOCP decoder is provably accurate.

**Theorem 2.3.2.** *Choose a coding matrix $A \in \mathbb{R}^{m\times n}$ with orthonormal columns spanning $V$, and let $(\delta_k)$ be the isometry constants of the rescaled matrix $\sqrt{\frac{m}{n}}\,A^*$. Suppose $\|P_{V^\perp} z\|_{\ell_2} \le \varepsilon$. Then the solution $\hat{x}$ to $(P_2)$ obeys*

$$\|\hat{x} - x\|_{\ell_2} \le C_2 \cdot \frac{\varepsilon}{\sqrt{1-\frac{n}{m}}} + \|x^{\mathrm{Ideal}} - x\|_{\ell_2} \tag{2.11}$$

*for some $C_2 = C_2(c)$ provided that the number $k$ of gross errors obeys $\delta_{3k} + \frac{1}{2}\delta_{2k} < \frac{c}{2}(\frac{m}{n} - 1)$ for some $c < 1$; $x^{\mathrm{Ideal}}$ is the ideal solution (2.3) one would get if no gross errors ever occurred ($e = 0$).*

*If the (orthonormal) columns of $A$ are selected uniformly at random, then with probability at least $1 - O(e^{-\gamma(m-n)})$ for some positive constant $\gamma$, the estimate (2.11) holds for $k \asymp \rho \cdot m$, provided $\rho \le \rho^*(n/m)$, which is a constant depending only $n/m$.* [1]

This theorem is of significant appeal because it says that the reconstruction error is in some sense within a constant factor of the ideal solution. Indeed, suppose all we know about $z$ is that $\|z\|_{\ell_2} \le \varepsilon$. Then $\|x^{\mathrm{Ideal}} - x\|_{\ell_2} = \|A^* z\|_{\ell_2}$ may be as large as $\varepsilon$. Thus for $m = 2n$, say, (2.11) asserts that *the reconstruction error is bounded by a constant times the ideal reconstruction error.* In addition, if one selects a coding matrix with random orthonormal columns (one way of doing so is to sample $X \in \mathbb{R}^{m\times n}$ with i.i.d. $N(0,1)$ entries and orthonormalize the columns by means of the QR factorization), then one can correct a positive fraction of arbitrarily corrupted entries, in a near ideal fashion.

Note that in the case where there are no small errors ($z = 0$), the decoding is exact since $\varepsilon = 0$ and $x^{\mathrm{Ideal}} = x$. Hence, this generalizes earlier results [28]. We

[1] Analysis shows $\rho^*$ to be of the form $\rho^* = O\left(\frac{n/m-1}{\log(1-n/m)}\right)$ but this is not informative because the constant is unknown. Determining the constant is extremely challenging; for an analysis with sparse errors see [39, 49].

would like to emphasize that there is nothing special about the fact that the columns of $A$ are taken to be orthonormal in Theorem 2.3.2. In fact, one could just as well obtain equivalent statements for general matrices. Our assumption only allows us to formulate simple and useful results.

While the previous result discussed arbitrary small errors, the next is about stochastic errors.

**Corollary 2.3.3.** *Suppose the small errors are i.i.d.* $N(0, \sigma^2)$ *and set* $\varepsilon := \sqrt{(m-n)(1+t)} \cdot \sigma$ *for some fixed* $t > 0$. *Then under the same hypotheses about the restricted isometry constants of* $A$ *and the number of gross errors as in Theorem 2.3.2, the solution to* $(P_2)$ *obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \leq C_2' \cdot m \cdot \sigma^2, \tag{2.12}$$

*for some numerical constant* $C_2'$ *with probability exceeding* $1 - e^{-\gamma^2(m-n)/2} - e^{-m/2}$ *where* $\gamma = \frac{\sqrt{1+2t}-1}{\sqrt{2}}$. *In particular, this last statement holds with overwhelming probability if* $A$ *is chosen at random as in Theorem 2.3.2.*

Suppose for instance that $m = 2n$ to make things concrete so that the MSE of the ideal estimate is equal to $m/2 \cdot \sigma^2$. Then the SOCP reconstruction is within a multiplicative factor $2C$ of the ideal MSE. Our experiments show that in practice the constant is small: e.g., when $m = 2n$, one can correct 15% of arbitrary errors, and in the overwhelming majority of cases obtain a decoded vector whose MSE is less than 3 times larger than the ideal MSE.

## 2.4 Decoding by linear programming

Another way to recover the signal $x$ from the corrupted vector $y$ (2.2) is by linear programming:

$$(P_\infty) \quad \min \|y - A\tilde{x} - \tilde{z}\|_{\ell_1} \text{ subject to } \|\tilde{z}\|_{\ell_\infty} \leq \lambda, \quad A^*\tilde{z} = 0, \tag{2.13}$$

with variables $\tilde{x} \in \mathbb{R}^n$ and $\tilde{z} \in \mathbb{R}^m$. As is well known, the program $(P_\infty)$ may also be re-expressed as a linear program by introducing slack variables just as in $(P_2)$; we omit the standard details. As with $(P_2)$, the parameter $\lambda$ here is related to the size of the small errors and will be discussed shortly. In the sequel, we shall also be interested in the more general formulation of $(P_\infty)$

$$\begin{aligned} \|y - A\tilde{x} - z\|_{\ell_1} \text{ subject to } |\tilde{z}|_i \le \lambda_i, \;\; 1 \le i \le m, \\ A^*\tilde{z} \qquad\qquad\qquad\qquad\qquad = 0, \end{aligned} \tag{2.14}$$

which gives additional flexibility for adjusting the thresholds $\lambda_1, \lambda_2, \ldots, \lambda_m$ to the noise level.

The same arguments as before prove that $(P_\infty)$ is equivalent to

$$(P'_\infty) \quad \min \|\tilde{e}\|_{\ell_1} \text{ subject to } \|QQ^*(y - \tilde{e})\|_{\ell_\infty} \le \lambda, \tag{2.15}$$

where we recall that $P_{V^\perp} = QQ^*$ is the orthonormal projector onto $V^\perp$ ($V$ is the column space of $A$); that is, if $\hat{e}$ is solution to $(P'_\infty)$, then there is a unique decomposition $y - \hat{e} = A\hat{x} + \hat{z}$ where $A^*\hat{z} = 0$ and $(\hat{x}, \hat{z})$ is solution to $(P_\infty)$. The converse is also true. Similarly, the more general program (2.14) is equivalent to minimizing the $\ell_1$ norm of $\tilde{e}$ under the constraint $|P_{V^\perp}(y - \tilde{e})|_i \le \lambda_i$, $1 \le i \le m$.

In statistics, the estimator $\hat{e}$ solution to $(P'_\infty)$ is known as the *Dantzig selector* [30]. It was originally introduced to estimate the vector $e$ from the data $y'$ and the model

$$y' = Q^*e + z' \tag{2.16}$$

where $z'$ is a vector of stochastic errors, e.g., independent mean-zero Gaussian random variables. The connection with our problem is clear since applying the parity-check matrix $Q^*$ on both sides of (2.2) gives

$$Q^*y = Q^*e + Q^*z$$

as before. If $z$ is stochastic noise, we can use the Dantzig selector to recover $e$ from $Q^*y$. Moreover, available statistical theory asserts that if $Q^*$ obeys nice restricted

isometry properties and $e$ is sufficiently sparse just as before, then this estimation procedure is extremely accurate and in some sense optimal.

It remains to discuss how one should specify the parameter $\lambda$ in (2.13)–(2.15) which is easy. Suppose the small errors are stochastic. Then we fix $\lambda$ so that the true vector $e$ is feasible for $(P'_\infty)$ with very high probability; i.e., we adjust $\lambda$ so that

$$\|P_{V^\perp}(y-e)\|_{\ell_\infty} = \|P_{V^\perp} z\|_{\ell_\infty} \le \lambda$$

with high probability. In the more general formulation, the thresholds are adjusted so that $\sup_{1\le i\le m} |P_{V^\perp} z|_i/\lambda_i \le 1$ with high probability.

The main result of this section is that the LP decoder is also provably accurate.

**Theorem 2.4.1.** *Choose a coding matrix $A \in \mathbb{R}^{m\times n}$ with orthonormal columns spanning $V$, and let $(\delta_k)$ be the isometry constants of the rescaled matrix $\sqrt{\frac{m}{n}}\, A^*$. Suppose $\|P_{V^\perp} z\|_{\ell_\infty} \le \lambda$. Then the solution $\hat{x}$ to $(P_\infty)$ obeys*

$$\|\hat{x} - x\|_{\ell_2} \le C_1 \sqrt{k} \cdot \frac{\lambda}{1-\frac{n}{m}} + \|x^{\mathrm{Ideal}} - x\|_{\ell_2} \tag{2.17}$$

*for some $C_1 = C_1(c)$ provided that the number $k$ of gross errors obeys $\delta_{3k} + \delta_{2k} < c\left(\frac{m}{n} - 1\right)$ for some $c < 1$; $x^{\mathrm{Ideal}}$ is the ideal solution (2.3) one would get if no gross errors ever occurred.*

*If the (orthonormal) columns of $A$ are selected uniformly at random, then with probability at least $1 - O(e^{-\gamma(m-n)})$ for some positive constant $\gamma$, the estimate (2.17) holds for $k \asymp \rho \cdot m$, provided $\rho \le \rho^*(n/m)$.*

In effect, the LP decoder efficiently corrects a positive fraction of arbitrarily corrupted entries. Again, when there are no small errors ($z = 0$), the decoding is exact. (Also and just as before, there is nothing special about the fact that the columns of $A$ are taken to be orthonormal.) We now consider the interesting case in which the small errors are stochastic. Below, we conveniently adjust the thresholds $\lambda_j$ so that the true vector $e$ is feasible with high probability, see Section 2.6.4 for details.

**Corollary 2.4.2.** *Choose a coding matrix $A$ with (orthonormal) columns selected*

*uniformly at random and suppose the small errors are i.i.d.* $N(0, \sigma^2)$. *Fix*

$$\lambda_i = \sqrt{2 \log m} \cdot \sqrt{1 - \|A_{i,\cdot}\|_{\ell_2}^2} \cdot \sigma$$

*in* (2.14), *where* $\|A_{i,\cdot}\|_{\ell_2} = (\sum_{1 \le j \le n} A_{i,j}^2)^{1/2}$ *is the* $\ell_2$ *norm of the* $i$*th row. Then if the number* $k$ *of gross errors is no more than a fraction of* $m$ *as in Theorem 2.4.1, the solution* $\hat{x}$ *obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \le [1 + C_1' s]^2 \cdot \|x^{\text{Ideal}} - x\|_{\ell_2}^2, \tag{2.18}$$

*with very large probability, where* $C_1'$ *is some numerical constant and*

$$s^2 = \frac{k}{m} \cdot \frac{\log m}{\frac{n}{m}(1 - \frac{n}{m})}.$$

*In effect,* $\|\hat{x} - x\|_{\ell_2}^2$ *is bounded by just about* $[1 + C_1' s]^2 \cdot n\sigma^2$ *since* $\|x^{\text{Ideal}} - x\|_{\ell_2}^2$ *is distributed as* $\sigma^2$ *times a chi-square with* $n$ *degrees of freedom, and is tightly concentrated around* $n\sigma^2$.

Recall that the MSE is equal to $n\sigma^2$ when there are no gross errors and, therefore, this last result asserts that *the reconstruction error is bounded by a constant times the ideal reconstruction error.* Suppose for instance that $m = 2n$. Then $s^2 = 4k(\log m)/m$ and we see that $s$ is small when there are few gross errors. In this case, the recovery error is very close to that attained by the ideal procedure. Our experiments show that in practice, the constant $C_1'$ is quite small: for instance, when $m = 2n$, one can correct 15% of arbitrary errors, and in the overwhelming majority of cases obtain a decoded vector whose MSE is less than 3 times larger than the ideal MSE.

Finally, this last result is in some way more subtle than the corresponding result for the SOCP decoder. Indeed, note the explicit dependence on $k$ of the scaling factor in (2.18) that is not present in the corresponding expression for the SOCP decoder (2.12). This says that in some sense the accuracy of the LP decoder *automatically adapts to the number* $k$ *of gross errors* which were introduced. The smaller this number, the smaller the recovery error. For small values of $k$, the bound in (2.18) may in fact be considerably smaller than its analog (2.12).

## 2.5 Numerical experiments

As mentioned earlier, numerical studies show that the empirical performance of the proposed decoding strategies is noticeable. To confirm these findings, this section discusses an experimental setup and presents numerical results. The reader wanting to reproduce our results may find the matlab file available at `http://www.acm.caltech.edu/~emmanuel/ConvexDecode.m` useful. Here are the steps we used:

1. Choose a pair $(n, m)$ and sample an $m$ by $n$ matrix $A$ with independent standard normal entries; the coding matrix is fixed throughout.

2. Choose a fraction $\rho$ of grossly corrupted entries and define the number of corrupted entries as $k = \texttt{round}(\rho \cdot m)$; e.g., if $m = 512$ and 10% of the entries are corrupted, $k = 51$.

3. Sample a block of information $x \in \mathbb{R}^n$ with independent and identically distributed Gaussian entries. Compute $Ax$.

4. Select $k$ locations uniformly at random and flip the signs of $Ax$ at these locations.

5. Sample the vector $z = (z_1, \ldots, z_m)$ of smaller errors with $z_i$ i.i.d. $N(0, \sigma^2)$, and add $z$ to the outcome of the previous step. Obtain $y$.

6. Obtain $\hat{x}$ by solving both $(P_2)$ and $(P_\infty)$ followed by a reprojection step discussed below [30].

7. Repeat steps (3)–(6) 500 times.

We briefly discuss the reprojection step. As observed in [30], both programs $(P_2')$ and $(P_\infty')$ have a tendency to underestimate the vector $e$ (they tend to be akin to soft-thresholding procedures). One can easily correct for this bias as follows:

1. Solve $(P_2')$ or $(P_\infty')$ and obtain $\hat{e}$.

2. Estimate the support of the gross errors $e$ via $I := \{i : |\hat{e}_i| > \sigma\}$, where $\sigma$ is the standard deviation of the smaller errors; recall that $y' := Q^*y = Q^*e + Q^*z$

and update the estimate by regressing $y'$ onto the selected columns of $Q^*$ via the method of least squares:

$$\hat{e} = \text{argmin } \|y' - Q^*\tilde{e}\|_{\ell_2}^2 \quad \text{subject to} \quad \tilde{e}_i = 0,\ i \in I^c.$$

3. Finally, obtain $\hat{x}$ via $(A^*A)^{-1}A^*(y - \hat{e})$ where $\hat{e}$ is the reprojected estimate calculated in the previous step.

In our series of experiments, we used $m = 2n = 512$ and a corruption rate of 10%. The standard deviation $\sigma$ is selected in such a way that just about the first three binary digits of each entry of the codeword $Ax$ are reliable. Formally $\sigma = \text{median}|Ax|/16$. Finally and to be complete, we set the threshold $\varepsilon$ in $(P_2)$ so that $\|Q^*z\|_{\ell_2} \le \varepsilon$ with probability .95; in other words, $\varepsilon^2 = \chi^2_{m-n}(.95) \cdot \sigma^2$, where $\chi^2_{m-n}(.95)$ is the 95th percentile of a chi-squared distribution with $m - n$ degrees of freedom. We also set the thresholds in the general formulation (2.14) of $(P_\infty)$ in a similar fashion. The distribution of $(QQ^*z)_i$ is normal with mean 0 and variance $s_i^2 = (QQ^*)_{i,i} \cdot \sigma^2$ so that the variable $z_i' = (QQ^*z)_i/s_i$ is standard normal. We choose $\lambda_i = \lambda \cdot s_i$ where $\lambda$ obeys

$$\sup_{1 \le i \le m} |z_i'| \le \lambda$$

with probability at least .95. In both cases, our selection makes the true vector $e$ of gross errors feasible with probability at least .95. In our simulations, the thresholds for the SOCP and LP decoders (the parameters $\chi^2_{m-n}(.95)$ and $\lambda$) were computed by Monte Carlo simulations.

To evaluate the accuracy of the decoders, we report two statistics

$$\rho^{\text{Ideal}} = \frac{\|\hat{x} - x\|}{\|x^{\text{Ideal}} - x\|}, \quad \text{and} \quad \rho^{\text{Oracle}} = \frac{\|\hat{x} - x\|}{\|x^{\text{Oracle}} - x\|}, \tag{2.19}$$

which compare the performance of our decoders with that of ideal strategies which assume either exact knowledge of the gross errors or exact knowledge of their locations. As discussed earlier, $x^{\text{Ideal}}$ is the reconstructed vector one would obtain if the gross errors were known to the receiver *exactly* (which is of course equivalent to

having no gross errors at all). The reconstruction $x^{\text{Oracle}}$ is that one would obtain if, instead, one had available an oracle supplying perfect information about the location of the gross errors (but not their value). Then one could simply delete the corrupted entries of the received codeword $y$ and reconstruct $x$ by the method of least squares, i.e., find the solution to $\|y^{\text{Oracle}} - A^{\text{Oracle}}\tilde{x}\|_{\ell_2}$, where $A^{\text{Oracle}}$ (resp. $y^{\text{Oracle}}$) is obtained from $A$ (resp. $y$) by deleting the corrupted rows.

The results are presented in Figure 2.1 and summarized in Table 2.1. These results show that both our approaches work extremely well. As one can see, our methods give reconstruction errors which are nearly as sharp as if no gross errors had occurred or as if one knew the locations of these large errors exactly. Put in a different way, the constants appearing in our quantitative bounds are in practice very small. Finally, the SOCP and LP decoders have about the same performance although upon closer inspection, one could argue that the LP decoder is perhaps a tiny bit more accurate.

| | median of $\rho^{\text{Ideal}}$ | mean of $\rho^{\text{Ideal}}$ | median of $\rho^{\text{Oracle}}$ | mean of $\rho^{\text{Oracle}}$ |
|---|---|---|---|---|
| SOCP decoder | 1.386 | 1.401 | 1.241 | 1.253 |
| LP decoder | 1.346 | 1.386 | 1.212 | 1.239 |

**Table 2.1.** Summary statistics of the ratios $\rho^{\text{Ideal}}$ and $\rho^{\text{Oracle}}$ (2.19) for the Gaussian coding matrix.

We also repeated the same experiment but with a coding matrix $A$ consisting of $n = 256$ randomly sampled columns of the $512 \times 512$ discrete Fourier transform, and obtained very similar results. The results are presented in Figure 2.2 and summarized in Table 2.2. The numbers are remarkably close to our earlier findings and again both our methods work extremely well (again the LP decoder is a tiny bit more accurate). This experiment is of special interest since it suggests that one can apply our decoding algorithms to very large data vectors, e.g., with sizes ranging in the hundred of thousands. The reason is that one can use off-the-shelf interior point algorithms which only need to be able to apply $A$ or $A^*$ to arbitrary vectors (and never need to manipulate the entries of $A$ or even store them). When $A$ is a partial Fourier transform, one can evaluate $Ax$ and $A^*y$ by means of the FFT and, hence, this is well suited for very large problems. See [14] for very large scale experiments

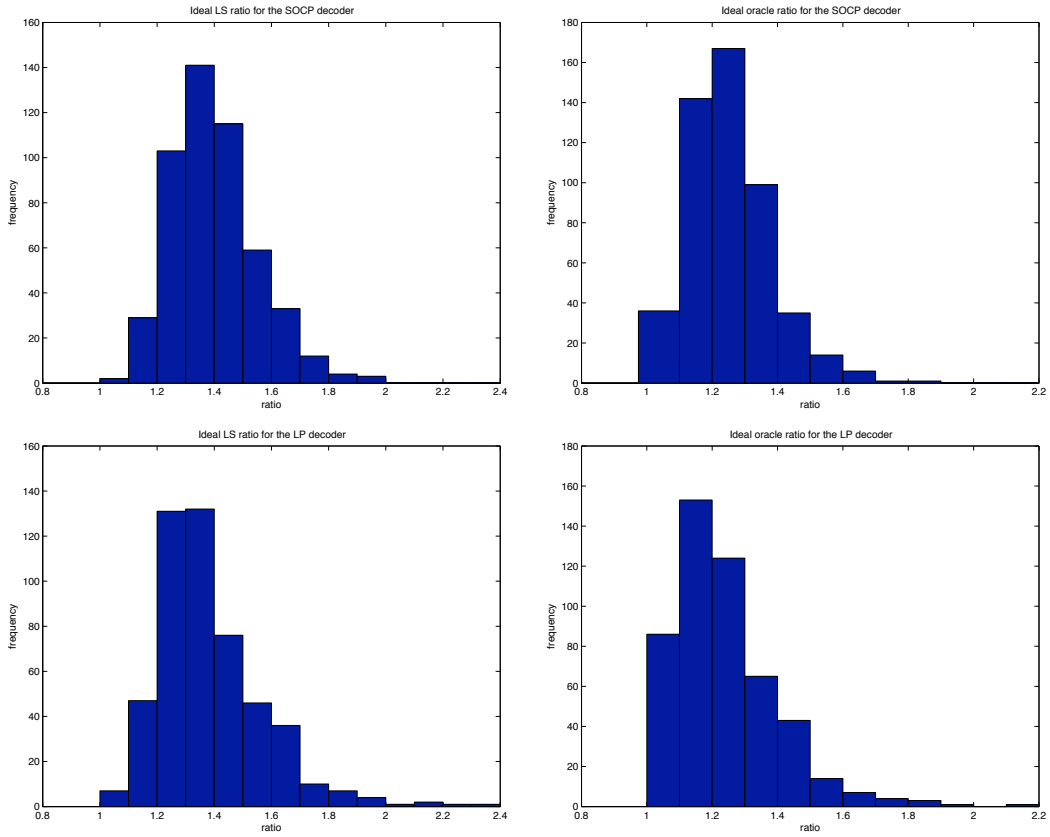



**Figure 2.1.** Statistics of the ratios (2.19) $\rho^{\text{Ideal}}$ (first column) and $\rho^{\text{Oracle}}$ (second column) which compare the performance of the proposed decoders with that of ideal strategies which assume either exact knowledge of the gross errors or exact knowledge of their locations. The first row shows the performance of the SOCP decoder, the second that of the LP decoder.

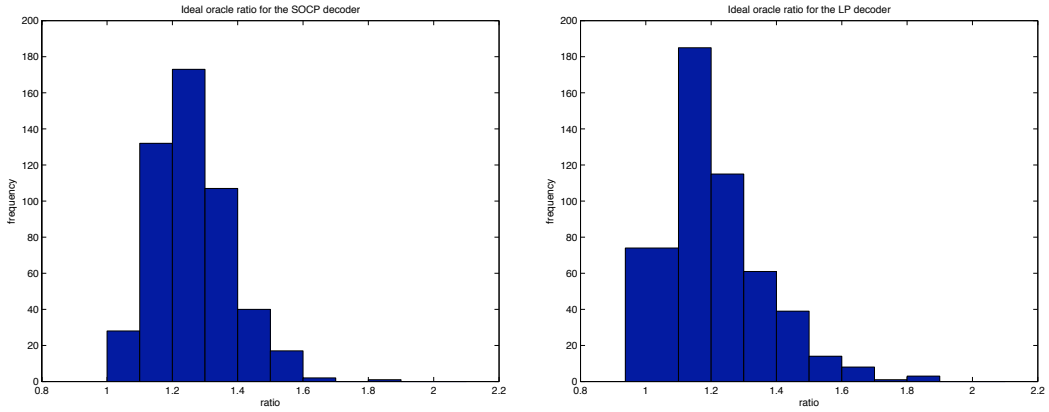



**Figure 2.2.** Statistics of the ratios $\rho^{\text{Oracle}}$ for the SOCP decoder (first column) and the LP decoder (second column) in the case where the coding matrix is a partial Fourier transform.

of a similar flavor.

| | median of $\rho^{\text{Ideal}}$ | mean of $\rho^{\text{Ideal}}$ | median of $\rho^{\text{Oracle}}$ | mean of $\rho^{\text{Oracle}}$ |
|---|---|---|---|---|
| SOCP decoder | 1.390 | 1.401 | 1.244 | 1.262 |
| LP decoder | 1.337 | 1.375 | 1.195 | 1.230 |

**Table 2.2.** Summary statistics of the ratios $\rho^{\text{Ideal}}$ and $\rho^{\text{Oracle}}$ (2.19) for the Fourier coding matrix.

## 2.6 Proofs

In this section, we prove all of our results. We begin with some preliminaries which will be used throughout, then prove the claims about the SOCP decoder, and end this section with the LP decoder. Our work builds on [27] and [30].

### 2.6.1 Preliminaries

We shall make extensive use of two simple lemmas that we now record.

**Lemma 2.6.1.** *Let $Y_d \sim \chi^2_d$ be distributed as a chi-squared random variable with $d$*

*degrees of freedom. Then for each* $t > 0$

$$\begin{aligned} &\mathbf{P}(Y_d - d \geq t\sqrt{2d} + t^2) \leq e^{-t^2/2} \quad \textit{and} \\ &\mathbf{P}(Y_d - d \leq -t\sqrt{2d}) \leq e^{-t^2/2}. \end{aligned} \tag{2.20}$$

This is fairly standard [62], see also [60] for very slightly refined estimates. We will use (2.20) as follows: for each $\epsilon \in (0,1)$ we have

$$\begin{aligned} &\mathbf{P}(Y_d \geq d\,(1-\epsilon)^{-1}) \leq e^{-\epsilon^2 d/4} \quad \text{and} \\ &\mathbf{P}(Y_d \leq d\,(1-\epsilon)) \leq e^{-\epsilon^2 d/4}. \end{aligned} \tag{2.21}$$

A consequence of these large deviation bounds is the estimate below.

**Lemma 2.6.2.** *Let* $(u_1, u_2, \ldots, u_m)$ *be a vector uniformly distributed on the unit sphere in* $m$ *dimensions and* $Z_n = u_1^2 + \ldots + u_n^2$ *be the squared length of its first* $n$ *components. Then for each* $t < 1$

$$\mathbf{P}\left(Z_n \leq \frac{n}{m}\,(1-t)\right) \leq e^{-nt^2/16} + e^{-mt^2/16}, \tag{2.22}$$

*and*

$$\mathbf{P}\left(Z_n \geq \frac{n}{m}\,\frac{1}{1-t}\right) \leq e^{-nt^2/16} + e^{-mt^2/16}. \tag{2.23}$$

*Proof.* A result of this kind would essentially follow from the measure concentration on the sphere [7], but we prefer giving a short and elementary argument. Suppose $X_1, X_2, \ldots, X_m$ are i.i.d. $N(0,1)$. Then the distribution of $(u_1, u_2, \ldots, u_m)$ is that of the vector $X/\|X\|_{\ell_2}$ and, therefore, the law of $Z_n$ is that of $Y_n/Y_m$, where $Y_k = \sum_{j\leq k} X_j^2$. For a fixed $t \in (0,1)$, define the events $A = \{Y_m \geq m/(1-t/2)\}$ and $B = \{Y_n/Y_m \leq n/m(1-t)\}$. We have

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(B \,|\, A^c)\,\mathbf{P}(A^c) + \mathbf{P}(B \,|\, A)\,\mathbf{P}(A) \\ &\leq \mathbf{P}(Y_n \leq n(1-t)/(1-t/2)) + \mathbf{P}(Y_m \geq m/(1-t/2)). \end{aligned}$$

For $0 \le t \le 1$, we have $(1-t)/(1-t/2) \le 1 - t/2$ and thus

$$\begin{aligned}
&\mathbf{P}(Z_n \le n/m(1-t)) \\
&\quad \le \mathbf{P}(Y_n \le n(1-t/2)) + \mathbf{P}(Y_m \ge m/(1-t/2)) \\
&\quad \le e^{-nt^2/16} + e^{-mt^2/16},
\end{aligned}$$

which follows from (2.21).

For the second inequality, we employ a similar strategy with $A = \{Y_m \le m(1 - t/2)\}$ and $B = \{Y_n/Y_m \ge n/m\,(1-t)^{-1}\}$, which leads to

$$\begin{aligned}
&\mathbf{P}(Z_n \ge n/m\,(1-t)^{-1}) \\
&\quad \le \mathbf{P}(Y_n \ge n/(1-t/2)) + \mathbf{P}(Y_m \le m(1-t/2)) \\
&\quad \le e^{-nt^2/16} + e^{-mt^2/16},
\end{aligned}$$

as claimed. □

### 2.6.2 Restricted isometries

For a matrix $\Phi$, define the sequences $(a_k)$ and $(b_k)$ as respectively the largest and smallest numbers obeying

$$a_k \|x\|_{\ell_2} \le \|\Phi x\|_{\ell_2} \le b_k \|x\|_{\ell_2}, \tag{2.24}$$

for all $k$-sparse vectors. In other words, if we list all the singular values of all the submatrices of $\Phi$ with $k$ columns, $a_k$ is the smallest element from that list and $b_k$ the largest. Note of course the resemblance with (2.10)—only this is slightly more general.

Restricted extremal singular values of random orthonormal projections will play an important role in the sequel. The following lemma states that for an $r \times m$ random orthogonal projection, the numbers $a_k$ and $b_k$ are about $\sqrt{r/m}$.

**Lemma 2.6.3.** *Let $\Phi$ be the first $r$ rows of a random orthogonal matrix (sampled*

*from the Haar measure). Then the restricted extremal singular values of* $\Phi$ *obey*

$$\mathbf{P}\left(a_k(\Phi) \leq \frac{7}{8}\sqrt{\frac{r}{m}}\right) \leq c_0\, e^{-\gamma_0 r} \tag{2.25}$$

*and*

$$\mathbf{P}\left(b_k(\Phi) \geq \frac{9}{8}\sqrt{\frac{r}{m}}\right) \leq c_0'\, e^{-\gamma_0' r} \tag{2.26}$$

*for some universal positive constants* $c_0, c_0', \gamma_0, \gamma_0'$ *provided that* $k \leq c_1 r/\log(m/r)$ *for some* $c_1 > 0$.

*Proof.* Put $\Sigma_k$ for the set of all unit-normed $k$-sparse vectors. By definition

$$a_k(\Phi) = \inf_{x \in \Sigma_k} \|\Phi x\|_{\ell_2}.$$

Take a fixed vector $x$ in $\Sigma_k$. Then $\|\Phi x\|_{\ell_2}^2$ is distributed as $Z_r$ in Lemma 2.6.2. To see why this is true, note that $\Phi x$ are the first $r$ components of $Ux$ where $U$ is an $m \times m$ random orthogonal matrix. The claim follows from the fact that $Ux$ is uniformly distributed on the $(m-1)$-dimensional unit sphere. This is useful because Lemma 2.6.2 can be employed to show that for a fixed $x \in \Sigma_k$, $\|\Phi x\|_{\ell_2}$ can not deviate much from $\sqrt{r/m}$. To develop an inequality concerning all sparse vectors, we now employ a covering number argument.

Consider an $\epsilon$-net $\mathcal{N}(\epsilon)$ of $\Sigma_k$. An $\epsilon$-net is a subset $\mathcal{N}(\epsilon)$ of $\Sigma_k$ such that for all $x \in \Sigma_k$, there is an $x_0 \in \mathcal{N}(\epsilon)$ such that $\|x - x_0\|_{\ell_2} \leq \epsilon$. In other words, $\mathcal{N}(\epsilon)$ approximates $\Sigma_k$ to within distance $\epsilon$. For each $x \in \Sigma_k$,

$$\|\Phi x\| \geq \|\Phi x_0\|_{\ell_2} - \|\Phi(x - x_0)\|_{\ell_2} \geq \|\Phi x_0\| - \epsilon,$$

for some $x_0 \in \mathcal{N}(\epsilon)$ obeying $\|x - x_0\|_{\ell_2} \leq \epsilon$, where the last inequality follows from the fact that the operator norm of $\Phi$ is bounded by 1. Hence,

$$a_k(\Phi) \geq \inf_{x_0 \in \mathcal{N}(\epsilon)} \|\Phi x_0\| - \epsilon.$$

Now set $\epsilon = 1/16 \cdot \sqrt{r/m}$. Then

$$\begin{aligned}
&\mathbf{P}\left(a_k(\Phi) < \frac{7}{8}\sqrt{\frac{r}{m}}\right) \\
&\quad \le \mathbf{P}\left(\inf_{x_0 \in \mathcal{N}(\epsilon)} \|\Phi x_0\|_{\ell_2} \le \sqrt{\frac{r}{m}}\left(1 - \frac{1}{16}\right)\right) \\
&\quad \le |\mathcal{N}(\epsilon)| \cdot \mathbf{P}\left(Z_r < \frac{r}{m}\left(1 - \frac{1}{16}\right)^2\right),
\end{aligned}$$

which comes from the union bound together with $\|\Phi x_0\|_{\ell_2}^2 \sim Z_r$ for each $x_0$. Further, one can find $\mathcal{N}(\epsilon)$ obeying

$$|\mathcal{N}(\epsilon)| \le (3/\epsilon)^k \binom{m}{k}.$$

The reason is simple. First, one can find an $\epsilon$-net of the $k-1$-dimensional sphere whose cardinality does not exceed $(3/\epsilon)^k$, see [74, Lemma 4.16]. And second, $\Sigma_k$ is a union of $\binom{m}{k}$ $k-1$-dimensional spheres. We then apply this fact together with Lemma 2.6.2, and obtain

$$\mathbf{P}\left(a_k(\Phi) < \frac{7}{8}\sqrt{\frac{r}{m}}\right) \le 48^k \, (m/r)^{k/2} \binom{m}{k} e^{-r/133}.$$

Next, there is a bound on binomial coefficients of the form $\log \binom{m}{k} \le k(\log(m/k)+1)$ so that

$$48^k \, (m/r)^{k/2} \binom{m}{k} \le \exp\left(k(5 + 0.5\log(m/r) + \log(m/k))\right).$$

One can check that if $k \le c_0 \cdot r/(\log m/r)$ for $c_0$ sufficiently small, the right-hand-side of the last inequality is bounded by $e^{\beta_0 r}$ for some $\beta_0 < 1/133$. This establishes the first part of the theorem, namely, (2.25).

The second part is nearly identical and is only sketched. We have that

$$b_k(\Phi) = \sup_{x \in \Sigma_k} \|\Phi x\|_{\ell_2} \le \sup_{x_0 \in \mathcal{N}(\epsilon)} \|\Phi x_0\|_{\ell_2} + \epsilon.$$

The proof now proceeds as before noting that (2.23) gives a bound on the probability that for each $x_0$, $\|\Phi x_0\|_{\ell_2}^2$ exceeds $r/m$ times a small multiplicative factor. □

Note that in this proof, we have not tried to derive the optimal constants, and

a more refined analysis would surely yield far better numerical constants.

### 2.6.3 The SOCP decoder

We begin by adapting an important result from [27].

**Lemma 2.6.4 (adapted from [27]).** *Set $\Phi \in \mathbb{R}^{r\times m}$ and let $(a_k)$ and $(b_k)$ be the restricted extremal singular values of $\Phi$ as in (2.24). Any point $\tilde{x} \in \mathbb{R}^m$ obeying*

$$\|\tilde{x}\|_{\ell_1} \le \|x\|_{\ell_1}, \quad \text{and} \quad \|\Phi\tilde{x} - \Phi x\|_{\ell_2} \le 2\varepsilon, \tag{2.27}$$

*also obeys*

$$\|\tilde{x} - x\|_{\ell_2} \le \frac{\sqrt{6}\,\varepsilon}{a_{3k}(\Phi) - \frac{1}{\sqrt{2}}\,b_{2k}(\Phi)}, \tag{2.28}$$

*provided that $x$ is $k$-sparse with $k$ such that $a_{3k}(\Phi) - \frac{1}{\sqrt{2}}\,b_{2k}(\Phi) > 0$. holds*

The proof follows the same steps as that of Theorem 1.1 in [27], and is omitted. In particular, it follows from (2.6) in the aforementioned reference with $M = 2|T_0|$ and $a_{M+|T_0|}$ (resp. $b_M$) in place of $\sqrt{1-\delta_{|T_0|+M}}$ (resp. $\sqrt{1+\delta_M}$) in the definition of $C_{|T_0|,M}$.

#### 2.6.3.1 Proof of Theorem 2.3.2

Recall that the solution $(\hat{x}, \hat{z})$ to $(P_2)$ obeys (2.8) where $\hat{e}$ is the solution to $(P_2')$. Replacing $y$ in (2.8) with $Ax + e + z$ gives

$$\begin{aligned}\hat{x} - x &= (A^*A)^{-1}A^*(e - \hat{e}) + (A^*A)^{-1}A^*z \\ &= (A^*A)^{-1}A^*(e - \hat{e}) + x^{\text{Ideal}} - x,\end{aligned} \tag{2.29}$$

and since $A^*A = I$,

$$\|\hat{x} - x\|_{\ell_2} \le \|A^*(e - \hat{e})\|_{\ell_2} + \|x^{\text{Ideal}} - x\|_{\ell_2}.$$

To prove (2.11), it then suffices to show that $\|e - \hat{e}\|_{\ell_2} \le \frac{C\,\varepsilon}{\sqrt{1-\frac{n}{m}}}$ since the 2-norm of $A^*$ is at most 1.

By assumption $\|Q^*(y-e)\|_{\ell_2} = \|Q^*z\|_{\ell_2} \leq \varepsilon$ and thus, $e$ is feasible for $(P_2')$ which implies $\|\hat{e}\|_{\ell_1} \leq \|e\|_{\ell_1}$. Moreover,

$$\|Q^*e - Q^*\hat{e}\|_{\ell_2} \leq \|Q^*(y-e)\|_{\ell_2} + \|Q^*(y-\hat{e})\|_{\ell_2} \leq 2\varepsilon.$$

We then apply Lemma 2.6.4 (with $\Phi = Q^*$) and obtain

$$\|e - \hat{e}\|_{\ell_2} \leq \frac{\sqrt{6}\,\varepsilon}{a_{3k}(Q^*) - \frac{1}{\sqrt{2}}\, b_{2k}(Q^*)}. \tag{2.30}$$

Now since the $m \times m$ matrix obtained by concatenating the columns of $A$ and $Q$ is an isometry, we have

$$\|A^*x\|_{\ell_2}^2 + \|Q^*x\|_{\ell_2}^2 = \|x\|_{\ell_2}^2 \quad \forall x \in \mathbb{R}^m,$$

whence

$$\begin{aligned} a_k^2(Q^*) &= 1 - b_k^2(A^*), \\ b_k^2(Q^*) &= 1 - a_k^2(A^*). \end{aligned}$$

Assuming that $a_{3k}(Q^*) \geq \frac{1}{\sqrt{2}} b_{2k}(Q^*)$, we deduce from (2.30) that

$$\begin{aligned} \|e - \hat{e}\|_{\ell_2} &\leq \sqrt{6}\,\varepsilon \cdot \frac{a_{3k}(Q^*) + \frac{1}{\sqrt{2}} b_{2k}(Q^*)}{1 - b_{3k}^2(A^*) - \frac{1}{2}(1 - a_{2k}^2(A^*))} \\ &\leq 2\sqrt{6}\varepsilon \cdot \frac{a_{3k}(Q^*)}{\frac{1}{2} + \frac{1}{2}a_{2k}^2(A^*) - b_{3k}^2(A^*)}. \end{aligned} \tag{2.31}$$

Recall that $(\delta_k)$ are the restricted isometry constants of $\sqrt{\frac{m}{n}}\, A^*$, and observe that by definition for each $k = 1, 2, \ldots,$

$$a_k^2(A^*) \geq \frac{n}{m}(1 - \delta_k), \qquad b_k^2(A^*) \leq \frac{n}{m}(1 + \delta_k).$$

It follows that the denominator on the right-hand side of (2.31) is greater or equal

to

$$\frac{1}{2} + \frac{n}{2m}(1-\delta_{2k}) - \frac{n}{m}(1+\delta_{3k}) \\ = \frac{1}{2}\left(1-\frac{n}{m}\right) - \frac{n}{m}\left(\delta_{3k} + \frac{1}{2}\delta_{2k}\right).$$

Now suppose that for some $0 < c < 1$,

$$\delta_{3k} + \frac{1}{2}\delta_{2k} \le \frac{c}{2} \cdot \left(\frac{m}{n} - 1\right).$$

This automatically implies $a_{3k}(Q^*) \ge \frac{1}{\sqrt{2}} b_{2k}(Q^*)$, and the denominator on the right-hand side of (2.31) is greater or equal to $\frac{1}{2}(1-c)(1-\frac{n}{m})$. The numerator obeys

$$a_{3k}^2(Q^*) = 1 - b_{3k}^2(A^*) \le 1 - a_{3k}^2(A^*) \le 1 - (1-\delta_{3k})\frac{n}{m}.$$

Since $\frac{n}{m}\,\delta_{3k} \le \frac{c}{2}(1-\frac{n}{m})$, we also have $a_{3k}^2(Q^*) \le (1+\frac{c}{2})(1-\frac{n}{m})$. In summary, (2.31) gives

$$\|e - \hat{e}\|_{\ell_2} \le C_2 \cdot \frac{\varepsilon}{\sqrt{1-\frac{n}{m}}},$$

where one can take $C_2$ as $4\sqrt{6(1+c/2)}/(1-c)$. This establishes the first part of the claim.

We now turn to the second part of the theorem and argue that if the orthonormal columns of $A$ are chosen uniformly at random, the error bound (2.11) is valid as long as we have a constant fraction of gross errors. Put $r = m - n$ and let $X$ be an $m$ by $r$ matrix with independent Gaussian entries with mean 0 and variance $1/m$. Consider now the reduced singular value decomposition of $X$

$$X = U\Sigma V^*, \quad U \in \mathbb{R}^{m\times r} \text{ and } \Sigma, V \in \mathbb{R}^{r\times r}.$$

Then the columns of $U$ are $r$ orthonormal vectors selected uniformly at random and thus $U$ and $Q$ have the same distribution. Thus we can think of $Q$ as being the left singular vectors of a Gaussian matrix $X$ with independent entries. From now on,

we identify $U$ with $Q$. Observe now that

$$\begin{aligned}\|X^*(\hat{e}-e)\|_{\ell_2} = \|V\Sigma Q^*(\hat{e}-e)\|_{\ell_2} &= \|\Sigma Q^*(\hat{e}-e)\|_{\ell_2} \\ &\leq \sigma_1(X)\,\|Q^*(\hat{e}-e)\|_{\ell_2},\end{aligned}$$

where $\sigma_1(X)$ is the largest singular value of $X$. The singular values of Gaussian matrices are well concentrated and a classical result [34, Theorem II.13] shows that

$$\mathbf{P}\left(\sigma_1(X) > 1 + \sqrt{\frac{r}{m}} + t\right) \leq e^{-mt^2/2}. \tag{2.32}$$

By choosing $t = 1$ in the above formula, we have

$$\|X^*(\hat{e}-e)\|_{\ell_2} \leq 3\|Q^*(\hat{e}-e)\|_{\ell_2} \leq 6\varepsilon$$

with probability at least $1 - e^{-m/2}$ since $\|Q^*(\hat{e}-e)\|_{\ell_2} \leq 2\epsilon$. We now apply Lemma 2.6.4 with $\Phi = X^*$, which gives

$$\begin{aligned}\|e-\hat{e}\|_{\ell_2} &\leq \frac{3\sqrt{6}\,\varepsilon}{a_{3k}(X^*) - \frac{1}{\sqrt{2}}\,b_{2k}(X^*)} \\ &= \sqrt{\frac{m}{r}} \cdot \frac{3\sqrt{6}\,\varepsilon}{a_{3k}(Y^*) - \frac{1}{\sqrt{2}}\,b_{2k}(Y^*)},\end{aligned} \tag{2.33}$$

where $Y = \sqrt{\frac{m}{r}}X$. The theorem is proved since it is well known that if $k \leq c_0 \cdot r/\log(m/r)$ for some constant $c_0$, we have $a_{3k}(Y^*) - \frac{1}{\sqrt{2}}\,b_{2k}(Y^*) \geq c_1$ with probability at least $1 - O(e^{-\gamma' r})$ for some universal constants $c_1$ and $\gamma$; this follows from available bounds on the restricted isometry constants of Gaussian matrices [29, 28, 42, 77].

#### 2.6.3.2 Proof of Corollary 2.3.3

First, we can just assume that $\sigma = 1$ as the general case is treated by a simple rescaling. Put $r = m - n$. Since the random vector $z$ follows a multivariate normal distribution with mean zero and covariance matrix $I_m$ ($I_m$ is the identity matrix in $m$ dimensions), $Q^*z$ is also multivariate normal with mean zero and covariance matrix $Q^*Q = I_r$. Consequently, $\|Q^*z\|_{\ell_2}^2$ is distributed as a chi-squared variable

with $r$ degrees of freedom. Pick $\lambda = \gamma\sqrt{r}$ in (2.20), and obtain

$$\mathbf{P}\left(\|Q^*z\|_{\ell_2}^2 \geq (1+\gamma\sqrt{2}+\gamma^2)r\right) \leq e^{-\gamma^2 r/2}.$$

With $t = \gamma\sqrt{2}+\gamma^2$ so that $\gamma = (\sqrt{1+2t}-1)/\sqrt{2}$, we have $\|Q^*z\|_{\ell_2} \leq \sqrt{r(1+t)}$ with probability at least $1-e^{-\gamma^2(m-n)/2}$. On this event, Theorem 2.3.2 asserts that

$$\|\hat{x}-x\|_{\ell_2} \leq C\sqrt{m(1+t)} + \|x-x^{\mathrm{Ideal}}\|_{\ell_2}.$$

This essentially concludes the proof of the corollary since the size of $\|x-x^{\mathrm{Ideal}}\|_{\ell_2}$ is about $\sqrt{n}$. Indeed, $\|x-x^{\mathrm{Ideal}}\|_{\ell_2}^2 = \|A^*z\|_{\ell_2}^2 \sim \chi_n^2$ as observed earlier. As a consequence, for each $t_0 > 0$, we have $\|x-x^{\mathrm{Ideal}}\|_{\ell_2} \leq \sqrt{n(1+t_0)}\cdot\sigma$ with probability at least $1-e^{-\gamma_0^2 n/2}$, where $\gamma_0$ is the same function of $t_0$ as before. Selecting $t_0$ as $t_0 = m/n$, say, gives the result.

### 2.6.4 The LP decoder

Before we begin, we introduce the number $\theta_{k,k'}$ of a matrix $\Phi \in \mathbb{R}^{r\times m}$ for $k+k' \leq m$ called the $k,k'$-*restricted orthogonality constants*. This is the smallest quantity such that

$$|\langle \Phi v, \Phi v'\rangle| \leq \theta_{k,k'} \cdot \|v\|_{\ell_2}\,\|v'\|_{\ell_2} \tag{2.34}$$

holds for all $k$ and $k'$-sparse vectors supported on disjoint sets. Small values of restricted orthogonality constants indicate that disjoint subsets of columns span nearly orthogonal subspaces. The following lemma which relates the number $\theta_{k,k'}$ to the extremal singular values will prove useful.

**Lemma 2.6.5.** *For any matrix $\Phi \in \mathbb{R}^{r\times m}$, we have*

$$\theta_{k,k'}(\Phi) \leq \frac{1}{2}\,(b_{k+k'}^2(\Phi) - a_{k+k'}^2(\Phi)).$$

*Proof.* Consider two vectors $v$ and $v'$ which are respectively $k$ and $k'$-sparse. By

definition we have

$$2a^2_{k+k'}(\Phi) \le \|\Phi v + \Phi v'\|^2_{\ell_2} \le 2b^2_{k+k'}(\Phi),$$
$$2a^2_{k+k'}(\Phi) \le \|\Phi v - \Phi v'\|^2_{\ell_2} \le 2b^2_{k+k'}(\Phi),$$

and the conclusion follows from the parallelogram identity

$$\begin{aligned} |\langle \Phi v, \Phi v' \rangle| &= \frac{1}{4} \left| \|\Phi v + \Phi v'\|^2_{\ell_2} - \|\Phi v - \Phi v'\|^2_{\ell_2} \right| \\ &\le \frac{1}{2} \left( b^2_{k+k'}(\Phi) - a^2_{k+k'}(\Phi) \right). \end{aligned}$$

□

The argument underlying Theorem 2.4.1 uses an intermediate result whose proof may be found in the Appendix. Here and in the remainder of this chapter, $x_I$ is the restriction of the vector $x$ to an index set $I$, and for a matrix $X$, $X_I$ is the submatrix formed by selecting the columns of $X$ with indices in $I$.

**Lemma 2.6.6.** *Let $\Phi$ be an $r \times m$-dimensional matrix and suppose $T_0$ is a set of cardinality $k$. For a vector $h \in \mathbb{R}^m$, we let $T_1$ be the $k'$ largest positions of $h$ outside of $T_0$. Put $T_{01} = T_0 \cup T_1$ and let $\Phi^*_{T_{01}}$ and $h_{T_{01}}$ be the coordinate restrictions of $\Phi^*$ and $h$ to $T_{01}$, respectively. Then*

$$\|h_{T_{01}}\|_{\ell_2} \le \frac{1}{a^2_{k+k'}(\Phi)} \|\Phi^*_{T_{01}} \Phi h\|_{\ell_2} + \frac{\theta_{k',k+k'}(\Phi)}{a^2_{k+k'}(\Phi)\sqrt{k'}} \|h_{T_0^c}\|_{\ell_1} \tag{2.35}$$

*and*

$$\|h\|^2_{\ell_2} \le \|h_{T_{01}}\|^2_{\ell_2} + \frac{1}{k'} \|h_{T_0^c}\|^2_{\ell_1}. \tag{2.36}$$

#### 2.6.4.1 Proof of Theorem 2.4.1

Just as before, it suffices to show that $\|e - \hat{e}\|_{\ell_2} \le C\sqrt{k} \cdot \lambda \cdot (1 - n/m)^{-1}$. Set $h = \hat{e} - e$ and let $T_0$ be the support of $e$ (which has size $k$). Because $e$ is feasible for $(P'_\infty)$ we

have on the one hand $\|\hat{e}\|_{\ell_1} \leq \|e\|_{\ell_1}$, which gives

$$\begin{aligned} \|e_{T_0}\|_{\ell_1} - \|h_{T_0}\|_{\ell_1} + \|h_{T_0^c}\|_{\ell_1} \leq \|e+h\|_{\ell_1} \leq \|e\|_{\ell_1} \\ \Rightarrow \|h_{T_0^c}\|_{\ell_1} \leq \|h_{T_0}\|_{\ell_1}. \end{aligned}$$

Note that this has an interesting consequence since

$$\|h_{T_0^c}\|_{\ell_1} \leq \|h_{T_0}\|_{\ell_1} \leq \sqrt{k} \cdot \|h_{T_0}\|_{\ell_2} \tag{2.37}$$

by Cauchy Schwarz. On the other hand

$$\|QQ^*h\|_{\ell_\infty} \leq \|QQ^*(\hat{e}-y)\|_{\ell_\infty} + \|QQ^*(y-e)\|_{\ell_\infty} \leq 2\lambda. \tag{2.38}$$

The ingredients are now in place to establish the claim. We set $k' = k$, apply Lemma (2.6.6) with $\Phi = Q^*$ to the vector $h = \hat{e} - e$, and obtain

$$\begin{aligned} \|h\|_{\ell_2} &\leq \sqrt{2}\,\|h_{T_{01}}\|_{\ell_2}, \quad \text{and} \\ \|h_{T_{01}}\|_{\ell_2} &\leq \frac{1}{a_{2k}^2(Q^*) - \theta_{k,2k}(Q^*)}\,\|Q_{T_{01}}Q^*h\|_{\ell_2}. \end{aligned} \tag{2.39}$$

Since each component of $Q_{T_{01}}Q^*h$ is at most equal to $2\lambda$, see (2.38), we have $\|Q_{T_{01}}Q^*h\|_{\ell_2} \leq \sqrt{2k} \cdot 2\lambda$. We then conclude from Lemma 2.6.5 that

$$\|h\|_{\ell_2} \leq 2\sqrt{k} \cdot \frac{2\lambda}{a_{2k}^2(Q^*) + \frac{1}{2}a_{3k}^2(Q^*) - \frac{1}{2}b_{3k}^2(Q^*)}. \tag{2.40}$$

For each $k$, recall the relations $a_k^2(Q^*) = 1 - b_k^2(A^*)$ and $b_k^2(Q^*) = 1 - a_k^2(A^*)$ which give

$$\begin{aligned} \|h\|_{\ell_2} &\leq 4\sqrt{k} \cdot \frac{\lambda}{D}, \\ D &:= 1 - b_{2k}^2(A^*) - \frac{1}{2}b_{3k}^2(A^*) + \frac{1}{2}a_{3k}^2(A^*). \end{aligned}$$

Now just as before, it follows from our definitions that for each $k$, $b_k^2(A^*) \leq \frac{n}{m}(1+\delta_k)$

and $a_k^2(A^*) \geq \frac{n}{m}(1-\delta_k)$. These inequalities imply

$$D \geq 1 - \frac{n}{m}(1+\delta_{2k}+\delta_{3k}).$$

Therefore, if one assumes that

$$\delta_{2k}+\delta_{3k} \leq c\left(\frac{m}{n}-1\right),$$

for some fixed constant $0 < c < 1$, then

$$\|e-\hat{e}\|_{\ell_2} = \|h\|_{\ell_2} \leq \frac{4\sqrt{k}}{1-c} \cdot \frac{\lambda}{1-\frac{n}{m}}.$$

This establishes the first part of the theorem.

We turn to the second part of the claim; if the orthonormal columns of $A$ are chosen uniformly at random, we show that the error bound (2.17) is valid with large probability as long as we have a constant fraction of gross errors. To do this, it suffices to show that the denominator $D$ in (2.40) obeys

$$D \geq \frac{3}{2}a_{3k}^2(Q^*) - \frac{1}{2}b_{3k}^2(Q^*) \geq \frac{r}{2m}.$$

This follows from Lemma 2.6.3. If $k$ is sufficiently small, we have that $a_{3k}^2(Q^*) \geq (7/8)^2 r/m$ and $b_{3k}^2(Q^*) \leq (9/8)^2 r/m$ except on a set of exponentially small probability, which gives

$$D \geq \frac{r}{m}\left(\frac{3}{2}\left(\frac{7}{8}\right)^2 - \frac{1}{2}\left(\frac{9}{8}\right)^2\right) \geq \frac{r}{2m}.$$

#### 2.6.4.2 Proof of Corollary 2.4.2

First, we can just assume that $\sigma = 1$ as the general case is treated by a simple rescaling. The random vector $QQ^*z$ follows a multivariate normal distribution with mean zero and covariance matrix $QQ^*$. In particular $(QQ^*z)_i \sim N(0, s_i^2)$, where $s_i^2 = (QQ^*)_{i,i}$. This implies that $z_i' = (QQ^*z)_i/s_i$ is standard normal with density $\phi(t) = (2\pi)^{-1/2}e^{-t^2/2}$. For each $i$, $\mathbf{P}(|z_i'| > t) \leq \phi(t)/t$ and thus

$$\mathbf{P}\left(\sup_{1\leq i\leq m}|z_i'| \geq t\right) \leq 2m \cdot \phi(t)/t.$$

With $t = \sqrt{2\log m}$, this gives $\mathbf{P}(\sup_{1\le i\le m} |z_i'| \ge \sqrt{2\log m}) \le 1/\sqrt{\pi \log m}$. Better bounds are possible but we will not pursue these refinements here. Observe now that $s_i^2 = \|Q_{i,\cdot}\|_{\ell_2}^2 = 1 - \|A_{i,\cdot}\|_{\ell_2}^2$, and since $\lambda_i = \sqrt{2\log m}\,\|Q_{i,\cdot}\|_{\ell_2}$, we have that

$$|QQ^* z_i| \le \lambda_i, \quad \forall i \tag{2.41}$$

with probability at least $1 - 1/\sqrt{\pi \log m}$.

On the event (2.41), Theorem 2.4.1 then shows that

$$\|\hat{x} - x\|_{\ell_2} \le C\,\sqrt{k} \cdot (m/r) \cdot \max_i |\lambda_i| + \|x - x^{\mathrm{Ideal}}\|_{\ell_2}. \tag{2.42}$$

We claim that

$$\frac{\max_i |\lambda_i|}{\sqrt{2\log m}} = \max_i \|Q_{i,\cdot}\|_{\ell_2} \le \sqrt{\frac{3r}{m}} \tag{2.43}$$

with probability at least $1 - 2e^{-\gamma m}$ for some positive constant $\gamma$. Combining (2.42) and (2.43) yields

$$\|\hat{x} - x\|_{\ell_2} \le 2C \cdot \sqrt{\frac{m\log m}{m-n}} \cdot \sqrt{k} + \|x - x^{\mathrm{Ideal}}\|_{\ell_2}.$$

This would essentially conclude the proof of the corollary since the size of $\|x - x^{\mathrm{Ideal}}\|_{\ell_2}$ is about $\sqrt{n}$. Exact bounds for $\|x - x^{\mathrm{Ideal}}\|_{\ell_2}$ are found in the proof of Corollary 2.3.3 and we do not repeat the argument.

It remains to check why (2.43) is true. For $r \ge m/3$ and since $\|Q_{i,\cdot}\|_{\ell_2} \le 1$, the claim holds with probability 1 because $3r/m \ge 1$! For $r \le m/3$, it follows from $\|Q_{i,\cdot}\|_{\ell_2}^2 + \|A_{i,\cdot}\|_{\ell_2}^2 = 1$ that

$$\begin{aligned}\mathbf{P}\left(\max_i \|Q_{i,\cdot}\|_{\ell_2}^2 \ge \frac{2r}{m}\right) &= \mathbf{P}\left(\min_i \|A_{i,\cdot}\|_{\ell_2}^2 \le \frac{n}{m}\left(1 - \frac{r}{n}\right)\right) \\ &\le m\,\mathbf{P}\left(\|A_{1,\cdot}\|_{\ell_2}^2 \le \frac{n}{m}\left(1 - \frac{r}{n}\right)\right).\end{aligned}$$

The claim follows by applying Lemma 2.6.2 since $r/n \le 1/2$.

## 2.7 Discussion

We have introduced two decoding strategies for recovering a block $x \in \mathbb{R}^n$ of $n$ pieces of information from a codeword $Ax$ which has been corrupted both by adversary and small errors. Our methods are concrete, efficient and guaranteed to perform well. Because we are working with real valued inputs, we emphasize that this work has nothing to do with the use of linear programming methods proposed by Feldman and his colleagues to decode binary codes such as turbo-codes or low-density parity check codes [51, 53, 52]. Instead, it has much to do with the recent literature on compressive sampling or compressed sensing [26, 29, 40, 87, 33, 76], see also [90, 67] for related work.

On the practical end, we truly recommend using the two-step refinement discussed in Section 2.5—the reprojection step—as this really tends to enhance the performance. We anticipate that other tweaks of this kind might also work and provide additional enhancement. On the theoretical end, we have not tried to obtain the best possible constants and there is little doubt that a more careful analysis will provide sharper constants. Also, we presented some results for coding matrices with orthonormal columns for ease of exposition but this is unessential. In fact, our results can be extended to nonorthogonal matrices. For instance, one could just as well obtain similar results for $m \times n$ coding matrices $A$ with independent Gaussian entries.

There are also variations on how one might want to decode. We focused on constraints of the form $\|P_{V^\perp}\tilde{z}\|$ where $\|\cdot\|$ is either the $\ell_2$ norm or the $\ell_\infty$ norm, and $P_{V^\perp}$ is the orthoprojector onto $V^\perp$, the orthogonal subspace to the column space of $A$. But one could also imagine choosing other types of constraints, e.g., of the form $\|X^*\tilde{z}\|_{\ell_2} \leq \varepsilon$ for $(P_2)$ or $\|XX^*\tilde{z}\|_{\ell_\infty} \leq \lambda$ for $(P_\infty)$ (or constraints about the individual magnitudes of the coordinates $(XX^*\tilde{z})_i$ in the more general formulation), where the columns of $X$ span $V^\perp$. In fact, one could choose the decoding matrix $X$ *first*, and then $A$ so that the ranges of $A$ and $X$ are orthogonal. Choosing $X \in R^{m\times r}$ with i.i.d. mean-zero Gaussian entries and applying the LP decoder with a constraint on $\|XX^*\tilde{z}\|_{\ell_\infty}$ instead of $\|\tilde{z}\|_{\ell_\infty}$ would simplify the argument since restricted isometry constants for Gaussian matrices are already readily available [29, 28, 42, 77]!

Finally, we discussed the use of coding matrices which have fast algorithms, thus enabling large scale problems. Exploring further opportunities in this area seems a worthy pursuit.

## 2.8 Appendix: Proof of Lemma 2.6.6

The proof is a variation on that of Lemma 3.1 in [30]. In the sequel, $T_0 \subset \{1, \ldots, m\}$ is a set of size $k$, $T_1$ is the $k'$ largest positions of $h$ outside of $T_0$, $T_{01} = T_0 \cup T_1$. Next, divide $T_0^c$ into subsets of size $k'$ and enumerate $T_0^c$ as $n_1, n_2, \ldots, n_{m-|T_0|}$ in decreasing order of magnitude of $h_{T_0^c}$. Set $T_j = \{n_\ell, (j-1)k' + 1 \le \ell \le jk'\}$. That is, $T_1$ is as before and contains the indices of the $k'$ largest coefficients of $h_{T_0^c}$, $T_2$ contains the indices of the next $k'$ largest coefficients, and so on.

Observe that $\Phi h_{T_{01}} = \Phi h - \sum_{j \ge 2} \Phi h_{T_j}$ so that

$$\|\Phi h_{T_{01}}\|_{\ell_2}^2 = \langle \Phi h_{T_{01}}, \Phi h \rangle - \sum_{j \ge 2} \langle \Phi h_{T_{01}}, \Phi h_{T_j} \rangle.$$

On the one hand, we have

$$|\langle \Phi h_{T_{01}}, \Phi h \rangle| = \langle h_{T_{01}}, \Phi_{T_{01}}^* \Phi h \rangle \le \|h_{T_{01}}\|_{\ell_2} \, \|\Phi_{T_{01}}^* \Phi h\|_{\ell_2},$$

and on the other

$$|\langle \Phi h_{T_{01}}, \Phi h_{T_j} \rangle| \le \theta_{k+k',k'} \|h_{T_{01}}\|_{\ell_2} \|h_{T_j}\|_{\ell_2}.$$

This gives

$$\begin{aligned} a_{k+k'}^2 \|h_{T_{01}}\|_{\ell_2}^2 &\le \|\Phi h_{T_{01}}\|_{\ell_2}^2 \\ &= \|h_{T_{01}}\|_{\ell_2} \, (\|\Phi_{T_{01}}^* \Phi h\|_{\ell_2} + \theta_{k+k',k'} \sum_{j \ge 2} \|h_{T_j}\|_{\ell_2}), \end{aligned} \tag{2.44}$$

where for simplicity, we have omitted the dependence on $\Phi$ in the constants $a_k(\Phi)$ and $\theta_{k,k'}(\Phi)$. We then develop an upper bound on $\sum_{j \ge 2} \|h_{T_j}\|_{\ell_2}$ as in [27]. By construction, the magnitude of each coefficient in $T_{j+1}$ is less than the average of

the magnitudes in $T_j$,

$$\|h_{T_{j+1}}\|_{\ell_\infty} \leq \|h_{T_j}\|_{\ell_1}/k' \quad \Rightarrow \quad \|h_{T_{j+1}}\|_{\ell_2}^2 \leq \|h_{T_j}\|_{\ell_1}^2/k'.$$

Therefore,

$$\sum_{j\geq 2} \|h_{T_j}\|_{\ell_2} \leq \sum_{j\geq 1} \|h_{T_j}\|_{\ell_1}/\sqrt{k'} = \|h\|_{\ell_1(T_0^c)}/\sqrt{k'}. \tag{2.45}$$

Hence, we deduce from (2.44) that

$$\|h_{T_{01}}\|_{\ell_2} \leq \frac{\|\Phi^*_{T_{01}}\Phi h\|_{\ell_2}}{a^2_{k+k'}} + \frac{\theta_{k+k',k'}\,\|h\|_{\ell_1(T_0^c)}}{a^2_{k+k'}\,\sqrt{k'}},$$

which proves the first part of the lemma.

For the second part, it follows from (2.45) that

$$\|h_{T_{01}^c}\|_{\ell_2} = \|\sum_{j\geq 2} h_{T_j}\|_{\ell_2} \leq \sum_{j\geq 2} \|h_{T_j}\|_{\ell_2} \leq \|h_{T_0^c}\|_{\ell_1}/\sqrt{k'}.$$

# Chapter 3

# Incoherence and sparsity oracle inequalities

## 3.1 Abstract

This chapter is concerned with the problem of recovering a sparse signal $x \in \mathbb{R}^n$ from a few noisy linear measurements $y \in \mathbb{R}^n$, or, alternatively, recovering a noisy signal $f \in \mathbb{R}^m$ that has a sparse representation, $f = Ax$. In other words, if $y = Ax + z$, where $A \in \mathbb{R}^{m \times n}$, $m < n$, and $z \in \mathbb{R}^m$ is a noise vector, we would like to find good estimates of $x$ or $Ax$. We require that the estimates be found in a computationally reasonable manner.

We explore what types of results are even possible by deriving loss bounds assuming one is provided with additional information. These bounds are known as sparsity oracle inequalities. We then detail existing results and examine under what conditions they come close to achieving oracle optimality, without access to the oracle information.

Finally, we introduce our results which obtain oracle optimality when estimating $x$ up to log factors and a factor of $\sqrt{s}$. We do this by solving convex optimization programs—the Dantzig selector if the noise is Gaussian, and a second-order cone program if the noise is deterministic and has a bounded $\ell_2$ norm. We require $A$ to obey an incoherence property, which allows the coherence of $A$ to be as large as $O(1/\log n)$ and still allows sparsities as large as $O(m/\log n)$. This is in contrast to other existing results involving coherence where the coherence can only be as large as $O(1/\sqrt{n})$ to allow sparsities as large as $O(\sqrt{n})$. We also do not make the common

assumption that the matrix $A$ obeys a uniform uncertainty principle, or some other eigenvalue restriction.

## 3.2 Introduction

We are interested in the following setup: we have noisy linear measurements of a sparse signal. The signal can be sparse in the sense that it has only a few non-zero coefficients, or sparse in the sense that its coefficients quickly decay. We will consider both stochastic noise, in which case we will assume the noise is Gaussian, $z \sim N(0, I\sigma^2)$, and deterministic noise with bounded $\ell_2$ norm, $\|z\|_{\ell_2} < \epsilon$. Because the measurements are linear, we think of having a measurement matrix $A \in \mathbb{R}^{m \times n}$. We are interested in the case where we take fewer measurements than the dimension of the signal, so $m < n$. Finally, we assume that the columns of $A$ have unit norm.

In the language of statistics, we are in the linear regression setup where the number of observations or samples is less than the number of parameters,

$$Ax + z = y.$$

This is also called parametric regression with fixed linear design, where $A$ is the design matrix, the columns of $A$ are predictors, $x$ is a vector of model coefficients or variables, and $y$ is a vector of observations called the response.

We would like to estimate the signal $x$ and quantify the coefficient loss, or estimate $Ax$ and quantify the prediction loss. We require that the method of obtaining the estimates be computationally tractable. In this chapter, we derive estimates for $x$ using a linear program known as the Dantzig selector [30] in the case of stochastic noise, and a second-order cone program in the case of deterministic noise, that are oracle optimal up to log factors and factors of $\sqrt{s}$. What is novel about our results is that we require the matrix $A$ to obey an incoherence property instead of the more common uniform uncertainty principle, or some other similar eigenvalue restriction.

### 3.2.1 Organization of chapter

The rest of this chapter is organized as follows. In Section 3.3 we discuss the notion of oracle optimality and derive benchmark estimates for various scenarios. In Section 3.4 we show that under a condition known as the uniform uncertainty principle, all the benchmark estimates can be met. In Section 3.5 we discuss a coherence condition and mention results that are able to achieve some oracle optimality success. In Section 3.6 we introduce our contributions and in Section 3.7 we briefly mention other, related work. In Section 3.8 we prove our theorems and, finally, in Section 3.9 we conclude with a discussion of our results.

## 3.3 Oracles and optimality

We are interested in finding good estimates of $x$ or $Ax$, but it is worthwhile to pause and first ask what types of results are even possible. The standard benchmark for problems of this type is to compare the performance of a particular method to what is possible if a little extra information is provided by an oracle. An oracle, in the tradition of the Greeks, is something that can provide information beyond what is usually known. We begin by calculating how well we can estimate $x$ or $Ax$ with the assistance of an oracle.

### 3.3.1 Exactly sparse signal, stochastic noise

For example, assuming that $x$ is sparse, if an oracle were to tell us the support $T$ of $x$ in advance, one could reconstruct an estimate of $x$, call it $x^\star$, using least-squares projection onto the support. This is a reasonable reconstruction strategy as $x^\star$ is the maximum-likelihood estimator over all signals supported on $T$. The estimate $x^\star$ is then given by

$$x^\star = \begin{cases} (A_T^* A_T)^{-1} A_T^* y & \text{on } T \\ 0 & \text{on } T^c, \end{cases}$$

where $A_T$, $T \subset \{1, \ldots, n\}$, is the $m \times |T|$ submatrix obtained by extracting the columns of $A$ corresponding to the indices in $T$. In other words, recalling that

$y = A_T x_T + z$ (we have used the fact that $x$ is only supported on $T$), we would have

$$x_T^\star = x_T + (A_T^* A_T)^{-1} A_T^* z$$

and $x_{T^c}^\star = 0$. Assuming that the noise is Gaussian, the estimated mean-squared error is then

$$\mathbf{E}\|x^\star - x\|_{\ell_2}^2 = \mathbf{E}\|(A_T^* A_T)^{-1} A_T^* z\|_{\ell_2}^2 = \sigma^2 \mathrm{Tr}((A_T^* A_T)^{-1}).$$

Making the reasonable assumption that the eigenvalues of $A_T^* A_T$ are well-behaved (meaning that they are clustered together away from zero), the ideal expected mean squared error obeys

$$\mathbf{E}\|x^\star - x\|_{\ell_2}^2 \geq \lambda_{\min}((A_T^* A_T)^{-1}) \cdot s \cdot \sigma^2 \geq C \cdot s \cdot \sigma^2.$$

In other words, we expect $\|x^\star - x\|_{\ell_2} \approx O(\sigma\sqrt{s})$. If we can achieve something similar to this bound for a method that does not know the support of $x$ in advance, we declare the method to be $\ell_2$ oracle optimal.

However, the $\ell_2$ norm is not the only metric by which to measure the reconstruction error. Instead of measuring the error between $x$ and $\hat{x}$ in the $\ell_2$ norm, we could measure it in the $\ell_1$ norm. Let $c = (A_T^* A_T)^{-1} A_T^* z$ and $c_j = e_j^* (A_T^* A_T)^{-1} A_T^* z = \langle A_T (A_T^* A_T)^{-1} e_j, z \rangle$, where $e_j$ is the $j$th canonical unit vector. Again, assuming the noise is Gaussian, we note that

$$c_j \sim N(0, \|A_T (A_T^* A_T)^{-1} e_j\|_{\ell_2}^2 \sigma^2),$$

and so write $c_j = \|(A_T A_T^*)^{-1/2} e_j\|_{\ell_2} \sigma \cdot w$ where $w \sim N(0,1)$. Thus we have

$$\begin{aligned}
\mathbf{E}\|(A_T^* A_T)^{-1} A_T^* z\|_{\ell_1} &= \sum_{j=1}^{s} \mathbf{E}|c_j| \\
&= \mathbf{E}|w| \sigma \sum_{j=1}^{s} \|(A_T^* A_T)^{-1/2} e_j\|_{\ell_2} \\
&\geq s \cdot \sigma \cdot \frac{\mathbf{E}|w|}{\sqrt{\|A_T^* A_T\|}}
\end{aligned}$$

where $\mathbf{E}|w|$ is some known constant. We have used the fact that

$$\|(A_T^* A_T)^{-1/2} e_j\|_{\ell_2} \geq \sqrt{\lambda_{\min}((A_T^* A_T)^{-1})} = 1/\sqrt{\lambda_{\max}(A_T^* A_T)}$$

and, again, the eigenvalues of $A_T^* A_T$ are well-behaved. Thus, we have

$$\mathbf{E}\|x^\star - x\|_{\ell_1} \geq C \cdot s \cdot \sigma.$$

Finally, it is possible that instead of being interested in how well the oracle can estimate $x$, one is interested in how well it can estimate $Ax$. In this case, a simple calculation shows that the expected mean-squared error is

$$\mathbf{E}\|Ax^\star - Ax\|_{\ell_2}^2 = \sigma^2 \mathrm{Tr}(A_T (A_T^* A_T)^{-1} A_T^*) = s \cdot \sigma^2.$$

In all of the cases we have discussed so far, the oracle has always given the support of $x$. However, if many of the coefficients of $x$ are very small, it is possible get a smaller expected mean squared error if the oracle instead provides the support of the *significant* coefficients of the signal. For example, consider the case where the non-zero entries of $x$ are such that $|x_i| \ll \sigma$ for all $i$. With this information we could set $x^\star = 0$. The $\ell_2$ squared loss is then $\|x\|^2$, which is smaller than $Cs\sigma^2$.

More generally, if the oracle gives the support of $x$ such that $x$ is above the noise level, i.e., the support $I$ of $x$ such that $|x_i| > \sigma$, we could estimate $x$ using this information as

$$x^\star = \begin{cases} (A_I^* A_I)^{-1} A_I^* y & \text{on } I \\ 0 & \text{on } I^c, \end{cases}$$

where $y = A_I x_I + A_{I^c} x_{I^c} + z$. This gives

$$\begin{aligned} \|x - x^\star\|_{\ell_2}^2 &= \|x_I - x_I^\star\|_{\ell_2}^2 + \|x_{I^c} - x_{I^c}^\star\|_{\ell_2}^2 \\ &= \|(A_I^* A_I)^{-1} A_I^* A_{I^c} x_{I^c} + (A_I^* A_I)^{-1} A_I^* z\|_{\ell_2}^2 + \|x_{I^c}\|_{\ell_2}^2 \\ &= \|(A_I^* A_I)^{-1} A_I^* A_{I^c} x_{I^c}\|_{\ell_2}^2 + \|(A_I^* A_I)^{-1} A_I^* z\|_{\ell_2}^2 \\ &\quad + 2\langle (A_I^* A_I)^{-1} A_I^* A_{I^c} x_{I^c}, (A_I^* A_I)^{-1} A_I^* z \rangle + \|x_{I^c}\|_{\ell_2}^2, \end{aligned}$$

and, again assuming that the noise is Gaussian,

$$\begin{aligned}\mathbf{E}\|x - x^\star\|_{\ell_2}^2 &= \|(A_I^* A_I)^{-1} A_I^* A_{I^c} x_{I^c}\|_{\ell_2}^2 + \sigma^2 \mathrm{Tr}((A_I^* A_I)^{-1}) + \|x_{I^c}\|_{\ell_2}^2 \\ &\geq C\left(\sigma^2 \cdot |I| + \|x_{I^c}\|_{\ell_2}^2\right) \\ &= C \cdot \sum_i \min(x_i^2, \sigma^2).\end{aligned}$$

In other words, we have

$$\|x - x^\star\|_{\ell_2}^2 \approx O\left(\sum_i \min(x_i^2, \sigma^2)\right)$$

which is potentially much smaller than $O(s\sigma^2)$. In some sense this gives a near optimal trade-off between the bias and variance coordinate by coordinate.

In this discussion we have restricted our attention to the $\ell_2$ loss of $x$ for the more refined oracle bound, but similar relations also hold for the $\ell_1$ loss and the $Ax$ loss. We will not go into them here in the interest of space considerations, but they take the expected form.

### 3.3.2 Extension to nearly sparse $x$

Thus far we have only considered estimating a signal $x$ that is exactly sparse. This is a bit unrealistic as many $x$ of interest probably do not satisfy this requirement. A perhaps more reasonable condition is that $x$ has only a few "large" components and the rest, while being non-zero, are small. For example, it is common to consider the situation where the entries of $x$ decay at a given rate.

In this case, instead of giving the support of $x$, the oracle might give the support $T$ of the $s$ largest components of $x$, in absolute value. One could estimate $x$ using this information in exactly the same way as for exactly sparse $x$, as

$$x^\star = \begin{cases} (A_T^* A_T)^{-1} A_T^* y & \text{on } T \\ 0 & \text{on } T^c, \end{cases}$$

where $y = A_T x_T + A_{T^c} x_{T^c} + z$. The math to calculate the expected mean squared error is exactly the same as for when the oracle gives the support of the significant

components of a sparse signal, and gives

$$\begin{aligned}\mathbf{E}\|x - x^\star\|_{\ell_2}^2 &= \|(A_T^* A_T)^{-1} A_T^* A_{T^c} x_{T^c}\|_{\ell_2}^2 + \sigma^2 \mathrm{Tr}((A_T^* A_T)^{-1}) + \|x_{T^c}\|_{\ell_2}^2 \\ &\geq C \cdot (\sigma^2 \cdot s + \|x_{T^c}\|_{\ell_2}^2).\end{aligned}$$

In other words, we have

$$\|x - x^\star\|_{\ell_2}^2 \approx O(\sigma\sqrt{s} + \|x_{T^c}\|_{\ell_2}).$$

If we assume the coefficients of $x$ decay at a given rate, for example if the entries of $x$ are rearranged by order of decreasing magnitude $|x_1| \geq |x_2| \geq \ldots \geq |x_n|$ and the $k$th largest entry satisfies

$$|x_k| < Ck^{-r} \tag{3.1}$$

for some positive constant $C$ and $r \geq 1$, we note that

$$\|x_{T^c}\|_{\ell_2} \leq C' s^{-r+1/2}$$

where $C'$ is another constant.

### 3.3.3 Other extensions: deterministic noise, no noise

So far we have only considered the case where the noise $z$ is Gaussian. However, it is possible to imagine situations where the only thing one knows about the noise is that its $\ell_2$ norm is bounded, i.e., $\|z\|_{\ell_2} < \epsilon$. In this case, following similar reasoning as above for Gaussian noise, it is not hard to show that one has the following approximate bounds for sparse $x$ and signals whose coefficients decay like (3.1), respectively. (See [27] for more details.)

$$\begin{aligned}\|x^\star - x\|_{\ell_2} &\approx O(\epsilon) \\ \|x^\star - x\|_{\ell_2} &\approx O(\epsilon + s^{-r+1/2}).\end{aligned}$$

Finally, we mention the case where there is no noise at all and $x$ is exactly $s$-sparse, so we have $y = Ax = A_T x_T$. In this situation, if the oracle gives the support

of $x$ we expect to be able to recover $x$ exactly, as $x_T = (A_T^* A_T)^{-1} A_T^* y$.

### 3.3.4 Summary of benchmarks

So in summary, we have the following benchmarks. If the noise is Gaussian and $x$ is exactly sparse, then with the assistance of an oracle we can achieve

$$\|x^\star - x\|_{\ell_2} \approx O(\sigma\sqrt{s}) \tag{3.2}$$

$$\|x^\star - x\|_{\ell_1} \approx O(\sigma s) \tag{3.3}$$

$$\|Ax^\star - Ax\|_{\ell_2} \approx O(\sigma\sqrt{s}) \tag{3.4}$$

which we refer to as $\ell_2$, $\ell_1$ *and* $Ax$ *optimality for* $s$*-sparse* $x$, respectively.

If $x$ is sparse but with coefficients below the noise level, we also have the refined benchmark

$$\|x^\star - x\|_{\ell_2}^2 \approx O\left(\sum_i \min(\sigma^2, x_i^2)\right). \tag{3.5}$$

If $x$ is not exactly sparse but has coefficients that decay like (3.1) we have

$$\|x^\star - x\|_{\ell_2} \approx O(\sigma\sqrt{s} + s^{-r+1/2}), \tag{3.6}$$

which we refer to as $\ell_2$ *optimality for non-sparse* $x$. We reiterate that in the case of decaying coefficients, $s$ is the size of the set the oracle returns, whether it be the $s$ largest coefficients of $x$ (in absolute value) or the number of entries of $x$ above the noise level.

If the noise instead is deterministic and satisfies $\|z\|_{\ell_2} < \epsilon$ we have the following $\ell_2$ optimality benchmarks

$$\|x^\star - x\|_{\ell_2} \approx O(\epsilon) \tag{3.7}$$

$$\|x^\star - x\|_{\ell_2} \approx O(\epsilon + s^{-r+1/2}), \tag{3.8}$$

for exactly $s$-sparse $x$ and $x$ with decaying coefficients, respectively. Finally, if there is no noise and $x$ is $s$-sparse, we have that $x^\star = x$.

We stress that we are being very imprecise here and only intend these benchmarks to roughly indicate how well we could hope to do with the assistance of an

oracle. We would of course like to know how close we can get to these optimal behaviors without the use of an oracle, using only computationally efficient methods. Clearly, one could not hope to do as well as with an oracle, but as we will see, it is possible to do nearly as well, up to possible log factors and other additional factors which we will discuss later, by solving convex optimization programs. One can informally think of these additional factors as being the price one pays for not knowing support information about the signal in advance.

## 3.4 The uniform uncertainty principle

In this section we show that if one assumes that the measurement matrix obeys a condition called the *uniform uncertainty principle*, also sometimes called the *restricted isometry property*, which we define below, one can achieve all the oracle inequality benchmarks up to log factors by solving either a linear program in the case of Gaussian noise or a second-order cone program in the case of $\ell_2$ bounded noise.

All of the results involving the restricted isometry property rely on the following definition, which we now state.

**Definition 3.4.1.** *The $s$-restricted isometry constant $\delta_s$ of $A$ is the smallest quantity such that*

$$(1 - \delta_s)\|x\|_{\ell_2}^2 \leq \|Ax\|_{\ell_2}^2 \leq (1 + \delta_s)\|x\|_{\ell_2}^2$$

*holds for all $s$-sparse vectors. A vector is said to be $s$-sparse if it has at most $s$ non-zero entries.*

If the $s$-restricted isometry constant $\delta_s$ is small, this implies that the matrix $A$ acts almost like an isometry on $s$-sparse vectors, hence the name. Another way to interpret the isometry constants is to view them as restrictions on the singular values of submatrices of $A$. If $T \subset \{1, 2, \ldots\}$, is a set that indexes the columns of $A$, and $A_T$ is the matrix formed from the columns of $A$ in $T$, then for all eigenvalues $\lambda$ of $A_T^* A_T$ we have

$$1 - \delta_s \leq \lambda(A_T^* A_T) \leq 1 + \delta_s$$

for all sets $T$ such that $|T| \leq s$.

Using this definition, we are now able to state the existing uniform uncertainty principle results.

### 3.4.1 Existing oracle inequality uniform uncertainty results

#### 3.4.1.1 Stochastic noise

If the noise $z$ is Gaussian, we let the estimate $\hat{x}$ be the solution to the following linear program, known as the Dantzig selector [30].

$$(P_1) \quad \min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|A^*(y - A\tilde{x})\|_{\ell_\infty} \leq \lambda_n \cdot \sigma$$

where $\lambda_n$ is a parameter that must be chosen. We will always select it to be $\sqrt{2 \log n}$ for reasons we discuss below.

The next three theorems presented in this section are minor modifications of Theorems 1.1, 1.2, and 1.3 of [30]. These modifications are not important to our discussion here, but might be of independent interest to other researchers, so we detail them in Appendices A and B at the end of this chapter.

**Theorem 3.4.2 (Theorem 1.1 of [30]).** *Suppose $x \in \mathbb{R}^n$ is any $s$-sparse vector and $\delta_{2s} < \sqrt{2} - 1$. Choose $\lambda_n = \sqrt{2 \log n}$. Then the solution $\hat{x}$ to $(P_1)$ obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \leq C_\delta \cdot \log n \cdot s \cdot \sigma^2$$

*with probability greater than $1 - (\sqrt{\pi \log n})^{-1}$, where $C_\delta$ is a constant that can be explicitly stated in terms of the constant $\delta_{2s}$.*

This theorem shows that $\|\hat{x} - x\|_{\ell_2} \approx O(\sqrt{s}\sigma \log n)$, and should be compared with (3.2). Thus, if the $2s$-restricted isometry constant of $A$ satisfies $\delta_{2s} < \sqrt{2} - 1$, then the solution $\hat{x}$ to the Dantzig selector is $\ell_2$ optimal for all $s$-sparse vectors $x$. We point out that the scaling is not exactly the same as for the oracle because of the presence of the $\log n$ term, but it is somewhat remarkable that without knowing anything at all in advance about the support of $x$ we are able to achieve anything close to the oracle bound.

We will subsequently refer to any bound requiring that a combination of restricted isometry constants be sufficiently small as the *uniform uncertainty prin-*

*ciple*, also sometimes referred to as the *restricted isometry property*. We take this liberty because it might be possible to improve the precise statement of the bound, however the role of the bound in the theorem is always the same.

We would also like to draw attention to the manner in which probability enters Theorem 3.4.2. If we assume that $x$ is a feasible solution to the Dantzig selector, in other words that $\|A^*(Ax-y)\|_{\ell_\infty} = \|A^*z\|_{\ell_\infty} < \sigma \cdot \lambda_n$, then the result would be deterministic. The probability only enters in because for Gaussian noise,

$$\mathbf{P}\left(\|A^*z\|_{\ell_\infty} > \sigma\sqrt{2\log n}\right) \leq (\pi \log n)^{-1/2},$$

hence our choice that $\lambda_n = \sqrt{2\log n}$.

Finally, we point out that if there is no noise, then Theorem 3.4.2 guarantees exact recovery of $s$-sparse $x$. This can be seen by letting $\sigma = 0$ in the theorem, which implies that $\hat{x} = x$. In this case $(P_1)$ reduces to

$$(P_2) \quad \min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad A\tilde{x} = y.$$

In addition to showing exact noiseless recovery, $\ell_2$-optimality, and $Ax$-optimality (this follows trivially as $\|A\hat{x} - Ax\|_{\ell_2} \leq \|A\| \, \|\hat{x} - x\|_{\ell_2}$ and $\hat{x}$ is $\ell_2$-optimal), Theorem 3.4.2 also shows that the solution to the Dantzig selector is $\ell_1$-optimal for Gaussian noise. To see this, we note that in the course of the proof of Theorem 3.4.2, it is shown that

$$\|h_{T_{01}}\|_{\ell_2} \leq C_\delta \cdot \sqrt{s} \cdot \sqrt{2\log n} \cdot \sigma$$

where $h = \hat{x} - x$ and $T_{01} = T \cup T_1$, where $T$ is the support of $x$ and $T_1 \subset \{1, \ldots, n\}$, $|T_1| = s$. $C_\delta$ is, as usual, a constant that can be stated in terms of restricted isometry constants.

Thus we have

$$\begin{aligned}
\|\hat{x} - x\|_{\ell_1} &\leq \|h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1} \\
&\leq 2\|h_T\|_{\ell_1} \\
&\leq 2\|h_{T_{01}}\|_{\ell_1} \\
&\leq C_\delta \cdot \sqrt{2\log n} \cdot \sigma \cdot s
\end{aligned}$$

where the second inequality follows from $\|h_{T^c}\|_{\ell_1} \leq \|h_T\|_{\ell_1}$ (the well known cone-constraint relation [45]) and we have used the fact that $\|h_{T_{01}}\|_{\ell_1} \leq \sqrt{2s}\|h_{T_{01}}\|_{\ell_2}$. Comparing $\|\hat{x} - x\|_{\ell_1} \approx O(\sqrt{\log n} \cdot s \cdot \sigma)$ with (3.3) shows that the solution to the Danzig selector is also $\ell_1$ optimal, up to the factor of $\log n$.

In addition to showing $\ell_2$ optimality, there is even a result in [30] that shows optimality for the more refined oracle estimate.

**Theorem 3.4.3 (Theorem 1.2 of [30]).** *Choose $t > 0$ and set $\lambda_n = (1 + \sqrt{2}t^{-1})\sqrt{2\log n}$. Suppose that $x \in \mathbb{R}^n$ is any $s$-sparse vector with $\delta_{2s} < \sqrt{2} - 1 - t$. Then the solution $\hat{x}$ to $(P_1)$ obeys*

$$\|\hat{x} - x\|_{\ell_2}^2 \leq C_\delta^2 \left(1 + \sum_i \min(x_i^2, \lambda_n^2 \sigma^2)\right)$$

*with large probability, where $C_\delta$ depends only on $\delta_{2s}$.*

This result should be compared with (3.5). We note in particular that the result is oracle optimal up to a log factor in the variance term only. This is slightly different than what appeared in [30] and we refer the reader to Appendix B of this chapter for an explanation of the minor differences.

Thus if the matrix $A$ satisfies the uniform uncertainty principle, the four types of optimality for stochastic noise and exactly $s$-sparse $x$ are achieved, up to factors of $\log n$.

If $x$ is not exactly sparse, but instead has coefficients that decay like (3.1), [30] also proves the following result.

**Theorem 3.4.4 (Theorem 1.3 of [30]).** *Suppose $x \in \mathbb{R}^n$ has coefficients that decay like (3.1) and $\delta_{2s} < \sqrt{2} - 1$ for some fixed $s$. Choose $\lambda_n = \sqrt{2\log n}$. Then $\hat{x}$, the solution to $(P_1)$, satisfies*

$$\|x - \hat{x}\|_{\ell_2}^2 \leq C_\delta \cdot \log n \cdot (s\sigma^2 + C_\delta^2 s^{-2r+1})$$

*with high probability, where $C_\delta$ is a constant which can be stated in terms of the restricted isometry constant $\delta_{2s}$.*

This theorem shows that $\|\hat{x} - x\|_{\ell_2}^2 \approx O(\log n \cdot s \cdot \sigma^2 + s^{-2r+1})$ which should

be compared with (3.6). We see that $\ell_2$ oracle optimality for signals with decaying coefficients is achieved, up to a factor of $\log n$.

#### 3.4.1.2 Deterministic noise

In the case where it is only known that the noise is bounded, i.e., $\|z\|_{\ell_2} < \epsilon$, we let the estimate $\hat{x}$ be the solution to the following convex program, a second-order cone program.

$$(P_3) \quad \min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|A\tilde{x} - y\|_{\ell_2} < \epsilon$$

In [27], and later slightly improved in [19], the following theorems are proven.

**Theorem 3.4.5 (Theorem 1 of [27]).** *Suppose $x \in \mathbb{R}^n$ is any $s$-sparse vector, $A$ satisfies $\delta_{2s} < \sqrt{2} - 1$ and $\|z\|_{\ell_2} \leq \epsilon$. Then the solution $\hat{x}$ to $(P_3)$ obeys*

$$\|\hat{x} - x\|_{\ell_2} \leq C_\delta \epsilon$$

*where $C_\delta$ is a constant that can be stated in terms of the restricted isometry constant $\delta_{2s}$.*

This result should be compared with (3.7).

**Theorem 3.4.6 (Theorem 2 of [27]).** *Suppose $x$ has coefficients that decay like (3.1), $\delta_{2s} < \sqrt{2} - 1$, and $\|z\|_{\ell_2} \leq \epsilon$. The the solution $\hat{x}$ to $(P_3)$ obeys*

$$\|\hat{x} - x\|_{\ell_2} \leq C_{\delta,1}\epsilon + C_{\delta,2} s^{-r+1/2}$$

*where $C_{\delta,1}$ and $C_{\delta,2}$ constants that can be stated in terms of the restricted isometry constant $\delta_{2s}$.*

This result should be compared with (3.8). We note that oracle optimality is achieved in these two theorems, without even spurious factors of $\log n$.

Thus we have shown in this section that if $A$ obeys the uniform uncertainty principle, all of the oracle optimality benchmarks are achieved via the solution of convex optimization programs, up to possible factors of $\log n$.

## 3.5 The coherence property and a statistical description of $x$

Whether a matrix obeys the uniform uncertainty principle or not may be difficult to check. For classes of random matrices one can show that the uniform uncertainty principle holds with high probability [28, 29, 34, 42]. For example, if the entries of $A$ are i.i.d. Gaussian $\sim N(0, 1/m)$ then the uniform uncertainty principle holds with high probability if $s \leq Cm/\log(n/m)$. Similar results hold for Bernoulli matrices (i.i.d. entries $\pm 1/\sqrt{m}$ with probability $1/2$) or a matrix formed from a Fourier matrix with randomly sampled rows.

However, for a given matrix of interest, it may be very difficult to verify that the uniform uncertainty principle holds and one could possibly be reduced to computing the singular values of all $A_T$ for $|T| \leq s$. The UUP is a very strong condition on the matrix and gives correspondingly strong results—results that are uniformly true for *any* $s$-sparse $x$.

It is potentially appealing, then, to weaken the condition on the matrix to something that is easily verified, and in return for this, accept results that are only true for an overwhelming majority of $x$. In other words, we will have weaker requirements on the matrix $A$, but will require a statistical description of $x$ and our results will only hold with a given high probability. The results discussed in this section no longer make use of the uniform uncertainty principle, but instead rely on an incoherence property which we now describe.

We denote by $X_i$ the $i$th column of a matrix $X$ with $r$ columns and introduce the notion of coherence, which essentially measures the maximum correlation between normalized columns of $X$, and is given by

$$\mu(X) = \max_{1 \leq i < j \leq r} \frac{|\langle X_i, X_j \rangle|}{\|X_i\|_{\ell_2} \|X_j\|_{\ell_2}}.$$

We have the following definition:

**Definition 3.5.1.** *A matrix $X$ with $r$ columns is said to obey the incoherence property with constant $c_0$ if*

$$\mu(X) \leq c_0 \cdot (\log r)^{-1}.$$

We also introduce a statistical description of exactly $s$-sparse $x$.

**Definition 3.5.2.** *If $x$ is exactly $s$-sparse, we say it comes from the following generic $s$-sparse model if:*

1. *The support $T \subset \{1, \ldots, n\}$ of $x$ is selected uniformly at random.*
2. *Conditional on $T$, the signs of the entries of $x$ on $T$ are independent and equally likely to be $\pm 1$.*

Later, when we consider signals that are not exactly sparse, we will also need the following statistical model.

**Definition 3.5.3.** *If $x$ is not exactly $s$-sparse, we say it comes from the following generic $s$-support model if:*

1. *The support $T \subset \{1, \ldots, n\}$ of the $s$ largest (in absolute value) coefficients of $x$ is selected uniformly at random.*
2. *Conditional on $T$, the signs of the entries of $x$ on $T$ are independent and equally likely to be $\pm 1$.*

Using these definitions, we are now able to state several results that rely on the incoherence property of a matrix instead of the uniform uncertainty principle.

### 3.5.1 Existing incoherence results

For Gaussian noise and $s$-sparse $x$, oracle optimality up to a log factor for $A\hat{x}$ is achieved in [22], when $\hat{x}$ is the solution of the following convex optimization program

$$(P_4) \quad \min_{\tilde{x}} \frac{1}{2}\|y - A\tilde{x}\|_{\ell_2}^2 + 2\lambda_n \sigma \|\tilde{x}\|_{\ell_1}.$$

This quadratic program, which is known as the lasso [83], is closely related to the Dantzig selector. Again, $\lambda_n$ is a parameter that must be chosen in advance.

What is proven in [22] is the following result showing $Ax$ optimality for $s$-sparse $x$.

**Theorem 3.5.4 (Theorem 1.2 of [22]).** *Suppose that $A$ obeys the incoherence property and $x$ is taken from the generic $s$-sparse model. Suppose that $s \leq c_0 n/(\|A\|^2 \log n)$. Then the solution to $(P_4)$ with $\lambda_n = \sqrt{2 \log n}$ obeys*

$$\|Ax - A\hat{x}\|_{\ell_2}^2 \leq C \cdot \lambda_n \cdot s \cdot \sigma^2$$

*with probability at least $1 - 3n^{-2\log 2} - n^{-1}(4\pi \log n)^{-1/2}$.*

Oracle optimality of $s$-sparse $x$ in $\ell_2$ and $\ell_1$ is also shown in [22], assuming the values of the non-zero components of $x$ are sufficiently above the noise level. This is a straightforward consequence of the following theorem.

**Theorem 3.5.5 (Theorem 1.3 of [22]).** *Let $T$ be the support of $x$ and suppose that*

$$\min_{i \in T} |x_i| > 8\sigma\sqrt{2 \log n}.$$

*Further suppose that $A$ obeys the incoherence property, $x$ is taken from the generic $s$-sparse model, and $s \leq c_o n/(\|A\|^2 \log n)$. Then the solution to $(P_2)$ with $\lambda_n = \sqrt{2 \log n}$ obeys*

$$supp(\hat{x}) = supp(x)$$
$$sgn(\hat{x}) = sgn(x)$$

*with probability at least $1 - 2n^{-1}((2\pi \log n)^{-1/2} + sn^{-1}) - O(n^{-2\log n})$.*

Because the lasso returns the correct support of $x$, one has the same information as given by the oracle and thus can obviously return oracle optimal results, without even having to pay the price of a log factor.

## 3.6 Contributions of this chapter

To summarize what we have discussed so far, if the matrix $A$ obeys the uniform uncertainty principle we have shown that, in the case of stochastic noise, the solution of the Dantzig selector $(P_1)$, obeys our benchmarks (up to log factors) for exactly sparse $x$, and also if $x$ has decaying coefficients. If the noise is deterministic and

bounded in $\ell_2$, then the solution to the second-order cone program ($P_3$) is oracle optimal for sparse $x$ and for $x$ with decaying coefficients. Finally, if the matrix $A$ instead satisfies the incoherence property, $x$ is $s$-sparse and comes from a proper statistical model, and the noise is Gaussian, we have shown that the lasso achieves $Ax$ optimality. In addition, if we assume that the non-zero entries of $x$ are sufficiently above the noise level, we also have $\ell_2$ and $\ell_1$ optimality.

This begs the following questions, which this chapter attempts to answer.

- Assuming $A$ satisfies the incoherence property and $x$ is exactly sparse, how close can we come to $\ell_2$ and $\ell_1$ optimality using the Dantzig selector if we don't assume the non-zero coefficients of $x$ are large? What if $x$ is not exactly sparse?
- What can we say if $A$ satisfies the incoherence property and we have bounded noise?

### 3.6.1 Common hypotheses of our theorems

All of the theorems we present in this section will require that a common set of hypotheses hold, which we now detail. The first two requirements are conditions on the coherence of $A$ and the sparsity of $x$.

H1. $\mu(A) < \dfrac{c_0}{\log n}$

H2. $s < \dfrac{c_1 n}{\|A\|^2 \log n}$

where $c_0$ and $c_1$ are positive constants. H1 requires that $A$ obeys the incoherence property while H2 is a limit on the sparsity of $x$. Furthermore, we require that the constants $c_0$ and $c_1$ that appear in H1 and H2 be sufficiently small so that they satisfy the following two relations

H3. $30c_0 + 12\sqrt{2c_1} + \dfrac{2c_1}{\log n} < \dfrac{1}{4}$

H4. $4c_0 + \sqrt{c_1} < \dfrac{1}{2\sqrt{8(2\log 2 + 1)}}$.

These last two relations are by no means optimal—they come from probability bounds which could most likely be improved, but which would complicate the proofs

of our theorems, and so we have chosen to leave them in this form. These hypotheses appear in all our theorems because for each theorem we need to show that the same certain conditions hold with high probability.

Finally, we point out that while Theorem 3.5.4 and Theorem 3.5.5 (Theorem 1.2 and 1.3 of [22]) also assume H1 and H2, it appears at first glance that H3 and H4 are not necessary in those results. However, a similar requirement that $c_0$ and $c_1$ be sufficiently small is actually implicitly assumed in their proofs.

With these assumptions in mind, we now are able to introduce our results.

### 3.6.2 No noise

We begin by considering the noiseless case, $Ax = y$. We show that if $x$ is from the generic $s$-sparse model and $A$ obeys the incoherence property, than the solution of $(P_2)$ recovers $x$ with high probability. We should point out that this result is essentially contained in Theorem 1.2 of [22], but it was not explicitly stated or proved there, and our proof of it, which is quite different than the proof given in [22], provided much inspiration for our other proofs in this section. Thus we include it here for completeness.

**Theorem 3.6.1.** *Suppose $x$ is taken from the generic $s$-sparse model and that hypotheses H1–H4 hold. Then the solution to*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \textit{such that} \quad A\tilde{x} = y$$

*satisfies $\hat{x} = x$ with probability at least $1 - 8n^{-2\log 2}$.*

Thus we are able to show optimality in the case of $s$-sparse $x$ if the matrix obeys the incoherence property.

### 3.6.3 Stochastic noise

In the following theorem, we are also able to show oracle optimality, almost, if the measurements are corrupted with Gaussian noise and the matrix $A$ obeys the incoherence property.

**Theorem 3.6.2.** *Suppose $x$ is taken from the generic $s$-sparse model and that hypotheses H1–H4 hold. Then the solution to*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \textit{such that} \quad \|A^*(y - A\tilde{x})\|_{\ell_\infty} \leq \lambda_n \cdot \sigma$$

*with $\lambda_n = \sqrt{2 \log n}$ satisfies the following bounds*

$$1. \quad \|x - \hat{x}\|_{\ell_2} \leq s \cdot \sqrt{2 \log n} \cdot \sigma \left(4/\sqrt{s} + 16\right)$$
$$2. \quad \|x - \hat{x}\|_{\ell_1} \leq \sqrt{2 \log n} \cdot \sigma \cdot s^{3/2} \left(8 + 12/s\right)$$

*with high probability.*

The first result of the theorem is of interest as it says that by simply solving the Dantzig selector, a linear program, the $\ell_2$ error between $x$ and its estimate is proportional to the true number of unknowns times the noise level $\sigma$ (up to log factors), where we have only assumed that the measurement matrix $A$ obeys the incoherence property. In other words, we have $\|\hat{x} - x\|_{\ell_2} \approx O(s\sigma\sqrt{2 \log n})$, which should be compared with (3.2). Without the assistance of an oracle and by finding an estimate $\hat{x}$ in a computationally reasonable manner, we have been able to achieve the ideal oracle $\ell_2$ behavior, up to a log factor and a factor of $\sqrt{s}$.

For the second part of the theorem we have $\|x - \hat{x}\|_{\ell_1} \approx O(\sqrt{2 \log n} \cdot \sigma \cdot s^{3/2})$, which should be compared with (3.3). Like in the $\ell_2$ case, we achieve the ideal $\ell_1$ benchmark, up to a log factor and a factor of $\sqrt{s}$.

We also have the following result when $x$ is not exactly $s$-sparse.

**Theorem 3.6.3.** *Suppose $x$ is taken from the generic $s$-support model and has coefficients that decay like* (3.1)*. Further suppose that hypotheses H1–H4 hold. Then the solution to*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \textit{such that} \quad \|A^*(y - A\tilde{x})\|_{\ell_\infty} \leq \lambda_n \cdot \sigma$$

*with $\lambda_n = \sqrt{2 \log n}$ satisfies*

$$\|x - \hat{x}\|_{\ell_2} \leq s \cdot \sqrt{2 \log n} \cdot \sigma \left(4/\sqrt{s} + 16\right) + O(s^{-r+1})$$

*with high probability.*

This shows that $\|x - \hat{x}\|_{\ell_2} \approx O(s\sigma\sqrt{2\log n} + s^{-r+1})$ and should be compared with (3.6). Again, we see that we have achieved the ideal oracle behavior up to a factor of $\sqrt{s}$ (plus a log factor).

### 3.6.4 Bounded noise

We now consider the case of deterministic instead of stochastic noise. If $x$ is exactly sparse we have the following theorem.

**Theorem 3.6.4.** *Suppose $x$ is taken from the generic $s$-sparse model and $\|z\|_{\ell_2} < \epsilon$. Further suppose that hypotheses H1–H4 hold. Then the solution to*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \textit{such that} \quad \|A\tilde{x} - y\|_{\ell_2} < \epsilon$$

*satisfies*

$$\|x - \hat{x}\|_{\ell_2} \leq (2 + 8\sqrt{s})\sqrt{6}\epsilon$$

*with high probability.*

We also have the following result when $x$ is not exactly sparse but has decaying coefficients.

**Theorem 3.6.5.** *Suppose $x$ is taken from the generic $s$-support model and has coefficients that decay like* (3.1)*. Further suppose that hypotheses H1–H4 hold and $\|z\|_{\ell_2} < \epsilon$. Then the solution to*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \textit{such that} \quad \|A\tilde{x} - y\|_{\ell_2} < \epsilon$$

*satisfies*

$$\|x - \hat{x}\|_{\ell_2} \leq \left(2/\sqrt{s} + 8\right)\sqrt{6}\epsilon + O(s^{-r+1})$$

*with high probability.*

These results should be compared with the oracle benchmarks (3.7) and (3.8). We see once again that we have achieved oracle optimality up to log factors and a

factor of $\sqrt{s}$, using only a computationally tractable method to find the estimate of $x$.

## 3.7 Connections with other work

Thus far in relation to our statistical estimation problem, we have almost exclusively mentioned results by Candès, Tao, Romberg, and Plan. This is by no means meant to imply that their work is the only that exists in this field, and we pause here to briefly mention other related results. We are of course unable to be exhaustive, but we hope to at least put our work in a better, more complete perspective.

We mentioned earlier that the restricted isometry constants that play a role in the uniform uncertainty principle can be thought of as bounds on the singular values of submatrices of $A$, where the submatrices are formed by taking subsets of columns of $A$. There are several results [68, 93, 94, 8] that also essentially impose restrictions on the singular values of submatrices of $A$. In [68] by solving the lasso they are able to get a bound on $\|\hat{x} - x\|_{\ell_2}$ by imposing what they call a sparse eigenvalue condition. In [93], a similar eigenvalue condition is imposed, which they call the sparse Riesz condition. They do not assume $x$ is exactly sparse but almost exactly sparse and get bounds on $\|Ax - A\hat{x}\|_{\ell_2}$ and $\|x - \hat{x}\|_{\ell_p}$ for $p \geq 1$. Similar results are also shown in [94] for $\|x - \hat{x}\|_{\ell_p}$. Finally, under a more general eigenvalue restriction, in [8] they bound $\|x - \hat{x}\|_{\ell_p}$ for $1 \leq p \leq 2$ and show oracle inequalities of $\|f - A\hat{x}\|_{\ell_2}$ for both Dantzig and lasso in the non-parametric regression model.

In addition to papers that impose eigenvalue constraints, there are also papers that instead impose a mutual coherence requirement [12, 13, 44]. Their requirement is different from the incoherence property we require, in that they assume $s \leq c/\mu(A)$. In [12, 13] they are able to get oracle inequality bounds on $\|Ax - A\hat{x}\|_{\ell_2}$ and $\|x - \hat{x}\|_{\ell_1}$ using a lasso-like penalty with stochastic noise. In the case of deterministic noise, [44] proves similar results. However, in order to allow as large of a sparsity as possible (the results are of course most interesting and widely applicable when $x$ does not have to be *that* sparse), the coherence must be minimal, i.e., $\mu \approx 1/\sqrt{n}$. Even in this maximum sparsity case, it is still required that $s \lesssim \sqrt{n}$ because $s \leq O(1/\mu(A))$. In contrast, the incoherence property of this chapter allows the coherence to be

as large as $O((\log n)^{-1})$, and still has the possibility that the result will hold for sparsities as large as $s = O(m/\log n)$. We discuss this further in Section 3.9.

We conclude this section by mentioning that in addition to the linear regression setup of this chapter, there are many related results involving variations on this theme, including non-parametric setups, extensions where the goodness of fit term is not the residual sum of squares [89], model selection scenarios where it is not required that the estimate be determined in a computationally feasible manner [9, 6], and regression with random instead of fixed design [61], to name just a few.

To elaborate this last result a little further, [61] is able to extend the results of [30] when the regression matrix has random design. In [30], as we have been discussing, it is shown that if the regression matrix satisfies the UUP then oracle optimality is achieved up to log factors and constants. Also, it is known that the UUP holds with high probability if the matrix has i.i.d. Gaussian or Bernoulli entries, for example. In [61], however, it is assumed that the columns of the design matrix are $a_k(X_j)$ where the $X_j$ are i.i.d. random variables in a measurable space with distribution $\Pi$. This of course includes Gaussian and Bernoulli matrices. Then the restricted isometry constants are given a slightly different definition of

$$(1-\delta)\|x\|_{\ell_2} \leq \|\sum_{j=1}^{p} x_j a_j\|_{L_2(\Pi)} \leq (1+\delta)\|x\|_{\ell_2}$$

for all $s$-sparse $x$. We note that in this new definition of restricted isometry, Gaussian or Bernoulli matrices have $\delta_s(\Pi) = 0$ because sparse column subsets are orthonormal systems in $L_2(\Pi)$, even though they are not orthornormal in $\ell_2$.

What [61] shows is that if $\delta_{3s}$ in this new definition is sufficiently small, oracle optimality is achieved. This extends the results of [30] to random design matrices beyond Gaussian and Bernoulli.

## 3.8 Proofs

All of our proofs require that the following three deterministic conditions hold. Define $w$ as $w = A_T(A_T^* A_T)^{-1}\text{sgn}x_T$.

1. *Invertibility condition.* The eigenvalues of $A_T^* A_T$ satisfy $\lambda(A_T^* A_T) \in [1/2, 3/2]$.

The bounds 1/2 and 3/2 are arbitrary; we just need that the smallest eigenvalue is bounded away from zero. Requiring $\lambda(A_T^* A_T) \in [1-\alpha, 1+\alpha]$ for $0 < \alpha < 1$ would also work.

2. *Duality condition.* The vector $w$ obeys $\|A_{T^c}^* w\|_{\ell_\infty} < 1/2$. Again, the number 1/2 is arbitrary; we just need $\|A_{T^c}^* w\|_{\ell_\infty} < \alpha$ for some $0 < \alpha < 1$.

3. *Noise condition.* The noise vector $z$ obeys $\|A^* z\|_{\ell_\infty} \leq \lambda_n$.

Note that an immediate consequence of the invertibility condition is that the submatrix $A_T^* A_T$ is invertible (and thus $w$ exists) and obeys

$$\|(A_T^* A_T)^{-1}\| \leq 2.$$

We also have $\|A_T^*\| = \|A_T\| \leq \sqrt{3/2}$.

We denote the columns of $A$ by $v_j$, $j = 1, \ldots, n$ and point out that $\|A_{T^c}^* w\|_{\ell_\infty} < 1/2$ is equivalent to $|\langle v_j, w\rangle| < 1/2$ for $j \in T^c$, and also that $A_T^* w = \mathrm{sgn} x_T$ is equivalent to $\langle v_j, w\rangle = \mathrm{sgn} x_j$ for $j \in T$. We will find it convenient in the course of the proofs to switch back and forth between these notations.

We will first prove our theorems assuming these conditions hold, and then prove that under the hypotheses of our theorems the conditions hold with large probability, hence proving our theorems.

### 3.8.1 Proof of Theorem 3.6.1

The proof of Theorem 3.6.1 is basically just a result of duality theory of convex optimization. Our argument basically follows that in Theorem 1.4 in [28], and is also related to the proof of Lemma 2.1 in [26].

Because our linear program is convex, we know at least one minimum exists, call it $\hat{x}$. Recall that the support of $x$ is $T$. Let $J = T \cup T^c$. Because $x$ and $\hat{x}$ are both feasible solutions and $y = A\hat{x} = Ax$ we have $\sum_{j \in T} x_j v_j = \sum_{j \in J} \hat{x}_j v_j = y$.

We need to show that $\hat{x} = x$. We will do this by showing first that $\|\hat{x}\|_{\ell_1} = \|x\|_{\ell_1}$ and then that $\mathrm{supp}(\hat{x}) = T$. Because $A_T^* A_T$ is invertible, there can only be one $x$ supported on $T$ such that $Ax = y$, and so we have that $\hat{x} = x$.

Because $\hat{x}$ is the minimum of our program, we have that

$$\|\hat{x}\|_{\ell_1} \leq \|x\|_{\ell_1}.$$

We also have that

$$\begin{aligned}
\|\hat{x}\|_{\ell_1} &= \sum_{j\in T} \hat{x}_j \text{sgn}(\hat{x}_j) + \sum_{j\in T^c} \hat{x}_j \text{sgn}(\hat{x}_j) \\
&\geq \sum_{j\in T} \hat{x}_j \text{sgn}(x_j) + \sum_{j\in T^c} \hat{x}_j \langle v_j, w\rangle \\
&= \sum_{j\in T} ((\hat{x}_j - x_j) + x_j)\text{sgn}(x_j) + \sum_{j\in T^c} \hat{x}_j \langle v_j, w\rangle \\
&= \sum_{j\in T} |x_j| + \langle w, \sum_{j\in J} \hat{x}_j v_j - \sum_{j\in T} x_j v_j\rangle \\
&= \sum_{j\in T} |x_j| + \langle w, y - y\rangle \\
&= \sum_{j\in T} |x_j| = \|x\|_{\ell_1}.
\end{aligned} \tag{3.9}$$

Thus we have that $\|\hat{x}\|_{\ell_1} = \|x\|_{\ell_1}$ and so the above inequality must hold with equality. So we have

$$\sum_{j\in T} \hat{x}_j \text{sgn}(\hat{x}_j) + \sum_{j\in T^c} \hat{x}_j \text{sgn}(\hat{x}_j) = \sum_{j\in T} \hat{x}_j \text{sgn}(x_j) + \sum_{j\in T^c} \hat{x}_j \langle v_j, w\rangle.$$

This can also be written as

$$\sum_{j\in T} \hat{x}_j (\text{sgn}(\hat{x}_j) - \text{sgn}(x_j)) + \sum_{j\in T^c} \hat{x}_j (\text{sgn}(\hat{x}_j) - \langle v_j, w\rangle) = 0.$$

A subtle but important point to note is that each term in each sum is nonnegative, and hence each term must be zero. Specifically, in the second sum, we must either have $\hat{x}_j = 0$ or $\text{sgn}(\hat{x}_j) - \langle v_j, w\rangle = 0$. However, because $|\langle v_j, w\rangle| < 1$ for all $j \in T^c$, we must have $\hat{x}_j = 0$ and thus the support of $\hat{x}$ is $T$, and the theorem is proven.

### 3.8.2 Proof of Theorem 3.6.2

We begin by proving the first part of Theorem 3.6.2, the $\ell_2$ bound $\|\hat{x} - x\|_{\ell_2}$. Let $h = \hat{x} - x$. We have that

$$\|x - \hat{x}\|_{\ell_2} = \|h\|_{\ell_2} \leq \|h_T\|_{\ell_2} + \|h_{T^c}\|_{\ell_2}.$$

We also have the following bound on $\|h_T\|_{\ell_2}$,

$$\begin{aligned} \|h_T\|_{\ell_2} &\leq 2\|A_T^* A_T h_T\|_{\ell_2} \\ &\leq 2\|A_T^* A h\|_{\ell_2} + 2\|A_T^* A_{T^c} h_{T^c}\|_{\ell_2} \\ &\leq 2\|A_T^* A h\|_{\ell_2} + \|h_{T^c}\|_{\ell_1}. \end{aligned} \tag{3.10}$$

The first inequality follows from

$$\begin{aligned} \|h_T\|_{\ell_2}^2 = \langle h_T, h_T \rangle &= \langle (A_T^* A_T)^{-1} A_T^* A_T h_T, h_T \rangle \\ &= \langle A_T^* A_T h_T, (A_T^* A_T)^{-1} h_T \rangle \\ &\leq 2\|A_T^* A_T h_T\|_{\ell_2} \|h_T\|_{\ell_2}, \end{aligned}$$

and the last inequality follows from

$$\begin{aligned} \|A_T^* A_{T^c} h_{T^c}\|_{\ell_2} &= \| \sum_{j \in T^c} A_T^* v_j h_j \|_{\ell_2} \\ &\leq \sum_{j \in T^c} \|A_T^* v_j\|_{\ell_2} |h_j| \\ &\leq \max_{j \in T^c} \|A_T^* v_j\|_{\ell_2} \|h_{T^c}\|_{\ell_1} \\ &\leq 1/2 \|h_{T^c}\|_{\ell_1}. \end{aligned}$$

We turn now to deriving a bound on $\|h_{T^c}\|_{\ell_1}$. We begin by noting that

$$\begin{aligned} |\hat{x}_j| = \hat{x}_j \mathrm{sgn}(\hat{x}_j) &\geq \hat{x}_j \mathrm{sgn}(x_j) \\ &= (\hat{x}_j - x_j)\mathrm{sgn}(x_j) + |x_j| \\ &= h_j \mathrm{sgn}(x_j) + |x_j|, \end{aligned}$$

which gives

$$\|\hat{x}_T\|_{\ell_1} \geq \|x_T\|_{\ell_1} + \langle \text{sgn} x_T, h_T \rangle.$$

Because $\|A^*(Ax - y)\|_{\ell_\infty} = \|A^* z\|_{\ell_\infty} \leq \lambda_n$ by the noise condition, we have that $x$ is feasible, and thus we know that $\|\hat{x}\|_{\ell_1} \leq \|x\|_{\ell_1}$ which implies

$$\begin{aligned} \|x\|_{\ell_1} &\geq \|\hat{x}_T\|_{\ell_1} + \|\hat{x}_{T^c}\|_{\ell_1} \\ &\geq \|x\|_{\ell_1} + \langle \text{sgn}(x_T), h_T \rangle + \|h_{T^c}\|_{\ell_1} \end{aligned}$$

where we have used the fact that $x$ is only supported on $T$. This implies

$$\|h_{T^c}\|_{\ell_1} \leq |\langle h_T, \text{sgn}(x_T) \rangle|.$$

We turn now to bounding the inner product.

$$\begin{aligned} |\langle h_T, \text{sgn}(x_T) \rangle| &= |\langle (A_T^* A_T)^{-1} (A_T^* A_T) h_T, \text{sgn}(x_T) \rangle| \\ &= |\langle A_T^* A_T h_T, (A_T^* A_T)^{-1} \text{sgn}(x_T) \rangle| \\ &\leq |\langle A_T^* A h, (A_T^* A_T)^{-1} \text{sgn}(x_T) \rangle| + |\langle A_T^* A_{T^c} h_{T^c}, (A_T^* A_T)^{-1} \text{sgn}(x_T) \rangle| \\ &\leq \|A_T^* A h\|_{\ell_2} \|(A_T^* A_T)^{-1} \text{sgn}(x_T)\|_{\ell_2} + |\langle h_{T^c}, A_{T^c}^* w \rangle| \\ &\leq 2\sqrt{s} \|A_T^* A h\|_{\ell_2} + 1/2 \|h_{T^c}\|_{\ell_1}. \end{aligned}$$

Thus we have the following bound on $\|h_{T^c}\|_{\ell_1}$,

$$\|h_{T^c}\|_{\ell_1} \leq 4\sqrt{s} \|A_T^* A h\|_{\ell_2}. \tag{3.11}$$

This implies

$$\|h\|_{\ell_2} \leq (2 + 8\sqrt{s}) \|A_T^* A h\|_{\ell_2},$$

and the first part of the theorem is proven as

$$\begin{aligned} \|A_T^* A h\|_{\ell_2} &\leq \sqrt{s}\|A_T^* A h\|_{\ell_\infty} \\ &= \sqrt{s}\|A^*(A\hat{x} - y) - A^*(Ax - y)\|_{\ell_\infty} \\ &\leq \sqrt{s}(\|A^*(A\hat{x} - y)\|_{\ell_\infty} + \|A^* z\|_{\ell_\infty}) \\ &\leq 2\sqrt{s}\lambda_n. \end{aligned} \tag{3.12}$$

This concludes the proof of the first part of Theorem 3.6.2. The proof of the second part of Theorem 3.6.2 is almost identical to the proof of the first part. We have

$$\begin{aligned} \|\hat{x} - x\|_{\ell_1} = \|h\|_{\ell_1} &\leq \|h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1} \\ &\leq \sqrt{s}\|h_T\|_{\ell_2} + \|h_{T^c}\|_{\ell_1}. \end{aligned}$$

The bounds on $\|h_T\|_{\ell_2}$, $\|h_{T^c}\|_{\ell_1}$, and $\|A_T^* A h\|_{\ell_2}$ go through in exactly the same way as in the proof of the first part of the theorem. Thus combining (3.10), (3.11), (3.12) gives

$$\|h\|_{\ell_1} \leq \lambda_n(8s^{3/2} + 12s)$$

and the theorem is proven.

### 3.8.3 Proof of Theorem 3.6.3

The proof of Theorem 3.6.3 is very similar to the proof of Theorem 3.6.2, and so we only outline here the differences which arise in the bound of $\|h_{T^c}\|_{\ell_1}$. We have

$$\|h_{T^c}\|_{\ell_1} = \|\hat{x}_{T^c} - x_{T^c}\|_{\ell_1} \leq \|\hat{x}_{T^c}\|_{\ell_1} + \|x_{T^c}\|_{\ell_1}.$$

Because $x$ is feasible by the noise condition, we know that $\|\hat{x}\|_{\ell_1} \leq \|x\|_{\ell_1}$ which implies

$$\begin{aligned} \|x\|_{\ell_1} = \|x_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1} &\geq \|\hat{x}_T\|_{\ell_1} + \|\hat{x}_{T^c}\|_{\ell_1} \\ &\geq \|x_T\|_{\ell_1} + \langle \mathrm{sgn}(x_T), h_T \rangle + \|h_{T^c}\|_{\ell_1} - \|x_{T^c}\|_{\ell_1}. \end{aligned}$$

Rearranging terms and canceling the $\|x_T\|_{\ell_1}$ implies the following bound on $h_{T^c}$,

$$\|h_{T^c}\|_{\ell_1} \leq 2\|x_{T^c}\|_{\ell_1} - \langle h_T, \mathrm{sgn}(x_T)\rangle \leq 2\|x_{T^c}\|_{\ell_1} + |\langle h_T, \mathrm{sgn}(x_T)\rangle|.$$

The bound on the inner product proceeds the same as Theorem 3.6.2 and thus we have the following bound on $\|h_{T^c}\|_{\ell_1}$

$$\|h_{T^c}\|_{\ell_1} \leq 4\sqrt{s}\|A_T^* Ah\|_{\ell_2} + 4\|x_{T^c}\|_{\ell_1}.$$

This implies

$$\|h\|_{\ell_2} \leq 2\sqrt{s}\lambda_n(2 + 8\sqrt{s}) + 8\|x_{T^c}\|_{\ell_1}.$$

Because the coefficients of $x$ decay like (3.1) we know

$$\|x_{T^c}\|_{\ell_1} < Cs^{-r+1}$$

and the theorem is proven.

### 3.8.4 Proof of Theorems 3.6.4 and 3.6.5

The proof of Theorems 3.6.4 and 3.6.5 is identical to the proofs of Theorems 3.6.2 and 3.6.3 except in the bounding of the term $\|A_T^* Ah\|_{\ell_2}$. Instead we now have

$$\begin{aligned}
\|A_T^* Ah\|_{\ell_2} &\leq \|A_T^*\|\|Ah\|_{\ell_2} \\
&\leq \sqrt{3/2}\|Ax - y - (A\hat{x} - y)\|_{\ell_2} \\
&\leq \sqrt{3/2}(\|z\|_{\ell_2} + \|A\hat{x} - y\|_{\ell_2} \\
&\leq \sqrt{3/2} \cdot 2\epsilon,
\end{aligned}$$

where the first inequality follows from the invertibility condition and the last because both $x$ and $\hat{x}$ are feasible. Thus the theorems are proven.

### 3.8.5 With high probability

We now turn to showing that our deterministic conditions hold with high probability under the hypotheses of our theorems. Our arguments basically follow proofs

presented in [22].

The proof of this relies on the following result of Joel Tropp [86]. The result we state is a slightly refined version which is explained in Section 3.3 of [22].

**Theorem 3.8.1.** *Suppose that $I \subset \{1, \dots, n\}$ is a random subset of columns of a matrix $X$ with at most $s$ elements, where the columns of $X$ have unit-norm. For $q = 2 \log n$,*

$$(\mathbf{E}\|X_I^* X_I - Id\|^q)^{1/q} \leq 2^{1/q} \left( 30\mu(X) \log n + 12 \left( \frac{2s\|X\|^2 \log n}{n} \right)^{1/2} + \frac{2s\|X\|^2}{n} \right). \tag{3.13}$$

*In addition, for the same value of $q$*

$$(\mathbf{E} \max_{i \in I^c} \|X_I^* X_i\|_{\ell_2}^q)^{1/q} \leq 2^{1/q} \left( 4\mu(X)\sqrt{\log n} + \sqrt{\frac{s}{n}}\|X\| \right). \tag{3.14}$$

Now we state and prove that our two conditions hold with high probability under the hypotheses of our theorems.

**Lemma 3.8.2.** *$\|(A_T^* A_T)^{-1}\| \leq 2$ with probability greater than $1 - 2n^{-2\log 2}$.*

*Proof.* Note that $T$ is a random set and put $Z = \|A_T^* A_T - \mathrm{Id}\|$. Clearly, if $Z \leq 1/2$, then all the eigenvalues of $A_T^* A_T$ are in the interval $[1/2, 3/2]$ and $\|(A_T^* A_T)^{-1}\| \leq 2$. Note that when $\mu(A)$ and $s$ obey the hypotheses of the theorem, we have

$$30\mu(A) \log n + 12 \left( \frac{2s\|A\|^2 \log n}{n} \right)^{1/2} + \frac{2s\|A\|^2}{n} < 30c_0 + 12\sqrt{2c_1} + \frac{2c_1}{\log n} < \frac{1}{4}.$$

By Markov's inequality and (3.13) we have

$$\mathbf{P}(Z > 1/2) \leq 2^q \mathbf{E} Z^q \leq 2(1/2)^q.$$

Letting $q = 2 \log n$, the invertibility condition thus holds with probability exceeding $1 - 2n^{-2\log 2}$. □

**Lemma 3.8.3.** *$\|A_{T^c}^* w\|_{\ell_\infty} < 1$ with probability greater than $1 - 4n^{-2\log 2}$.*

*Proof.* We have the following relation

$$\begin{aligned}\|A_{T^c}^* w\|_{\ell_\infty} &= \|A_{T^c}^* A_T (A_T^* A_T)^{-1} \mathrm{sgn} x_T\|_{\ell_\infty} \\ &= \max_{i \in T^c} |Z_i|,\end{aligned}$$

where $Z_i = \sum_{j \in T} W_{ij} \mathrm{sgn} x_j$ and $W_i = (A_T^* A_T)^{-1} A_T^* v_i$. Note that $Z_i$ has two sources of randomness—the set $T$ and the signs of the entries of $x_T$.

Recall the definition of $Z$ from Lemma 3.8.2 and consider the event

$$E = \{Z \leq 1/2\} \cap \{\max_{i \in T^c} \|A_T^* v_i\|_{\ell_2} \leq \gamma\},$$

for some positive $\gamma$. On this event we have

$$\|W_i\|_{\ell_2} \leq \|(A_T^* A_T)^{-1}\| \|A_T^* v_i\|_{\ell_2} \leq 2\gamma.$$

Now

$$\begin{aligned}\mathbf{P}(\{|Z_i| \geq t\} \cap E) &= \mathbf{E}(\mathbf{1}_E \mathbf{1}_{|Z_i| \geq t}) \\ &= \mathbf{E}_T(\mathbf{1}_E \mathbf{E}_{\mathrm{sgn} x_T} \mathbf{1}_{|Z_i| \geq t}) \\ &= \mathbf{E}_T(\mathbf{1}_E \mathbf{P}_{\mathrm{sgn} x_T}(|Z_i| \geq t)).\end{aligned}$$

In order to show that $Z_i$ is small on the event $E$ with high probability we will use a theorem of Hoeffding [57].

**Theorem 3.8.4 (Hoeffding's inequality).** *If $X_1, X_2, \ldots, X_n$ are independent random variables and $a_i \leq X_i \leq b_i$ $(i = 1, 2, \ldots, n)$, then for $t > 0$*

$$\mathbf{P}(S - \mathbf{E}S \geq t) \leq e^{-2t^2 / \sum_{i=1}^n (b_i - a_i)^2},$$

*where $S = \sum_{i=1}^n X_i$.*

For fixed $T$, $Z_i$ is a sum of independent random variables with zero mean because we have assumed that the signs of the entries of $x_T$ are i.i.d. symmetric variables. Also, clearly for each $j \in T$ we have $-|W_{ij}| \leq W_{ij} \mathrm{sgn} x_j \leq |W_{ij}|$. Thus, applying Hoeffding's inequality we have

$$\mathbf{P}(\{|Z_i| \geq t\} \cap E) \leq 2e^{-\frac{t^2}{8\gamma^2}}.$$

Setting $t = 1$ and applying the union bound, we have

$$\mathbf{P}(\{\|(A^*_{T^c} w\|_{\ell_\infty} \geq 1\} \cap E) \leq 2(n-s)e^{-\frac{1}{8\gamma^2}}.$$

Now

$$\begin{aligned}
\mathbf{P}(\|A^*_{T^c} w\|_{\ell_\infty} \geq 1) &\leq \mathbf{P}(\{\|A^*_{T^c} w\|_{\ell_\infty} \geq 1\} \cap E) + \mathbf{P}(E^c) \\
&\leq 2(n-s)e^{-1/8\gamma^2} + \mathbf{P}(Z > 1/2) + \mathbf{P}\left(\max_{i \in T^c} \|A^*_T v_i\| > \gamma\right) \\
&\leq 2ne^{-1/8\gamma^2} + 2n^{-2\log 2} + \mathbf{P}\left(\max_{i \in T^c} \|A^*_T v_i\| > \gamma\right).
\end{aligned}$$

For the first term, if $\gamma$ satisfies

$$\gamma < \frac{1}{\sqrt{8(2\log 2 + 1)}\sqrt{\log n}},$$

then $ne^{-1/8\gamma^2} < n^{-2\log 2}$.

We treat the last term using Markov's inequality. Letting $q = 2\log n$, under the hypotheses of the theorems we have

$$\mathbf{P}\left(\max_{i \in T^c} \|A^*_T v_i\| > \gamma\right) \leq \gamma^{-q}\ \mathbf{E}\left(\max_{i \in T^c} \|A^*_T v_i\|^q\right) \leq 2(\gamma_0/\gamma)^q$$

where

$$\gamma_0 = \frac{4c_0 + \sqrt{c_1}}{\sqrt{\log n}}$$

and we have used (3.14).

Therefore, if $\gamma_0 < \gamma/2$ we have that the last term does not exceed $1 - 2n^{-2\log 2}$. We have indeed that $\gamma_0 < \gamma/2$ as one of the hypotheses of the theorems is that $4c_0 + \sqrt{c_1} < 1/(2\sqrt{8(2\log 2 + 1)})$.

Thus the duality condition holds with probability exceeding $1 - 6n^{-2\log 2}$.

□

## 3.9 Discussion

In this chapter we have derived bounds on the loss of signal estimates for both deterministic and stochastic noise, and with exactly sparse signals and signals with decaying coefficients, when the measurement matrix obeys an incoherence property instead of the more standard uniform uncertainty principle. Having a condition that involves the mutual coherence of a matrix instead of a condition on the singular values of subsets of columns of the matrix is potentially appealing because it is easier to check for a given matrix of interest. Also, uniform uncertainty results are very strong in the sense that they hold uniformly for all sufficiently sparse signals. Introducing a statistical signal model and getting results for signals "on average" has the potential to give results that hold for larger sparsities and are more in line with what is observed in numerical experiments, even for matrices and sparsities for which the UUP no longer holds.

Our results are especially interesting because the incoherence property we require allows the coherence of a matrix to be as large as $O((\log n)^{-1})$, and still permits sparsities as large as $O(m/\log n)$. This sparsity bound comes from the fact that we require that $s < O(n/\|A\|^2 \log n)$ and $\|A\|^2 \geq n/m$ when $A$ has unit-normed columns. (See Chapter 4 of this thesis for a derivation of this bound.) The bound is met when $A$ is a Gaussian matrix, for example [22]. Sparsities as large as $O(m/\log n)$ are also attained when $A$ is the catenation of the spike orthobasis and the Fourier matrix, $A = [IF]$. In this case, $\|A\| = \sqrt{2}$ and $m = 2n$. (Again, see Chapter 4 for more details.) This is an improvement of other results involving coherence which require minimal coherence of $O(1/\sqrt{n})$ in order to achieve maximal sparsities of $O(\sqrt{n})$, as discussed in Section 3.7.

The major deficiency of our results is that they are all a factor of $\sqrt{s}$ from the ideal oracle inequality behavior. It is an open question whether it is possible to remove this undesirable factor.

## 3.10 Appendix A

In this appendix we show that the form of the uniform uncertainty principle stated in Theorems 3.4.2, 3.4.3, and 3.4.4 comes from a simple modification of the proofs

of Theorems 1.1, 1.2 and 1.3 of [30]. We will need the following definition of the $s, s'$-restricted orthogonality constants.

**Definition 3.10.1.** *The $s, s'$-restricted orthogonality constant $\theta_{s,s'}$ of $A$ for $s + s' < n$ is the smallest quantity such that*

$$|\langle A_T x, A_{T'} x' \rangle| \leq \theta_{s,s'} \cdot \|x\|_{\ell_2} \|x'\|_{\ell_2}$$

*holds for all disjoint sets $T, T' \subseteq \{1, \ldots, n\}$ of cardinality $|T| \leq s$ and $|T'| \leq s'$.*

Roughly speaking, a small value of the restricted orthogonality constant $\theta_{s,s'}$ indicates that disjoint subsets of columns of $A$ of size $s$ and $s'$, respectively, span nearly orthogonal subspaces.

In Theorems 1.1, 1.2, and 1.3 of [30], the uniform uncertainty principle is stated as $\delta_{2s} + \theta_{s,2s} < 1$, instead of as $\delta_{2s} < \sqrt{2} - 1$ as it is in Theorems 3.4.2, 3.4.3, and 3.4.4 of this chapter. However, the orthogonality constant $\theta_{s,2s}$ only enters the proofs of Theorems 1.1, 1.2, and 1.3 of [30] in the following bound

$$|\langle A h_{T_{01}}, A h_{T_j} \rangle| \leq \theta_{s,2s} \|h_{T_j}\|_{\ell_2} \|h_{T_{01}}\|_{\ell_2}$$

where $T_{01} = T_0 \cup T_1$ and $T_0$, $T_1$, $T_j$ are disjoint sets of size $s$.

Now, we claim that

$$|\langle A h_{T_{01}}, A h_{T_j} \rangle| \leq \sqrt{2}\delta_{2s} \|h_{T_j}\|_{\ell_2} \|h_{T_{01}}\|_{\ell_2},$$

and so wherever $\theta_{s,2s}$ appears in a proof in [30], it can be replaced by $\sqrt{2}\delta_{2s}$. This then gives the form of the UUP that appears in this chapter.

This modified form of the UUP is perhaps preferable because it does not require the introduction of orthogonality constants, and so the theorem can be stated only in terms of the restricted isometry constant $\delta_{2s}$. It is also appealing to write the UUP as $\delta_{2s} < \sqrt{2} - 1 \approx 0.414$ because in order to guarantee uniqueness of the sparse $x$ we need $\delta_{2s} < 1$. Thus we require $\delta_{2s}$ to be just slightly smaller than what is required for uniqueness. (See [30, 28] for a further discussion of this.)

To prove our claim we will need the following lemma.

**Lemma 3.10.2 (Lemma 2.1 of [19]).** *We have*

$$|\langle Ax, Ax'\rangle| \leq \delta_{s+s'}\|x\|_{\ell_2}\|x'\|_{\ell_2}$$

*for all $x, x'$ supported on disjoint subsets $T, T' \subset \{1, \ldots, n\}$ with $|T| \leq s,\ |T'| \leq s'$.*

*Proof.* This is an application of the parallelogram identity. Suppose $x$ and $x'$ are unit vectors with disjoint support of size $s$ and $s'$, respectively. Then restricted isometry gives

$$2(1-\delta_{s+s'}) = \|x \pm x'\|_{\ell_2}^2(1-\delta_{s+s'}) \leq \|Ax \pm Ax'\|_{\ell_2}^2 \leq \|x \pm x'\|_{\ell_2}^2(1+\delta_{s+s'}) = 2(1+\delta_{s+s'}). \tag{3.15}$$

Now the parallelogram identity asserts that

$$|\langle Ax, Ax'\rangle| = \frac{1}{4}\left|\|Ax + Ax'\|_{\ell_2}^2 - \|Ax - Ax'\|_{\ell_2}^2\right| \leq \delta_{s+s'}$$

where the inequality follows from (3.15). Thus we have

$$|\langle Ax, Ax'\rangle| \leq \delta_{s+s'}$$

for all unit vectors with disjoint support, and the lemma follows from the fact that for any $x$, we can form a unit vector $x/\|x\|_{\ell_2}$, combined with the linearity of the inner product. □

Now we can prove the claim. We have

$$|\langle Ah_{T_{01}}, Ah_{T_j}\rangle| \leq |\langle Ah_{T_0}, Ah_{T_j}\rangle| + |\langle Ah_{T_1}, Ah_{T_j}\rangle|, \tag{3.16}$$

and from Lemma 3.10.2 we know

$$|\langle Ah_{T_x}, Ah_{T_j}\rangle| \leq \delta_{2s}\|h_{T_j}\|_{\ell_2}\|h_{T_x}\|_{\ell_2} \tag{3.17}$$

for $x = 0, 1$. Again using the fact that $T_0$ and $T_1$ are disjoint we also know that

$$\|h_{T_0}\|_{\ell_2} + \|h_{T_1}\|_{\ell_2} \leq \sqrt{2}\|h_{T_{01}}\|_{\ell_2},$$

and the claim follows by combining this with (3.16) and (3.17).

## 3.11 Appendix B

In this appendix we show how minor modifications of the proof of Theorem 1.2 in [30] give the form of Theorem 3.4.3 presented in this chapter. This form is preferable to what appears in [30] because the log factor only multiplies the variance term, instead of both the bias and variance terms. (See [94] for further discussion.)

For ease of exposition, we restate Theorem 3.4.3 here.

**Theorem 3.11.1.** *Choose $t > 0$ and set $\lambda_n = (1 + \sqrt{2}t^{-1})\sqrt{2\log n}$. Then if $x$ is $s$-sparse, $y = Ax + z$ where $z \sim N(0, I\sigma^2)$, $\delta_{2s} < (1-t)(\sqrt{2}-1)$ and $\hat{x}$ is the solution of*

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|A^*(A\tilde{x} - y)\|_{\ell_\infty} < \lambda_n \sigma \tag{3.18}$$

*then*

$$\|\hat{x} - x\|_{\ell_2}^2 \leq C_\delta^2 \left(1 + \sum_i \min(x_i^2, \lambda_n^2 \sigma^2)\right)$$

*with large probability, where $C_\delta$ depends only on $\delta_{2s}$.*

**Remark 3.11.1.** *The 1 in the bound on $\|\hat{x} - x\|_{\ell_2}$ can be removed from the theorem if it is assumed that the largest entry of $\beta$ is strictly greater than the second largest. Alternatively, it can be replaced with any $\epsilon > 0$.*

**Remark 3.11.2.** *We replace $\theta_{s,2s}$ with $\delta_{2s}$ in the theorem without further comment. (See Appendix A.)*

*Proof.* Without loss of generality let $\sigma = 1$ and order the $x_i$s in decreasing order of magnitude

$$|x_1| \geq |x_2| \geq \ldots \geq |x_n|.$$

Set $\lambda = \sqrt{2\log n}$ and let $S_0$ be the largest integer such that

$$\lambda^2 S_0 - 1 \leq \sum_j \min(x_j^2, \lambda^2). \tag{3.19}$$

This implies

$$\lambda^2(S_0+1)-1 \geq \sum_j \min(x_j^2, \lambda^2).$$

As

$$(S_0+1)\min(x_{S_0+1}^2, \lambda^2) \leq \sum_{j=1}^{S_0+1} \min(x_j^2, \lambda^2) \leq \lambda^2(S_0+1)-1,$$

we have

$$\min(x_{S_0+1}^2, \lambda^2) \leq \lambda^2 - \frac{1}{S_0+1} < \lambda^2,$$

which in turn implies that $x_j < \lambda$ for all $j > S_0$.

Write $x = x^{(1)} + x^{(2)}$ where

$$x_j^{(1)} = x_j \cdot 1_{1 \leq j \leq S_0}$$
$$x_j^{(2)} = x_j \cdot 1_{j > S_0}.$$

Note that $x^{(1)}$ is the hard-thresholded version of $x$ on the set $T_0 = \{1, \dots S_0\}$. As $x^{(2)}$ is $s$-sparse (this is immediate as $x$ is $s$-sparse) and

$$\|x^{(2)}\|_{\ell_2}^2 = \sum_{j > S_0} \min(x_j^2, \lambda^2) \leq \lambda^2(S_0+1)-1 \leq 2\lambda^2 S_0, \tag{3.20}$$

we can apply Corollary 6.3 of [30] and decompose $x^{(2)} = x' + x''$, where

$$\|x'\|_{\ell_2} \leq \frac{1+\delta_{2s}}{1-\delta_{2s}(1+\sqrt{2})}\sqrt{2}\lambda\sqrt{S_0}$$
$$\|x'\|_{\ell_1} \leq \frac{1+\delta_{2s}}{1-\delta_{2s}(1+\sqrt{2})}\sqrt{2}\lambda S_0$$
$$\|A^*Ax''\|_{\ell_\infty} < \frac{1-\delta_{2s}^2}{1-\delta_{2s}(1+\sqrt{2})}\sqrt{2}\lambda.$$

Noting that

$$A^*(A(x^{(1)}+x')-y) = -A^*z - A^*Ax'',$$

we have

$$\|A^*(A(x^{(1)}+x')-y)\|_{\ell_\infty} \leq \left(1 + \frac{1-\delta_{2s}^2}{1-\delta_{2s}(1+\sqrt{2})}\sqrt{2}\right)\lambda.$$

By assumption, $t < 1-\delta_{2s}(1+\sqrt{2})$ which implies $x^{(1)}+x'$ is a feasible solution of

(3.18).

The rest of the proof of the theorem is identical to that of Theorem 1.2 of [30], carrying around extra factors of $\sqrt{2}$ which come from the extra factor of 2 in (3.20). We will not repeat these details here. Eventually one has

$$\|x - \hat{x}\|_{\ell_2} \leq C_\delta S_0^{1/2} \lambda \tag{3.21}$$

where $C_\delta$ is a constant that can be explicitly stated in terms of $\delta_{2s}$.

Rearranging (3.19) gives

$$\lambda^2 S_0 \leq 1 + \sum_i \min(x_i^2, \lambda^2)$$

and combining this with (3.21) proves the theorem. □

# Chapter 4

# Compressed sensing of signals with sparse dictionary representations

## 4.1 Abstract

This chapter considers the problem of reconstructing a generic signal $f \in \mathbb{R}^n$ from a limited number of random linear measurements. We are interested in signals that can be sparsely represented by an overcomplete dictionary $\Psi \in \mathbb{R}^{n \times p}$. Instead of requiring that the dictionary obey a restricted isometry property, which loosely speaking requires that the matrix norm of every sparse subset of columns from $\Psi$ be well behaved, we only require that it obey a weaker incoherence property. This incoherence property basically ensures that the columns of the dictionary are not too colinear.

We show that by solving a linear program, we have that $\hat{f} = f$ with high probability if the signal representation in the dictionary is sufficiently sparse, and we explore under what conditions we will be able to reconstruct the signal from an optimal number of measurements. By optimal we mean that the number of measurements is proportional to the sparsity of the signal up to log factors.

We also explore the implications of our results for several example dictionaries and complement our study with numerical simulations showing our method works well.

## 4.2 Introduction

In the past, there has been much interest in finding orthobases in which large classes of signals are sparse or nearly sparse. This is because if a signal has a sparse or nearly sparse representation, compression, approximation and estimation are possible. For example, what underlies JPEG compression of images and transform coders is that a signal is transformed into a basis where it has quickly decaying coefficients and then the small coefficients are set to zero.

Eventually, however, there was a move away from orthobases to overcomplete signal representations as it was realized that certain features of signals are extremely well represented with certain representations (sinusoids by Fourier, point-like singularities by wavelets [65], curves by curvelets [20, 21]), but one single basis seemed unable to represent well *all* features of interest. It was hoped that by forming an overcomplete dictionary of waveforms—a union of several representations, perhaps—one would be able to well-approximate large classes of more complicated signals. However, in an overcomplete dictionary signals no longer have a unique representation, and focus turned to finding the sparsest representation of a signal in a dictionary [31, 66].

Relatively recently there has also been considerable interest in compressed sensing, also known as compressive sampling. Here it is asked whether from relatively few measurements of a sparse signal, it is possible to reconstruct the signal, see [40, 18] and references therein.

This chapter in some sense combines sparse representations and compressed sensing and asks whether from just a few measurements of a signal known to be well represented by a given dictionary, is it possible to reconstruct the signal.

Of course if the combined measurement/dictionary matrix obeys the uniform uncertainty principle, also known as the restricted isometry property, the answer is yes [27, 19]. However, in some sense this isn't very realistic. Classes of matrices known to obey the uniform uncertainty principle with high probability are all random, and while random measurements may make sense, random dictionaries do not. This is because, generally speaking, the waveforms that make up the dictionary need to look like various features of the signal. Random waveforms will look like

noise, and it is unlikely that signals of interest will resemble noise.

In this chapter, instead of requiring the dictionary to obey the uniform uncertainty principle, we require that it obey a weaker incoherence property, which we show is obeyed by dictionaries of interest. For example, an often discussed dictionary is the spikes and sines dictionary, $\Psi = [\mathrm{I}\ \mathrm{F}]$, where I is the identity matrix and F is a basis of sinusoids (a discrete cosine transform). If $n$ is a perfect square, this dictionary does not obey the uniform uncertainty principle for sparsities greater than $\sqrt{n}$ (basically because the Dirac comb can be expressed as a superposition of $\sqrt{n}$ terms in the canonical basis or in the sinusoid basis), but it does obey our incoherence property. We ask whether, if we have a signal that is sparse in the spikes and sines dictionary, it is possible from optimally few measurements (by which we mean that the number of measurements is proportional to the number of dictionary elements in the sparse representation of the signal, up to log factors) to perfectly reconstruct the signal with high probability. The answer this chapter gives is yes.

Before explaining this further, we first turn to detailing our setup and precisely stating our theorem.

### 4.2.1 Setup and statement of theorem

We are interested in the following setup: We have a signal $f \in \mathbb{R}^n$ that is a sparse superposition of elements of a dictionary $\Psi \in \mathbb{R}^{n \times p}$, so that $f$ can be written as $f = \Psi x$, where $x \in \mathbb{R}^p$ is an $s$-sparse vector with support $T$. We further assume that $x$ comes from a random signal model defined as follows:

1. The support $T \subset \{1, \ldots, p\}$ of the $s$ nonzero coefficients of $x$ is selected uniformly at random.

2. Conditional on $T$, the signs of the nonzero entries of $x$ are independent and equally likely to be $\pm 1$.

We assume without loss of generality that $\Psi$ has unit-normed columns. This is possible because we are only interested in estimating $f = \Psi x$ and not $x$. By rescaling $\Psi$ we rescale $x$ but do not affect $f$: $f = \sum_i \psi_i x_i = \sum_i \psi_i / \|\psi_i\|_{\ell_2} \cdot x_i \|\psi_i\|_{\ell_2}$. We further assume that $\Psi$ obeys an incoherence property which we now describe.

We denote by $X_i$ the $i$th column of a matrix $X \in \mathbb{R}^{t\times r}$ and introduce the notion of coherence, which essentially measures the maximum correlation between normalized columns of $X$, and is defined by

$$\mu(X) = \max_{1\le i<j\le r} \frac{|\langle X_i, X_j\rangle|}{\|X_i\|_{\ell_2}\|X_j\|_{\ell_2}}.$$

We have the following definition:

**Definition 4.2.1.** *A matrix $X$ with $r$ columns is said to obey the incoherence property with constant $c_0$ if*

$$\mu(X) \le c_0 \cdot (\log r)^{-1}.$$

We are interested in obtaining an estimate $\hat{f}$ from measurements $y = \Phi f$. We will assume the measurement matrix $\Phi \in \mathbb{R}^{m\times n}$ is a Gaussian matrix properly normalized, i.e., $\phi_{ij} \sim N(0, 1/m)$. This Gaussian assumption is not strictly necessary; what we need is that the entries of $\Phi$ are i.i.d. random variables that obey certain moment bounds, which in turn can be used to show that the random variables satisfy certain concentration inequalities. We will discuss this further in Section 4.7, but note that other random matrix ensembles, such as random projections and the Bernoulli ensemble, would also work. We will also assume that $\Phi$ obeys $m > m_0 = O(\log p/\epsilon^2)$ with $\epsilon = (\log p)^{-(1+\alpha)}$ for some fixed $\alpha > 0$. Thus when $\alpha$ is small, $m \approx (\log p)^3$.

Finally, we will assume that the combined measurement/dictionary matrix $A$ is the product $\Phi\Psi$ which then has its columns normalized to one, so $A = \Phi\Psi D$ where $D = \mathrm{diag}(1/\|\Phi\Psi_i\|_{\ell_2})$. This assumption is not at all necessary, but we make it to simplify the statement of our theorem and our proofs. (In fact, a simple calculation shows that $\mathbf{E}(\|\Phi\Psi_i\|_{\ell_2}^2) = 1$ and so the columns of $A$ are close to being unit-normed without any normalization.)

To get our estimate $\hat{f}$ we first solve the following linear program

$$(P_1) \quad \min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad A\tilde{x} = y$$

to obtain an estimate $\hat{x}$, and then let $\hat{f} = \sum_i \Psi_i \hat{x}_i / \|\Phi\Psi_i\|_{\ell_2}$.

**Theorem 4.2.2.** *Suppose that $\Psi$ obeys the incoherence property with constant $c_0$ and assume that $x$ is taken from the generic $s$-sparse model. Suppose that*

$$s \leq \frac{c_1 p(1-\epsilon)}{(1+\sqrt{n/m}+\delta)^2 \|\Psi\|^2 \log p}$$

*for some positive numerical constants $c_1$ and $\delta$. Finally, suppose that $c_0$, $c_1$ and $p$ satisfy the following two relations:*

1. $30c_0 + \dfrac{60(\log p)^{-\alpha}}{1-(\log p)^{-(1+\alpha)}} + 12\sqrt{2c_1} + \dfrac{2c_1}{\log p} < \dfrac{1}{4}$
2. $4c_0 + \dfrac{8(\log p)^{-\alpha}}{1-(\log p)^{-(1+\alpha)}} + c_1 < \dfrac{1}{2\sqrt{8(2\log 2+1)}}$.

*Then the solution to $(P_1)$ satisfies $\hat{f} = f$ with probability at least $1 - 2e^{-m\delta^2/2} - 4e^{-\frac{m}{2}(\epsilon^2/2-\epsilon^3/3)} - 8p^{-2\log 2}$.*

### 4.2.2 Organization of the chapter

The rest of this chapter is organized as follows. In Section 4.3 we discuss the relation between recovery from an optimal number of measurements and the norm of the dictionary, in Section 4.4 we discuss several example dictionaries, in Section 4.5 we perform numerical experiments to show our method works well, in Section 4.6 we prove our theorem, and, finally, in Section 4.7 we discuss how this work fits in with other, prior work.

## 4.3 Signal recovery from optimal number of measurements

As $s$, the sparsity of the signal, depends on the matrix norm of the dictionary, it is interesting to ask what conditions on $\|\Psi\|$ lead to signal recovery from an optimal number of measurements. Again, by optimal we mean that the number of measurements is proportional to the sparsity of $f$, up to log factors.

We first prove the following proposition:

**Proposition 4.3.1.** *For any $n \times p$ dictionary $\Psi$ with unit normed columns, we have*

*that*

$$\sqrt{\frac{p}{n}} \leq \|\Psi\| \leq \sqrt{p}.$$

*Proof.* We have the following relation between the standard matrix norm and the Frobenius norm, where the Frobenius norm of a matrix is the square-root of the sum of the squares of all of its entries. For a matrix $X$ with $p$ columns we have

$$\|X\| \leq \|X\|_F \leq \sqrt{p}\|X\|.$$

Because all the columns of $\Psi$ have unit-norm, we know $\|\Psi\|_F = \sqrt{p}$ and thus $\|\Psi\| \leq \sqrt{p}$ follows directly.

This upper bound is actually achieved for a dictionary with all identical columns. Let $\psi_i$ be the column that is repeated $p$ times to make up $\Psi$. Then we have

$$\|\Psi\| = \max_{\|x\|_{\ell_2}=1} \|\psi_i \sum_{i=1}^{p} x_i\|_{\ell_2} = \max_{\|x\|_{\ell_2}=1} \left|\sum_{i=1}^{p} x_i\right| = \sqrt{p}.$$

For the lower bound on $\|\Psi\|$, again, because the Frobenius norm of the matrix is $\sqrt{p}$, we know that

$$\sum_{i=1}^{n} \|\psi_i^T\|_{\ell_2}^2 = p$$

where $\psi_i^T$ is the $i$th row of $\Psi$. Thus, there exists a row such that $\|\psi_i^T\|_{\ell_2} \geq \sqrt{p/n}$, and we have

$$\begin{aligned}
\|\Psi\| = \|\Psi^*\| &= \max_{\|x\|_{\ell_2}=1} \|\Psi^* x\|_{\ell_2} \\
&\geq \max_{x=e_i} \|\Psi^* x\|_{\ell_2} \\
&= \max_i \|\psi_i^T\|_{\ell_2} \\
&\geq \sqrt{p/n},
\end{aligned}$$

where $e_i$ is the $i$th standard basis vector.

This lower bound is achieved for a union of orthobases. For a dictionary that is a union of $k$ $n \times n$ orthobases, $\Psi = [O_1 O_2 \cdots O_k]$ (so $p = kn$), we have that $\|\Psi\| = \sqrt{k} = \sqrt{p/n}$. This follows from

$$\begin{aligned}
\|\Psi\|^2 = \|\Psi^*\|^2 &= \max_{\|x\|_{\ell_2}=1} \|\Psi^* x\|_{\ell_2}^2 \\
&= \max_{\|x\|_{\ell_2}=1} \left\| \begin{vmatrix} O_1^* x \\ O_2^* x \\ \vdots \\ O_k^* x \end{vmatrix} \right\|_{\ell_2}^2 \\
&= \max_{\|x\|_{\ell_2}=1} \sum_{j=1}^{k} \|O_j^* x\|_{\ell_2}^2 \\
&= k = \frac{p}{n}
\end{aligned}$$

and thus the proposition is proven. □

If $\|\Psi\|$ is close to its lower bound of $\sqrt{p/n}$, our theorem gives, ignoring $\epsilon$, $\delta$, and assuming $\sqrt{n/m} \gg 1$,

$$s \leq \frac{Cp}{n/m \cdot p/n \cdot \log p} \leq C \frac{m}{\log p}$$

and we attain recovery from an optimal number of measurements.

If $\|\Psi\|$ is instead close to its upper bound of $\sqrt{p}$, we have (again, ignoring $\epsilon$, $\delta$, and assuming $\sqrt{n/m} \gg 1$)

$$s \leq C \frac{m}{n \log p}$$

and we no longer have a guarantee for recovery from a limited number of measurements as the number of measurements must now be greater than an expression that depends on $n$, the length of the signal.

However, it is very important to note that in order to reconstruct $f$ with the optimal number of measurements, we still require that $\Psi$ obey the incoherence property. It is possible that if our dictionary elements are somehow "too similar" the incoherence requirement will not be met even if the dictionary norm is well-behaved. For example, consider the union of two copies of the same orthobasis. Obviously the incoherence property is not satisfied as $\mu(\Psi) = 1$, and so our theorem does not apply, despite the fact that $\|\Psi\| = \sqrt{p/n}$ is optimal.

It is interesting to note, however, that while $\|\Psi\|^2 = p/n$ does not imply that the incoherence property will be met, $\|\Psi\|^2 = p$ implies that the incoherence property will not be met. We show this fact in the next proposition.

**Proposition 4.3.2.** *If $\|\Psi\|^2 = p$ then the incoherence property will not hold.*

*Proof.* We can write the coherence of $\Psi$ as $\mu(\Psi) = \|\Psi^*\Psi - I\|_{\max}$ where the max norm is the maximum of the absolute values of a matrix. We have the following relation between the standard matrix norm and the max norm of a matrix $X \in \mathbb{R}^{r\times t}$

$$\frac{1}{\sqrt{rt}}\|X\| \leq \|X\|_{\max}.$$

Because $\Psi^*\Psi \in \mathbb{R}^{p\times p}$ we thus have

$$\frac{1}{p}\|\Psi^*\Psi - I\| \leq \mu(\Psi).$$

A simple calculation shows that $\|\Psi^*\Psi - I\| = \max\{1, \|\Psi\|^2 - 1\}$. Now, if $\|\Psi\|^2 = p$ then $\|\Psi\|^2 - 1 > 1$ for any reasonable sized $p$ and so we have

$$1 - \frac{1}{p} \leq \mu(\Psi) \leq 1.$$

Thus $\mu(\Psi)$ will not satisfy the incoherence condition if $\|\Psi\| = p$ for reasonable sized $p$. □

## 4.4 Example dictionaries

We discuss in this section several example dictionaries that highlight the uses and limitations of Theorem 4.2.2.

### 4.4.1 Spikes and sines

Returning to the example dictionary of spikes and sines mentioned in the introduction, we first note that the coherence of this dictionary is $\mu(\Psi) = \sqrt{2/n} = 2/\sqrt{p}$. So we will have $\mu(\Psi) \leq c_0/\log p$ with small $c_0$ for reasonable sized $p$, and hence our coherence requirement will be satisfied. Also, we have that $\|\Psi\| = \sqrt{2}$. Thus,

ignoring $\epsilon$, $\delta$, and assuming $\sqrt{n/m} \gg 1$, we have that if

$$s \leq C \frac{m}{\log p}$$

we will be able to reconstruct $f$ with high probability. In other words, if the number of measurements is just slightly more than the sparsity of $f$ in our dictionary, up to log factors, we will be able to recover $f$, and thus we attain optimal recovery.

### 4.4.2 Fourier, wavelet, and ridgelet dictionary

As another example, consider a dictionary that is the union of a Fourier basis, the fine scales of a 2-dimensional orthonormal wavelet basis, and the fine scales of the orthonormal ridgelet basis [38]. In the frequency domain, the basis elements of ridgelets are localized near angular wedges which, at radius $r = 2^j$, have radial extent $\Delta r \approx 2^j$ and angular extent $\Delta\theta \approx 2^{-j}$. In the spatial domain they look like sums of ridge functions.

This dictionary should be good at representing signals that are combinations of point-singularities, oscillations and curves or edges. (We should mention that orthonormal ridgelets are precursors of curvelets [20, 21] which have provably optimal coefficient decay for $C_1$ curves in $\mathbb{R}^2$, i.e., curvelets optimally represent edges. This would be the ideal representation to use, but unfortunately curvelets in their current form are redundant and have high coherence, even at fine scales. We thus choose orthonormal ridgelets which have zero coherence with each other. However, if in the future someone constructs an almost orthonormal representation of curvelets, this could be used in place of ridgelets.)

This dictionary will also satisfy our incoherence requirement, assuming the scales of wavelets and ridgelets are taken to be sufficiently fine. To justify this claim, we first show that sinusoids are incoherent with fine scale wavelets. We will illustrate this in 1-d; the extension to 2-d will be straightforward. Labeling the scales of the wavelet transform by $j \geq j_0$, where $j = j_0$ is the coarsest scale and larger $j$ correspond to increasingly fine scales, and labeling the shift parameter $k = 1, 2, \ldots$, we have

$$\psi_{j,k}(t) = 2^{j/2}\psi(t2^j - k),$$

see [65] for more details about wavelets. Thus we have

$$\hat{\psi}_{j,k}(\xi) = 2^{-j/2}\hat{\psi}(2^{-j}\xi)e^{-i2\pi 2^{-j}k\xi}$$

where $\hat{\psi}(x)$ is some bounded function, $|\hat{\psi}(x)| < C$, and so

$$|\langle \psi_{j,k}, f_\xi \rangle| \leq \frac{C}{2^{j/2}}$$

where $f_\xi$ is an element of the Fourier basis at frequency $\xi$. The extension to 2-d wavelets and 2-d sinusoids is similar and gives $|\langle \psi_{j,k}, f_\xi \rangle| \leq C2^{-j}$. As long as $C2^{-j} < c_0/\log p$, which it will be for sufficiently large $j$ (sufficiently small wavelet scales), our coherence condition will be met for fine scale wavelets and sinusoids of any frequency. Intuitively, this is because at finer and finer scales wavelets start to look like diracs, which are of course incoherent with sinusoids.

Similar calculations for Fourier and ridgelets at scale $j$ give [45]

$$|\langle \rho_\lambda, f_\xi \rangle| < C2^{-j/2}$$

and for ridgelets and wavelets

$$|\langle \rho_\lambda, \psi_{j,k} \rangle| < C2^{-j/2}.$$

Thus, as long as we only include sufficiently fine scale ridgelets and wavelets in our dictionary (sufficiently fine so that the incoherence property is satisfied by the dictionary), we will be able to reconstruct signals that are a sparse superposition of these dictionary elements.

Another possibility instead of including only the fine scale wavelets and ridgelets in the dictionary is, if the coarse scale elements of the different representations span the same space which is orthogonal to the space spanned by the rest of the elements in the dictionary, to measure the projection of the signal onto that space directly and then take random measurements of the residual signal. In other words, measure $P_{\Psi_0} f$ directly (i.e., low pass filter the signal and then finely sample it) and then solve $(P_1)$ using $\Phi(f - P_{\Psi_0} f) = y$.

### 4.4.3 Tight-frame dictionaries

So far we have only discussed example dictionaries where the dictionary elements have unit-norm. As we mentioned earlier, this is not at all a necessary assumption. In fact, one frequently gives up the assumption of unit-normed columns and instead is interested in tight-frame dictionaries with frame-bound 1. This means the dictionary is an isometry and obeys a generalized Parseval's formula

$$\|\Psi^* f\|_{\ell_2} = \|f\|_{\ell_2},$$

for all $f \in \mathbb{R}^n$. Thus we have $\|\Psi\| = 1$.

If we require $\Psi$ to be a tight frame instead of having unit-normed columns, the proof of Theorem 4.2.2 goes through exactly the same as when we assume unit-normed dictionary elements, only we require instead that

$$s \leq \frac{c_1 p(1-\epsilon) \min_i \|\Psi_i\|_{\ell_2}^2}{(1+\sqrt{n/m}+\delta)^2 \log p}.$$

If the tight-frame dictionary has a redundancy related to the length of the dictionary elements, for example $p = cn$ or $p = cn \log n$ for some constant $c$, we again will be able to achieve reconstruction with the optimal number of measurements, assuming that the column norms do not scale too badly.

The only problem that might arise is if the coherence of the dictionary is too high, and it is possible that a dictionary of interest does not satisfy our incoherence requirement. For example, consider the time-frequency Gabor dictionary whose elements consist of a scaled, translated and modulated window function. One would expect this dictionary to be good at representing signals composed of pulses—for example reflection radar signals.

However, the coherence of the dictionary is almost maximal, i.e., $\mu(\Psi_{\text{Gabor}}) \approx 1$. This is because dictionary elements that are at the same location but neighboring scales will have large absolute inner product. To motivate this, we will consider continuous-time Gabor dictionary elements and let the window function be Gaussian and the scales be dyadic. In other words, let the dictionary elements of our Gabor

dictionary look like

$$g_{j,u,\xi}(t) = e^{-\pi\left(\frac{t-u}{2^j}\right)^2} e^{i\xi t}.$$

A simple calculation shows $\|g_{j,u,\xi}\|_{\ell_2}^2 = 2^{j-1/2}$. Also, we have

$$\begin{aligned}\langle g_{j,u,\xi}, g_{j+1,u,\xi+\Delta\xi}\rangle &= \int_{-\infty}^{\infty} e^{-5\pi/4\left(\frac{t-u}{2^j}\right)^2} e^{-i\Delta\xi t}\, dt \\ &= \frac{2^{j+1}}{\sqrt{5}} e^{-\frac{(\Delta\xi)^2 2^2 2^j}{5\pi}} e^{-i\Delta\xi u}.\end{aligned}$$

Putting this all together gives

$$\frac{|\langle g_{j,u,\xi}, g_{j+1,u,\xi+\Delta\xi}\rangle|}{\|g_{j,u,\xi}\|_{\ell_2}\ \|g_{j+1,u,\xi}\|_{\ell_2}} = \frac{1}{\sqrt{1+2^{-2}}} e^{-\frac{(\Delta\xi)^2 2^2 2^j}{5\pi}}.$$

Thus for our dictionary, $\mu(\Psi_{\text{Gabor}}) \gtrsim 0.8944$ for two elements at different scales, even when they are off-frequency, as long as $\Delta\xi$ is not too big and so the exponential term is $\approx 1$. Hence the Gabor dictionary does not obey the incoherence property of our theorem, and our theorem does not apply.

## 4.5 Numerical experiments

In order to investigate how well our method performs empirically, we performed a series of experiments designed as follows:

1. Select $n$ (the size of the input signal) and $m$ (the number of measurements). Select the dictionary $\Psi \in \mathbb{R}^{n\times p}$. Sample the measurement matrix $\Phi \in \mathbb{R}^{m\times n}$ with i.i.d. Gaussian entries.

2. Select $s$, the sparsity of $x$. Select the support of $x$ of size $s$ at random and sample $x$ on its support with i.i.d. Gaussian entries. (The entries do not need to be Gaussian, they should just have random signs.)

3. Make $f = \Psi x$ and $y = \Phi f$.

4. Solve $(P_1)$ to obtain $\hat{f}$ and compare $f$ to $\hat{f}$. Declare success if $\|f - \hat{f}\|_{\ell_2} < 10^{-3}$.

5. Repeat 100 times for each $s$.

6. Repeat for various sizes of $n$ and $m$ and various dictionaries $\Psi$.

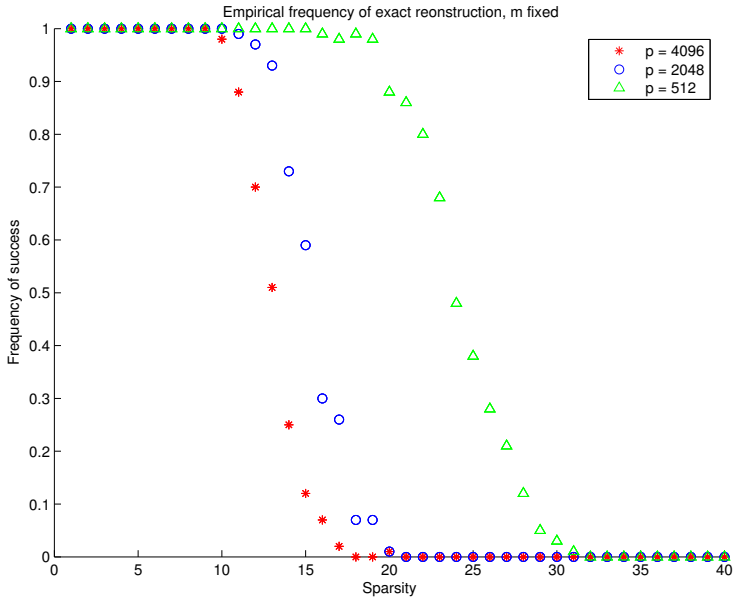



**Figure 4.1.** Empirical frequency of recovery of an input signal from $y = \Phi f$ where $\Phi$ is an $m$ by $n$ matrix with independent Gaussian entries and $f$ is sparse in the spikes and sines dictionary. Here $m = 100$ is fixed, $p = 2n$, and $p$ varies.

The results are presented in Figure 4.1. Figure 4.1 examines the situation where the signal $f$ has a sparse representation in the spikes and sines dictionary, $\Psi$ =[I F]. Here the number of measurements $m$ is kept fixed, $p = 2n$, and $p$ is allowed to vary. Our experiments show that the cutoff point where one no longer recovers $f$ is further out the smaller $p$ is. This agrees with the idea that the sparsity needs to be smaller than something that depends inversely on $p$, i.e., $s \leq Cm/\log p$, if $m$ is fixed.

## 4.6 Proof of Theorem 4.2.2

Our proof requires that the following two deterministic conditions hold. Recall that $A = \Phi\Psi$ and define $w \equiv A_T(A_T^* A_T)^{-1}\mathrm{sgn}x_T$.

1. *Invertibility condition.* The submatrix $A_T^* A_T$ is invertible (and thus $w$ exists) and obeys

$$\|(A_T^* A_T)^{-1}\| \leq 2.$$

The number 2 is arbitrary; we just need the smallest eigenvalue of $A_T^* A_T$ to be bounded away from zero.

2. *Duality condition.* The vector $w$ obeys $\|A_{T^c}^* w\|_{\ell_\infty} < 1$.

We will first prove our theorem assuming these conditions hold, and then prove that under the hypotheses of our theorem the conditions hold with large probability.

### 4.6.1 Proof assuming deterministic conditions

The proof of Theorem 4.2.2, assuming the invertibility and duality conditions, is the same as the proof of Theorem 3.6.1 of Chapter 3 of this thesis, and so we will not repeat the details here.

### 4.6.2 With high probability

We now turn to showing that our deterministic conditions hold with high probability under the hypotheses of Theorem 4.2.2. Our general strategy follows the methods of Section 3.8.5 of Chapter 3 of this thesis, only now we have the added complication of the measurement matrix $\Phi$ combined with the dictionary $\Psi$.

Before we begin, however, we need the following lemma, which we will use to relate the coherence of $A$ to the coherence of $\Psi$.

**Lemma 4.6.1.** *(Johnson-Lindenstrauss.) [59, 3, 5] Let $\Omega$ be the set of vectors*

$$\Omega = \{v \mid v = \Psi_i - \Psi_j,\ i < j \text{ or } \Psi_j = 0\}.$$

*Then because we have assumed $m > m_0 = O(\log p/\epsilon^2)$ for the measurement matrix $\Phi$, where $\epsilon = (\log p)^{-(1+\alpha)}$, we have*

$$(1-\epsilon)\|v\|_{\ell_2}^2 \leq \|\Phi v\|_{\ell_2}^2 \leq (1+\epsilon)\|v\|_{\ell_2}^2$$

*for all $v \in \Omega$ with probability greater than $1 - 2e^{-\frac{m}{2}(\epsilon^2/2 - \epsilon^3/3)}$.*

The lemma allows us to show that if $\Psi$ obeys the incoherence property, then $A$

also obeys the incoherence property with high probability, as we have

$$\begin{aligned}
\mu(A) &= \max_{1\le i<j\le p} |\langle A_i, A_j\rangle| \\
&= \max_{1\le i<j\le p} \frac{|\langle \Phi\Psi_i, \Phi\Psi_j\rangle|}{\|\Phi\Psi_i\|_{\ell_2}\ \|\Phi\Psi_j\|_{\ell_2}} \\
&= \max_{1\le i<j\le p} \frac{|\|\Phi\Psi_i\|_{\ell_2}^2 + \|\Phi\Psi_j\|_{\ell_2}^2 - \|\Phi(\Psi_i-\Psi_j)\|_{\ell_2}^2|}{2\|\Phi\Psi_i\|_{\ell_2}\ \|\Phi\Psi_j\|_{\ell_2}} \\
&\le \max_{1\le i<j\le p} \frac{|(1+\epsilon)(\|\Psi_i\|_{\ell_2}^2 + \|\Psi_j\|_{\ell_2}^2) - (1-\epsilon)\|\Psi_i-\Psi_j\|_{\ell_2}^2|}{2(1-\epsilon)\|\Psi_i\|_{\ell_2}\ \|\Psi_j\|_{\ell_2}} \\
&\le \max_{1\le i<j\le p} \frac{|\|\Psi_i\|_{\ell_2}^2 + \|\Psi_j\|_{\ell_2}^2 - \|\Psi_i-\Psi_j\|_{\ell_2}^2|}{2\|\Psi_i\|_{\ell_2}\ \|\Psi_j\|_{\ell_2}} + \frac{\epsilon}{1-\epsilon}\frac{\|\Psi_i\|_{\ell_2}^2 + \|\Psi_j\|_{\ell_2}^2}{\|\Psi_i\|_{\ell_2}\ \|\Psi_j\|_{\ell_2}} \\
&= \mu(\Psi) + \frac{2\epsilon}{1-\epsilon}
\end{aligned} \tag{4.1}$$

with probability greater than $1 - 2e^{-\frac{m}{2}(\epsilon^2/2-\epsilon^3/3)}$.

We will also need the following theorem of Joel Tropp. The result we use is a slightly refined version of what appears in [86], explained in section 3.3 of [22].

**Theorem 4.6.2.** *Suppose that $I \subset \{1,\dots,n\}$ is a random subset of columns of a matrix $X$ with at most $s$ elements, where the columns of $X$ have unit-norm. For $q = 2\log n$,*

$$(\mathbf{E}\|X_I^* X_I - Id\|^q)^{1/q} \le 2^{1/q}\left(30\mu(X)\log n + 12\left(\frac{2s\|X\|^2\log n}{n}\right)^{1/2} + \frac{2s\|X\|^2}{p}\right). \tag{4.2}$$

*In addition, for the same value of $q$*

$$(\mathbf{E}\max_{i\in I^c}\|X_I^* X_i\|_{\ell_2}^q)^{1/q} \le 2^{1/q}\left(4\mu(X)\sqrt{\log n} + \sqrt{\frac{s}{n}}\|X\|\right). \tag{4.3}$$

Now we state and prove that our two deterministic conditions hold with high probability under the hypotheses of our theorem.

**Lemma 4.6.3.** *$\|(A_T^* A_T)^{-1}\| \le 2$ for sufficiently large $p$ and $n$, with probability greater than $1 - 2p^{-2\log 2} - 2e^{-\frac{m}{2}(\epsilon^2/2-\epsilon^3/3)} - e^{-m\delta^2/2}$.*

*Proof.* Let $Z = \|A_T^* A_T - \mathrm{Id}\|$ and note that it has two sources of randomness—$T$ is a random set and $\Phi$ is a random matrix. Clearly, if $Z \le 1/2$, then all the eigenvalues of $A_T^* A_T$ are in the interval $[1/2, 3/2]$ and $\|(A_T^* A_T)^{-1}\| \le 2$.

Let $E_1$ be the event $E_1 = \{\forall v \in \Omega : |\ \|\Phi v\|_{\ell_2}^2 - \|v\|_{\ell_2}^2\ | < \epsilon \|v\|_{\ell_2}^2\}$ and $E_2$ be the event $E_2 = \{\|\Phi\| \leq 1 + \sqrt{n/m} + \delta\}$.

We have

$$\begin{aligned}\mathbf{P}(Z > 1/2) &= \mathbf{P}(Z > 1/2 \cap E_1 \cap E_2) + \mathbf{P}(Z > 1/2 \cap (E_1 \cap E_2)^c) \\ &\leq \mathbf{P}(Z > 1/2 \cap E_1 \cap E_2) + \mathbf{P}(E_1^c) + \mathbf{P}(E_2^c).\end{aligned}$$

We will first show that

$$\mathbf{P}(Z > 1/2 \cap E_1 \cap E_2) = \mathbf{E}_\Phi(\mathbf{1}_{Z>1/2}\mathbf{1}_{E_1\cap E_2}) = \mathbf{E}_\Phi(\mathbf{1}_{E_1\cap E_2}\mathbf{E}_T\mathbf{1}_{Z>1/2}) \leq 2p^{-2\log 2}.$$

Using Markov's inequality and (4.2) we have

$$\mathbf{E}_T\mathbf{1}_{Z>1/2} \leq 2^q \mathbf{E}_T Z^q \leq 2^{q+1}\left(30\mu(A)\log p + 12\left(\frac{2s\|A\|^2\log p}{p}\right)^{1/2} + \frac{2s\|X\|^2}{p}\right)^q$$

where $q = 2\log p$.

On event $E_1$, $\|D\| = \max_i 1/\|\Phi\Psi_i\|_{\ell_2} \leq 1/\sqrt{1-\epsilon}$ (recall that $\|\Psi_i\|_{\ell_2} = 1,\ \forall i$), and on event $E_2$, $\|\Phi\| \leq 1 + \sqrt{n/m} + \delta$. Thus on $\{E_1 \cap E_2\}$ we have that

$$\|A\| = \|\Phi\Psi D\| \leq \|\Phi\|\|\Psi\|\|D\| \leq (1 + \sqrt{n/m} + \delta)\|\Psi\|/\sqrt{1-\epsilon}.$$

At first glance, bounding the norm of the product by the product of the norms in this step might seem to be too mild—perhaps $\|\Phi\Psi\| \ll \|\Phi\|\|\Psi\|$. However, we are primarily interested in cases where $\Psi$ is a union of orthobases and/or tight frames. In this situation, it is not hard to show that in fact $\|\Phi\Psi\| = \|\Phi\|\|\Psi\|$.

Also on event $E_1$ we have

$$\mu(A) \leq \mu(\Psi) + \frac{2\epsilon}{1-\epsilon} \leq \frac{c_0}{\log p} + \frac{2\epsilon}{1-\epsilon}$$

where the second inequality follows because $\Psi$ obeys the incoherence property and we have used (4.1).

Thus we have

$$\begin{aligned}\mathbf{E}_\Phi(\mathbf{1}_{E_1\cap E_2}\mathbf{E}_T\mathbf{1}_{Z>1/2}) &< 2^{q+1}\left(30c_0+\frac{60(\log p)^{-\alpha}}{1-(\log p)^{-(1+\alpha)}}+12\sqrt{2c_1}+\frac{2c_1}{\log p}\right)^q\\ &< 2^{q+1}\left(\frac{1}{4}\right)^q = 2p^{-2\log 2}\end{aligned}$$

where the last inequality follows from the hypotheses of our theorem.

Now, as $\mathbf{P}(E_1^c) \le 2^{-\frac{m}{2}(\epsilon^2/2-\epsilon^3/3)}$ by Lemma 4.6.1, we turn to showing $\mathbf{P}(E_2^c) \le e^{-m\delta^2/2}$. To do this we will use the well-known Gaussian concentration inequality. (See, for example, [63].) Given i.i.d. standard normal random variables $X_1, X_2, \ldots, X_n$, a Lipschitz function $F : \mathbb{R}^n \to \mathbb{R}$ with Lipschitz constant $L$, i.e.,

$$|F(x) - F(x')| \le L\|x - x'\|_2 \text{ for every } x, x' \text{ in } \mathbb{R}^n,$$

and letting $Y = F(X_1, X_2, \ldots, X_n)$ we have

$$\mathbf{P}(Y - \mathbf{E}Y \ge t) \le e^{-t^2/2L^2}.$$

The function $F$ we are interested in for our purposes is the matrix norm, which has Lipschitz constant 1 as

$$|\ \|M_1\|_2 - \|M_2\|_2\ | \le \|M_1 - M_2\|_2 \le \|M_1 - M_2\|_{\mathrm{F}}$$

where $M_1$ and $M_2$ are matrices in $\mathbb{R}^{m\times n}$ and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm (the Euclidian norm for the matrix treated as a vector in $\mathbb{R}^{mn}$).

Letting $Z \in \mathbb{R}^{m\times n}$, $Z_{ij} \sim N(0,1)$, gives

$$\mathbf{P}(\|Z\| > \mathbf{E}\|Z\| + t) \le e^{-t^2/2}.$$

In Theorem 2.13 of [34] it is shown that

$$\mathbf{E}\|Z\| \le \sqrt{m} + \sqrt{n}.$$

As $\Phi = 1/\sqrt{m}\ Z$, we have

$$\mathbf{P}(\|\Phi\| > 1 + \sqrt{n/m} + \delta) \leq e^{-m\delta^2/2},$$

as desired, and hence the lemma is proven.

□

**Lemma 4.6.4.** *$\|A^*_{T^c} w\|_{\ell_\infty} < 1$ with probability greater than $1 - 6p^{-2\log 2} - 4e^{-\frac{m}{2}(\epsilon^2/2 - \epsilon^3/3)} - 2e^{-m\delta^2/2}$ for sufficiently large $n, p$.*

*Proof.* We have the following

$$\begin{aligned} \|A^*_{T^c} w\|_{\ell_\infty} &= \|A^*_{T^c} A_T (A^*_T A_T)^{-1} \mathrm{sgn} x_T\|_{\ell_\infty} \\ &= \max_{i \in T^c} |Z_i|, \end{aligned}$$

where $Z_i = \sum_{j \in T} W_{ij} \mathrm{sgn} x_j$ and $W_i = (A^*_T A_T)^{-1} A^*_T A_i$.

Recall the definition of $Z$ from Lemma 4.6.3 and consider the event

$$E_3 = \{Z \leq 1/2\} \cap \{\max_{i \in T^c} \|A^*_T A_i\|_{\ell_2} \leq \gamma\},$$

for some positive $\gamma$.

On this event we have

$$\|W_i\|_{\ell_2} \leq \|(A^*_T A_T)^{-1}\| \|A^*_T A_i\|_{\ell_2} \leq 2\gamma.$$

Note that $Z_i$ is a sum of independent random variables with zero mean because we have assumed that the signs of the entries of $x_T$ are i.i.d. symmetric variables. Also, clearly for each $j \in T$ we have $W_{ij} \mathrm{sgn} x_j \in [-|W_{ij}|, |W_{ij}|]$. Thus, applying Hoeffding's inequality we have

$$\mathbf{P}(\{|Z_i| \geq t\} \cap E_3) \leq 2e^{-\frac{t^2}{8\gamma^2}}.$$

Setting $t = 1$ and applying the union bound, we have

$$\mathbf{P}(\{\|(A^* w)_{T^c}\|_{\ell_\infty} \geq 1\} \cap E_3) \leq 2(p - s) e^{-\frac{1}{8\gamma^2}},$$

and thus

$$\begin{aligned}\mathbf{P}(\|(A^*w)_{T^c}\|_{\ell_\infty} \geq 1) &\leq \mathbf{P}(\{\|(A^*w)_{T^c}\|_{\ell_\infty} \geq 1\} \cap E_3) + \mathbf{P}(E_3^c) \\ &\leq 2(p-s)e^{-1/8\gamma^2} + \mathbf{P}(Z > 1/2) + \mathbf{P}\left(\max_{i \in T^c} \|A_T^* A_i\|_{\ell_2} > \gamma\right) \\ &\leq 2pe^{-1/8\gamma^2} + \mathbf{P}(Z > 1/2) + \mathbf{P}\left(\max_{i \in T^c} \|A_T^* A_i\|_{\ell_2} > \gamma\right).\end{aligned}$$

For the first term, if $\gamma$ satisfies

$$\gamma < \frac{1}{\sqrt{8(2\log 2 + 1)}\sqrt{\log p}},$$

then $pe^{-1/8\gamma^2} < p^{-2\log 2}$. Also, we have already shown in Lemma 4.6.3 that $\mathbf{P}(Z > 1/2) \leq 2p^{-2\log 2} + 2e^{-\frac{m}{2}(\epsilon^2/2 - \epsilon^3/3)} + e^{-m\delta^2/2}$.

Thus we need to just bound the last term. With an application of Markov's inequality similar to what is done in the proof of Lemma 2.1 we have

$$\begin{aligned}\mathbf{P}\left(\max_{i \in T^c} \|A_T^* A_i\|_{\ell_2} > \gamma\right) &\leq \mathbf{E}_\Phi\left(\mathbf{E}_T \mathbf{1}_{\max_{i \in T^c} \|A_T^* A_i\|_{\ell_2} > \gamma} \mathbf{1}_{E_1 \cap E_2}\right) + \mathbf{P}(E_1^c) + \mathbf{P}(E_2^c) \\ &\leq 2\left(\frac{\gamma_0}{\gamma}\right)^q + e^{-m\delta^2/2} + 2^{-m/2(\epsilon^2/2 - \epsilon^3/3)},\end{aligned}$$

where $q = 2\log p$ and

$$\gamma_0 = \frac{4c_0}{\sqrt{\log p}} + \frac{8(\log p)^{-\alpha - 1/2}}{1 - (\log p)^{-(\alpha+1)}} + \sqrt{\frac{c_1}{\log p}}.$$

Therefore, if $\gamma_0 < \gamma/2$ we have that the last term does not exceed $(\gamma_0/\gamma)^q \leq 2p^{-2\log 2}$. Indeed, $\gamma_0 < \gamma/2$ if

$$4c_0 + \frac{8(\log p)^{-\alpha}}{1 - (\log p)^{-(\alpha+1)}} + c_1 < \frac{1}{2\sqrt{8(2\log 2 + 1)}},$$

which is precisely one of the hypotheses of Theorem 4.2.2.

Thus the duality condition holds with probability exceeding $1 - 6p^{-2\log 2} - 4e^{-m/2(\epsilon^2/2 - \epsilon^3/3)} - 2e^{-m\delta^2/2}$.

□

## 4.7 Discussion

### 4.7.1 Contributions and relationship to prior work

The main contribution of this chapter is to show that it is possible to recover a signal from measurements when it is known that the signal is sparse in a given dictionary, and where the dictionary is not required to obey the uniform uncertainty principle but is instead required to obey a weaker incoherence condition. Results not requiring the uniform uncertainty principle or some other strict, difficult to verify condition are not very common and we mention here three other works and discuss this chapter's relationship to them.

In [25] it is shown that if a signal has a sparse representation in $\Psi = [\mathrm{I}\ \mathrm{F}]$ then by solving a linear program one can recover the sparse representation of the signal with high probability. As noted earlier, $\Psi = [\mathrm{I}\ \mathrm{F}]$ does not obey the uniform uncertainty principle. As in this chapter, is required that the sparse representation have random support and signs. This is nearly the identical setup to what we have in this chapter, only now we take *measurements* of the signal and want to recover the signal itself. (Although it is interesting to note that we end up recovering the sparse representation and from that we obtain the signal.)

The paper [15] is also related to this work. Before we detail how the two are related, we recall what was mentioned in the introduction of this chapter, namely that the assumption that the measurement matrix $\Phi$ is a Gaussian matrix is not strictly necessary. From an examination of the proof of Theorem 4.2.2, it is easy to see that what is actually required is that the matrix obeys the Johnson-Lindenstrauss requirement (see Lemma 4.6.1) and has a well-behaved top singular value. Other matrices that satisfy these requirements are random orthogonal matrices and matrices from the Bernoulli ensemble, to name a few.

Now, an application given in the introduction of [15] is to let $M$ and $\Psi$ be orthogonal matrices and let $\Phi$ be $m$ randomly sampled rows of $M$. Thus, for comparison with this chapter, we will let $\Phi$ be the first $m$ rows of a random orthogonal matrix. Letting $y = \Phi\Psi x$ where $x$ has random support and signs and then solving

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \Phi\Psi\tilde{x} = y$$

gives $x = \hat{x}$ with high probability if $m \geq C\bar{\mu}^2 s \log n$ where $\bar{\mu} = \max_{j,k} |\langle \Psi_k, \Phi_j \rangle|$ and is a rough measure of how similar $\Phi$ and $\Psi$ are.

Applying this setup to this chapter and noting that $\mu(\Psi) = 0$ and hence trivially obeys our incoherence property, and that $\|\Phi\| = 1$, we get that $\hat{x} = x$ with high probability if $m \geq Cs \log n$, where we have ignored $\delta$ and $\epsilon$. Thus if $\bar{\mu} = O(1)$, then [15] can be viewed as a special case of this chapter where the dictionary is taken to be an orthonormal matrix. Note that this is not entirely a fair comparison, as the results in [15] apply to *any* orthogonal $\Phi$, while in our comparison we have taken $\Phi$ to be a random orthogonal matrix. However, the results in this chapter are in some sense nicer, because they apply to any matrix that obeys the Johnson-Lindenstrauss requirement, and not just orthogonal matrices. (It is also of interest to note that in [4] an algorithm is given to quickly apply a matrix that satisfies the Johnson-Lindenstrauss property, which could have practical applications.)

Finally, [22] also has connections with this work. The setup is similar to that of this chapter in that $x$ is a sparse vector with random support and signs and $\Psi$ obeys our same incoherence property, but there is added noise so $y = \Psi x + z$. If the noise is set to zero, however, Theorem 1.3 of [22] basically says that $\hat{x} = x$ with high probability. Again, the main difference with this chapter is that we recover the signal from measurements $y = \Phi\Psi x$.

### 4.7.2 Future work

In this chapter we require that the signal $f$ be an exactly sparse superposition of dictionary elements. This is a clear shortcoming as usually signals are not exactly sparse in a dictionary but only compressible, meaning that there is a representation of $f$ in $\Psi$, $f = \Psi x$, such that the coefficients of $x$ decay quickly. Thus $x$ has only a few large coefficients, but it is not exactly sparse. Ideally, one would like to be able to estimate $f$ well given measurements $y = \Phi f$ where $f$ is only compressible in a dictionary $\Psi$. Note that one gives up being able to reconstruct $f$ exactly.

Also, it is unclear if incoherence is the correct condition to put on the dictionary $\Psi$ even if $f$ is an exactly sparse superposition of dictionary elements, if one is only interested in reconstructing $f$ and does not care about getting a good reconstruction of $x$. A simple example is the dictionary consisting of one element (column) $\Psi_1$

repeated $p$ times. This matrix clearly does not obey the incoherence property, but because $f = \Psi x = \Psi_1 \sum_{i=1}^{n} x_i$, when we solve

$$\min_{\tilde{x}} \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \Phi\Psi_1 \sum_{i=1}^{n} x_i = \Phi\Psi_1 \sum_{i=1}^{n} \tilde{x}_i$$

we will always get that $\hat{f} = f$, even if $\hat{x} \neq x$.

# Chapter 5

# Compressed sensing and the method of $\ell_1$-analysis

## 5.1 Abstract

This chapter deals with the problem of recovering a signal $f \in \mathbb{R}^n$ from a few noisy measurements $y \in \mathbb{R}^m$,

$$y = \Phi f + z,$$

where $m \ll n$ and $\Phi$ is a measurement matrix. We assume the signal $f$ can be well represented by a dictionary $\Psi$ and the noise $z$ satisfies $\|z\|_{\ell_2} < \epsilon$.

We show that the method of $\ell_1$-analysis is guaranteed to give good recovery, where the analysis formulation of the problem derives an estimate $\hat{f}$ for the signal $f$ from the solution to the following optimization problem

$$\min_{\tilde{f}} \|\Psi^* \tilde{f}\|_{\ell_1} \quad \text{such that} \quad \|\Phi \tilde{f} - y\|_{\ell_2} < \epsilon.$$

This can be written as a convex optimization problem and the well-established machinery of convex optimization used to solve it efficiently.

A typical result in this chapter says that if the combined matrix $\Phi\Psi$ satisfies a condition called the *uniform uncertainty principle*, also referred to as the *restricted isometry property*, then we are guaranteed good recovery, both when $f$ is an exactly sparse superposition of dictionary elements, and in the more realistic case when $f$ can only be well represented by the dictionary, meaning that $\Psi^* f$ decays quickly.

We compare our results with the more standard $\ell_1$-synthesis approach, where the

estimate of $f$ is synthesized from an estimate of weights of elements in the dictionary, and complement our study with numerical experiments. Our results are of interest because numerically $\ell_1$-analysis seems to show promise in certain scenarios, but unlike $\ell_1$-synthesis, very little is known about the method theoretically.

## 5.2 Introduction

A standard method of data acquisition and compression is to sample a signal, transform it into a basis where it has only a few large coefficients (in other words, it has a sparse representation), throw away most of the transformed coefficients, and keep only the very few large coefficients. To reconstruct the signal, one simply applies the inverse transformation, and as long as the coefficients that were thrown away were small, one gets a close approximation of the signal. However, it seems somehow wasteful to take so many signal samples and keep so few coefficients. This raises the following question. Is it possible to combine the sampling and compression steps, and instead just take very few "compressed samples"? This is the subject of a growing field, known as *compressed sensing* or *compressive sampling* [40, 18].

To be more precise, in this chapter we are interested in estimating a signal $f \in \mathbb{R}^n$ from a few linear measurements $y \in \mathbb{R}^m$ which may or may not be noisy. In other words, $y = \Phi f$ or $y = \Phi f + z$, where $\Phi \in \mathbb{R}^{m \times n}$ is a measurement matrix, $m \ll n$, and $z \in \mathbb{R}^m$ is a noise vector such that $\|z\|_{\ell_2} < \epsilon$. From these measurements, we would like to find a good reconstruction of the signal. (Of course we are only interested in reconstructing the signal in a computationally feasible way!)

It is standard in compressed sensing scenarios to assume that the signal $f$ can be sparsely represented in an orthonormal basis, meaning that it has an exactly sparse representation, or perhaps more realistically, has coefficients that quickly decay in the basis. This is reasonable if the signal is very smooth (in which case Fourier is a good choice of basis) or is piecewise continuous (in which case wavelets are a good choice of basis [65]). However, for larger classes of more complicated signals with combinations of features, frequently no one basis is suitable. For example, the Fourier basis is good at representing sinusoids, while the Dirac basis is good at representing impulses. However, neither basis is good at representing signals that

contain both sinusoids and impulses.

Thus, in this chapter we will not assume signals can be well represented in a basis, but instead will focus on overcomplete signal representations, known as dictionaries. A dictionary is a collection of elements, also known as atoms, that is full rank. Because they are overcomplete, dictionaries allow more flexibility and richness in signal representations and are able to well represent larger, more interesting classes of signals. Also, for a given application, there is often a dictionary that arises naturally from the problem. For example, reflected radar signals are often a train of pulses of different widths at different frequencies. In this case the Gabor dictionary, whose elements are an overcomplete collection of translated windowed sinusoids where the windows have various widths and the sinusoids have various frequencies, is a natural choice of dictionary.

For the purposes of this chapter, we will assume we have a tight-frame dictionary $\Psi \in \mathbb{R}^{n\times p}$, $p \gg n$, with frame bound 1. Thus $\Psi\Psi^* = I$. (Note that obviously $\Psi^*\Psi \neq I$.) Many dictionaries of interest are tight frames or can be designed to be a tight frame, for example, curvelets, the Gabor dictionary, wavelet frames, the overcomplete discrete cosine transform, etc. However, unlike in a basis, signals no longer have a unique representation in a dictionary. There are infinitely many $\alpha$ such that $f = \Psi\alpha$, and this adds another layer of complexity to the problem.

### 5.2.1 The method of $\ell_1$-synthesis

Our challenge, then, is to reconstruct a signal that can be well represented by a dictionary, from measurements of the signal. The standard way of approaching this problem is known as *synthesis*. From the measurements, one first determines a "good" set of dictionary coefficients, $\hat{\alpha}$, and then builds the signal as a superposition of these atoms, $\hat{f} = \Psi\hat{\alpha}$. In other words, the signal is synthesized from atoms in the dictionary.

In the noiseless case, $y = \Phi\Psi\alpha$, a popular and successful method of selecting the set of dictionary coefficients is given by the solution of the following linear program

$$\min_{\tilde{\alpha}} \|\tilde{\alpha}\|_{\ell_1} \quad \text{such that} \quad \Phi\Psi\tilde{\alpha} = y, \tag{5.1}$$

while in the noisy case, $y = \Phi\Psi\alpha + z$ where $\|z\|_{\ell_2} < \epsilon$, the estimate is given by the following second-order cone program

$$\min_{\tilde{\alpha}} \|\tilde{\alpha}\|_{\ell_1} \quad \text{such that} \quad \|\Phi\Psi\tilde{\alpha} - y\|_{\ell_2} < \epsilon. \tag{5.2}$$

Both of these optimization programs are convex, and can be solved using the standard tools of convex optimization [10]. These synthesis methods work by achieving an estimate $\hat{\alpha}$ which leads to a good estimate $\hat{f} = \Psi\hat{\alpha}$.

There are many known results for the $\ell_1$-synthesis setup, but we pause here and mention three in particular. In [27] it is shown that if $\Phi\Psi$ obeys a condition called the *uniform uncertainty principle*, also known as the *restricted isometry property*, then $\hat{\alpha}$ is a good estimator of $\alpha$. The uniform uncertainty principle requires that the restricted isometry constants of $A = \Phi\Psi$ be sufficiently small, where the restricted isometry constants of a matrix are defined as follows.

**Definition 5.2.1.** *For each integer $s = 1, 2, \ldots$, define the restricted isometry constant $\delta_s$ of a matrix $A$ as the smallest number such that*

$$(1 - \delta_s)\|x\|_{\ell_2}^2 \leq \|Ax\|_{\ell_2}^2 \leq (1 + \delta_s)\|x\|_{\ell_2}^2$$

*holds for all $s$-sparse vectors $x$. A vector is said to be $s$-sparse if it has at most $s$ nonzero entries.*

These restricted isometry constants are named as such because for every $s$, $\delta_s$ essentially measures how far from an isometry $A$ is, acting only on $s$-sparse vectors.

What is proved in [27] and slightly refined in [19] is the following theorem.

**Theorem 5.2.2 ([27, 19]).** *Assume $\delta_{2s}(\Phi\Psi) < \sqrt{2} - 1$. Let $\alpha_s$ be the truncated vector corresponding to the $s$ largest (in absolute value) coefficients of $\alpha$. Then the solution $\hat{\alpha}$ to* (5.2) *obeys*

$$\|\alpha - \hat{\alpha}\|_{\ell_2} \leq C_0 s^{-1/2}\|\alpha - \alpha_s\|_{\ell_1} + C_1\epsilon$$

*where the constants $C_0$ and $C_1$ are given in terms of restricted isometry constants.*

Because $\Psi$ is a tight frame with frame bound 1 we have

$$\|f - \hat{f}\|_{\ell_2} = \|\Psi(\alpha - \hat{\alpha})\|_{\ell_2} = \|\alpha - \hat{\alpha}\|_{\ell_2},$$

and so the theorem relates the estimation error of $f$ to the $s$-term approximation error of $\alpha$, plus a noise term.

This result is appealing because it handles the case where $\alpha$ is exactly sparse and also when $\alpha$ has decaying coefficients. In addition, the measurements may or may not be noisy. (If they are not noisy, then $\epsilon = 0$ and (5.2) reduces to (5.1).) The most obvious drawback of the result is that it requires the combined measurement matrix/dictionary to obey a restricted isometry property. (We refer to any condition that requires the restricted isometry constants to be sufficiently small as a restricted isometry property or uniform uncertainty principle.) It is possible that for measurements and dictionaries of interest this condition will not hold.

A result that addresses this shortcoming is Theorem 4.2.2 of Chapter 4 of this thesis. Theorem 4.2.2 assumes that the measurement matrix obeys a restricted isometry property, while the dictionary must only obey a weaker coherence property, which essentially says that the columns of the dictionary can not be too colinear. In compressed sensing setups like we have discussed, this is a bit more realistic as the dictionary is usually determined by the types of signals in which one is interested, while one has a bit more freedom in the design of the measurements. Unfortunately, the result in Chapter 4 requires that $\alpha$ be exactly sparse. This is a bit unrealistic as many signals of interest will probably not be an exact superposition of just a few dictionary elements. It is more reasonable to expect that signals of interest can be well-approximated by just a few dictionary elements, so that there is a representation of the signal with $\alpha$ quickly decaying, but this situation is not dealt with in that result.

Finally, we mention a result that assumes the joint measurement/dictionary matrix $\Phi\Psi$ obeys a coherence property, namely Theorem 1.3 of [22]. It is a nice result in that neither the measurements nor the dictionary must obey a restricted isometry property, but it also suffers from the requirement that $\alpha$ be exactly sparse, and even adds the requirement that the nonzero coefficients of $\alpha$ be sufficiently large.

### 5.2.2 The method of $\ell_1$-analysis

All of the synthesis results we have mentioned rely on finding a good estimate of $\alpha$, which in turn leads to a good estimate of $f$. In some sense, however, it seems strange to estimate $\alpha$ at all, when what we would really like is to estimate $f$. Moreover, it is possible that there are times when a good estimate of $\alpha$ does not exist, but good estimates of $f$ do exist. For example, if two atoms in the dictionary are identical and equal to $f$, then, forgetting the measurements, it is impossible to estimate $\alpha$ well—we do not know which atom formed $f$. But obviously the dictionary is still able to well represent $f$. This leads to the question, if we are interested in $f$, why bother estimating $\alpha$ in the first place?

To address this, we introduce the $\ell_1$-*analysis* formulation of the problem. In the noiseless case, $y = \Phi f$, the estimate $\hat{f}$ is given by the minimizer of the following linear program

$$\min_{\tilde{f}} \|\Psi^* \tilde{f}\|_{\ell_1} \quad \text{such that} \quad \Phi \tilde{f} = y, \tag{5.3}$$

while in the noisy case, $y = \Phi f + z$ where $\|z\|_{\ell_2} < \epsilon$, the estimate $\hat{f}$ is given by

$$\min_{\tilde{f}} \|\Psi^* \tilde{f}\|_{\ell_1} \quad \text{such that} \quad \|\Phi \tilde{f} - y\|_{\ell_2} < \epsilon. \tag{5.4}$$

Note that in this case the estimator for $f$ is acquired directly as the minimum of a convex optimization program. Instead of being synthesized from atoms in the dictionary, analysis associates each signal $\tilde{f}$ with a vector of coefficients $\Psi^* \tilde{f}$. It is the $\ell_1$ norm of this vector that analysis minimizes. Heuristically, this seems like a reasonable strategy, as $\ell_1$ is known to be sparsity promoting [31] and $\Psi^* \tilde{f}$ is correlating the candidate signal $\tilde{f}$ with the elements of the dictionary. Since we are assuming $f$ can be well represented by the dictionary, it makes sense that $\Psi^* f$ would be "sparse."

It is interesting to note that analysis and synthesis are actually closely related. In fact, because any function $f$ can be written as $f = \Psi\alpha$ for some $\alpha$, we can rewrite (5.4) as

$$\min_{\tilde{\alpha}} \|P\tilde{\alpha}\|_{\ell_1} \quad \text{such that} \quad \|\Phi\Psi\tilde{\alpha} - y\|_{\ell_2} < \epsilon$$

where $P$ is the projector $\Psi^*\Psi$. This is very similar in form to (5.2), only instead of

minimizing over all vectors in $\mathbb{R}^p$, the minimum is taken over $p$-dimensional vectors projected into an $n$-dimensional subspace determined by the dictionary $\Psi$.

### 5.2.3 Statement of results

Besides avoiding estimating $\alpha$, there are other compelling reasons for considering the analysis method. In addition to involving fewer unknowns than $\ell_1$-synthesis, numerical simulations show $\ell_1$-analysis works well, and in certain situations outperforms synthesis. For example, in [50], when the dictionary is an overcomplete discrete cosine transform and the signals are standard test images, analysis performs better than synthesis, and the improvement increases as the redundancy of the dictionary increases. Also, in [17] when the dictionary is a Gabor tight frame and the signals are pulses, analysis seems to slightly outperform synthesis. When $\ell_1$ reweighting is added, the increase in performance becomes more dramatic. See also [79, 80] for other analysis-based numerical experiments.

As far as we know, however, no results exist giving conditions on the measurements or dictionary that guarantee good recovery of $f$. (Although [50] does give theoretical results showing differences between the analysis and synthesis methods.) The contribution of this chapter is to make a first modest step at addressing this gap. Our results are of interest because, while the synthesis setup has been extensively studied, the analysis setup is not very well understood and seems to show promise in certain scenarios.

More specifically, we show that if the combined matrix $\Phi\Psi$ obeys a uniform uncertainty principle, then by solving (5.3) or (5.4), one gets a good estimate of $f$. We are also able to get similar, slightly stronger, results if we assume a particular form for the measurement matrix $\Phi$, that takes the dictionary into account.

We have the following theorem:

**Theorem 5.2.3.** *Assume $\delta_{2s}(\Phi\Psi) < \sqrt{2} - 1$. Let $(\Psi^* f)_s$ be the truncated vector corresponding to the $s$ largest (in absolute value) coefficients of $\Psi^* f$. Then the solution $\hat{f}$ to* (5.4) *obeys*

$$\|f - \hat{f}\|_{\ell_2} \leq C_0 s^{-1/2} \|\Psi^* f - (\Psi^* f)_s\|_{\ell_1} + C_1 \epsilon$$

*with constants $C_0$ and $C_1$ given explicitly in terms of restricted isometry constants in the proof.*

Theorem 5.2.3 relates the approximation error in $f$ to the best $s$-term approximation error of $\Psi^* f$, measured in the $\ell_1$ norm, plus a noise term, when $\hat{f}$ is the solution to $\ell_1$-analysis.

If something is known about $\Psi^* f$, either that it is sparse or compressible (meaning that its reordered coefficients quickly decay), we have the following corollaries.

**Corollary 5.2.4.** *Under the same conditions as Theorem 5.2.3, if $\Psi^* f$ is exactly $s$-sparse, we have*

$$\|f - \hat{f}\|_{\ell_2} \leq C\epsilon.$$

*Moreover, if the measurements of $f$ are noiseless, we recover $f$ exactly.*

The proof of the corollary is immediate, as if $\Psi^* f$ is exactly $s$-sparse, then $\Psi^* f = (\Psi^* f)_s$, and if the measurements are noiseless, $\epsilon = 0$. If $\Psi^* f$ is not exactly sparse, but instead has decaying coefficients, we have the following corollary.

**Corollary 5.2.5.** *Under the same conditions as Theorem 5.2.3, if the $k$th largest coefficient in absolute value of $\Psi^* f$, $|\Psi^* f|_{(k)}$, satisfies*

$$|\Psi^* f|_{(k)} \leq C_r \cdot k^{-r}$$

*for $r \geq 1$, then the solution to (5.4) satisfies*

$$\|f - \hat{f}\|_{\ell_2} \leq C_0 s^{-r+1/2} + C_1 \epsilon$$

*for constants $C_0$ and $C_1$ in terms of restricted isometry constants and $r$.*

This follows from Theorem 5.2.3 because, approximating the sum $\|\Psi^* f - (\Psi^* f)_s\|_{\ell_1}$ as an integral and integrating, gives

$$\frac{\|\Psi^* f - (\Psi^* f)_s\|_{\ell_1}}{\sqrt{s}} \leq C_r \cdot s^{-r+1/2}.$$

Similarly, we have $\|\Psi^* f - (\Psi^* f)_s\|_{\ell_2} \leq C_r' \cdot s^{-r+1/2}$. Thus the result says that our approximation error is almost the same as the approximation error made by keeping

the $s$ largest coefficients of $\Psi^* f$. The faster the entries of $\Psi^* f$ decay, the better our approximation.

These results should be compared with the synthesis result, Theorem 5.2.2, which relates the approximation error of $f$ to the decay of the coefficients of $\alpha$. We have already mentioned earlier that an exactly sparse $\alpha$ is not very realistic in applications because signals in general will not be an exact superposition of a handful of elements in the dictionary. It is reasonable, however, to assume that $\alpha$ decays; after all, we are assuming that the dictionary is able to well represent signals of interest. It is also reasonable to assume that the coefficients of $\Psi^* f$ decay. This is, again, because our signals can be well represented by the dictionary. $\Psi^* f$ takes the atoms of the dictionary and correlates them with the signal. It is reasonable to assume that a few atoms are highly correlated with the signal, while the rest are only weakly correlated.

We would also like to point out that for certain applications it might even be reasonable to assume that $\Psi^* f$ is exactly sparse. While it might not be that the signal is exactly a superposition of dictionary elements, it might very well be that only relatively few dictionary elements have a non-zero correlation with the signal—think of the Gabor dictionary with a signal that is a fairly narrow pulse that looks similar to an element in the dictionary, but is not the exact element. Each dictionary element has a finite support, and only elements whose support intersects the support of the pulse signal will have a non-zero inner product, while the rest of the correlations will be zero.

There is of course a relation between the coefficient decay of $\alpha$ and $\Psi^* f$. In fact, if $\Psi^* f$ decays, then there exists an $\alpha$ that also decays. (Take $\alpha = \Psi^* f$. Then $\Psi\alpha = \Psi\Psi^* f = f$.) The reverse is not necessarily true, however. If $\alpha$ quickly decays, then there is no guarantee that $\Psi^* f = \Psi^*\Psi\alpha = P\alpha$ also decays, although of course it is possible for this to be true in certain cases.

Finally, we mention that our result suffers from the same drawback as Theorem 5.2.2 in that it requires that the combined measurement/dictionary matrix $\Phi\Psi$ obeys the uniform uncertainty principle, which might not hold for measurements and dictionaries of interest. As an attempt to address this shortcoming, we are also able to prove a similar result to Theorem 5.2.3 if we assume a particular form of the

measurement matrix $\Phi$, namely that $\Phi = A\Psi^*$ where only $A$ is required to obey the uniform uncertainty principle.

**Theorem 5.2.6.** *Assume $\Phi = A\Psi^*$, and $\delta_{2s}(A) < \sqrt{2} - 1$. Let $(\Psi^* f)_s$ be the truncated vector corresponding to the $s$ largest (in absolute value) coefficients of $\Psi^* f$. Then the solution $\hat{f}$ to (5.4) obeys*

$$\|f - \hat{f}\|_{\ell_2} \leq C_0 s^{-1/2} \|\Psi^* f - (\Psi^* f)_s\|_{\ell_1} + C_1 \epsilon$$

*with constants $C_0, C_1$ given explicitly in the proof.*

Note that Theorem 5.2.6 is not just a specific case of Theorem 5.2.3 because if we simply let $\Phi = A\Psi^*$ in Theorem 5.2.3 we require that $\delta_{2s}(AP) < \sqrt{2} - 1$ where $P$ is the projector $\Psi^*\Psi$, as opposed to requiring that $\delta_{2s}(A) < \sqrt{2} - 1$.

An nice feature of this second result is that there is no condition at all on the dictionary $\Psi$. Our method works even if $\Psi$ has extremely highly correlated columns, or even identical columns. Also, if $\Psi$ can be quickly applied to vectors, then $\Phi = A\Psi^* = (\Psi A^*)^*$ can be built efficiently.

## 5.3 Numerical Experiments

In order to further explore our theoretical results, we performed the following numerical experiments. We took as our dictionary the Gabor dictionary. This dictionary consists of windowed sinusoids where the windows have dyadic widths and the sinusoids varying frequencies. In other words the elements look like

$$\psi_{j,k,l}(t) = w\left(\frac{t - u_k}{2^j}\right) \sin\left(2\pi\omega_l(t - u_k)\right).$$

The window $w(t)$ is taken to be an iterated sine function so it forms a partition of unity (and hence the dictionary is a tight frame), while the translations $u_k$ and frequencies $\omega_l$ are three and two times redundant at each scale, respectively.

Thus the elements of the Gabor dictionary look like pulses of varying widths and frequencies and we expect that this dictionary will be good at representing trains of pluses. The test signals we considered are shown in Figure 5.1. We emphasize that

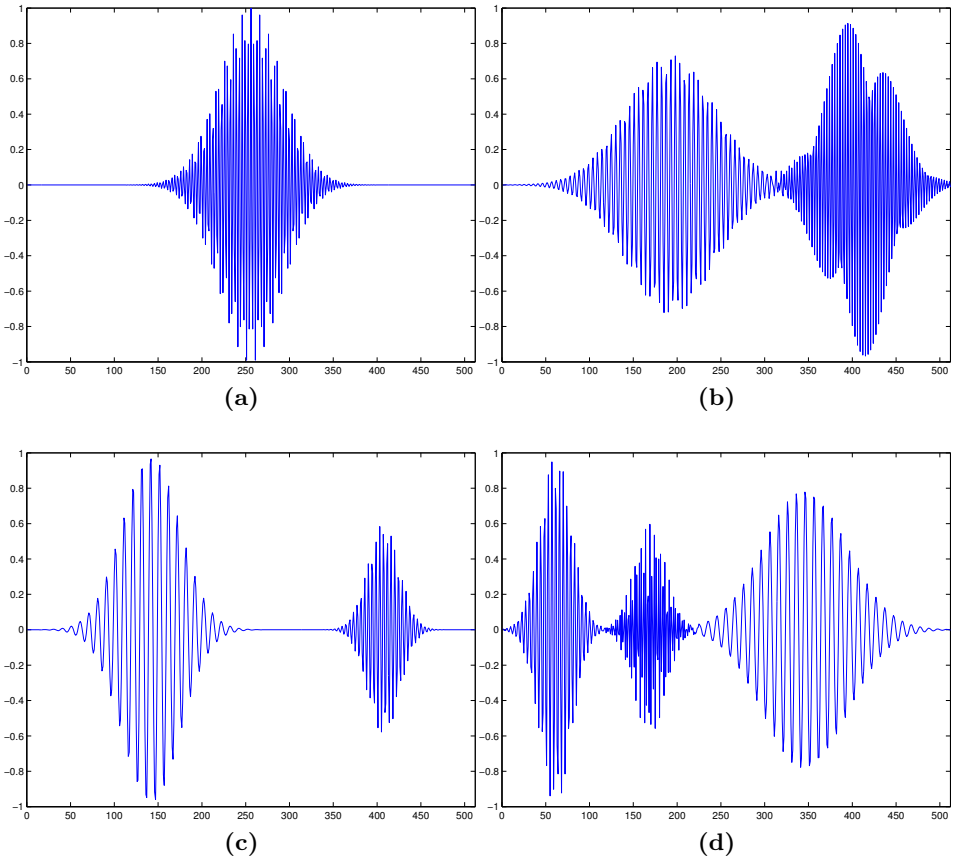



**Figure 5.1.** Four sample test pulses. We expect that the Gabor dictionary will be able to well represent these signals. We note, however, that none of the signals is an exact element of the dictionary, or the sum of a few elements.

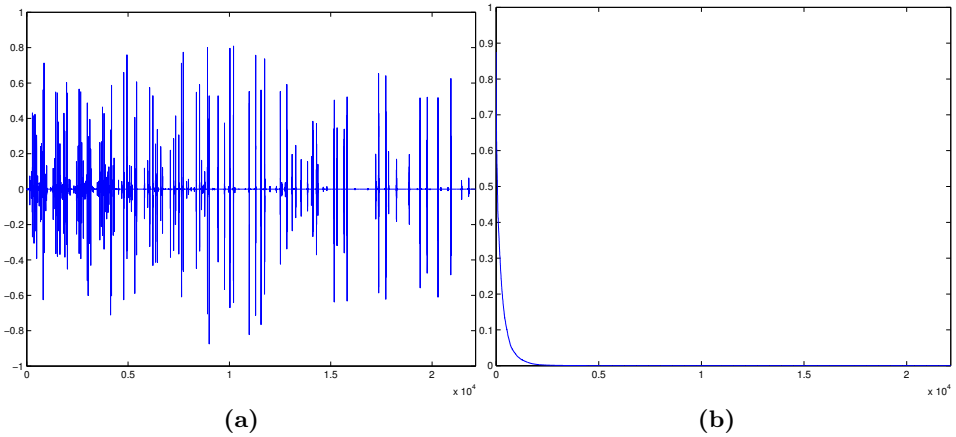
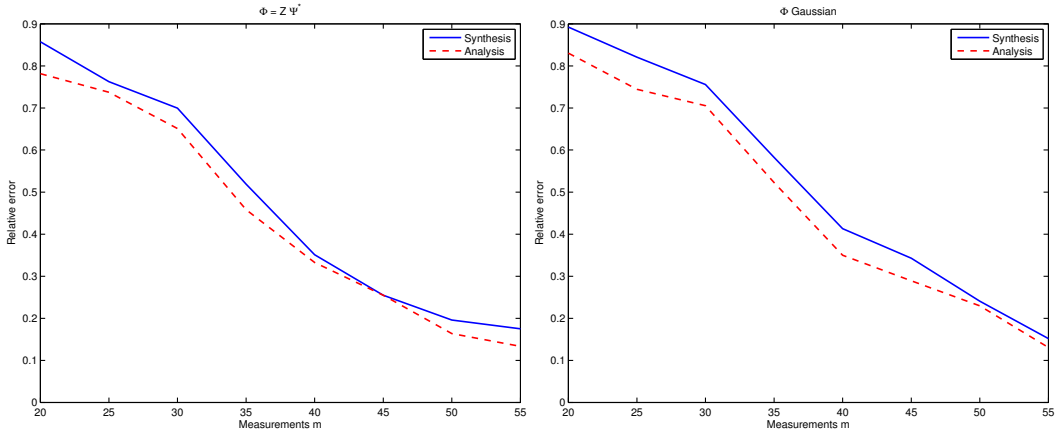



**Figure 5.2.** The coefficients of $\Psi^* f$, when $f$ is the test pulse shown in Figure 5.1b. In 5.2a we see the coefficients, and in 5.2b we see the coefficients sorted by absolute value. All of the test signals in Figure 5.1 showed similar fast decay of $\Psi^* f$.




**(a)** **(b)**

**Figure 5.3.** Numerical experiment results. 5.3a shows the averaged relative error of 20 experiments for various numbers of measurements $m$ when the measurement matrix $\Phi = Z\Psi^*$ and $Z$ is Gaussian. The blue solid line shows the results for $\ell_1$-synthesis, while the broken red line shows the results of $\ell_1$-analysis. 5.3b shows the same, only $\Phi$ is simply taken to be a Gaussian matrix.

these test signals are not elements of the dictionary or exactly sparse combinations of elements of the dictionary. Figure 5.2 shows that indeed $\Psi^* f$ has quickly decaying coefficients.

We considered the noiseless case $y = \Phi f$, and let the measurement matrix $\Phi$ be either a matrix with i.i.d. $N(0,1)$ entries, or $\Phi = Z\Psi^*$ where $Z$ is a matrix with i.i.d. $N(0,1)$ entries. (This second measurement matrix satisfies the conditions of Theorem 5.2.6 with high probability.) We then solved both $\ell_1$ analysis (5.3) and $\ell_1$ synthesis (5.1) using optimization code which can be found at [1], formed the estimate $\hat{f}$, and calculated the relative error

$$\text{err}_{\text{rel}} = \frac{\|\hat{f} - f\|_{\ell_2}}{\|f\|_{\ell_2}}.$$

We did this for various number of measurements $m$, and for each $m$ we repeated the experiment 20 times and averaged the result. Our results are presented in Figure 5.3. The results we show are for the test signal Figure 5.1b, although all the test signals we tried showed similar behavior.

We see that analysis slightly outperforms synthesis, and that surprisingly both methods do about the same when the measurements are simply Gaussian versus

when they take into account the dictionary. This is particularly interesting because in this case $\Phi\Psi$ certainly does not obey the uniform uncertainty principle, and suggests that our results should be able to be extended.

## 5.4 Proofs

### 5.4.1 Proof of Theorem 5.2.3

Note that the proof of Theorem 5.2.3 is basically contained in [19] which in turn is based on a proof in [28]. However, it is interesting to note that in (5.4) we are minimizing the $\ell_1$ norm over an $n$-dimensional subspace of $\mathbb{R}^p$ (vectors in the orthogonal complement of the range of $\Psi$) instead of all of $\mathbb{R}^p$, yet because $f$ and $\hat{f}$ live in that subspace, we are still able to follow the methods of the proof. We include the proof here for completeness.

Before proving the theorem, however, we start with some notation. Throughout the chapter, $x_T$ is a vector equal to $x$ on an index set $T$ and zero elsewhere. Set $h = \hat{f} - f$ and decompose $\Psi^* h$ into a sum of vectors $(\Psi^* h)_{T_0}, (\Psi^* h)_{T_1}, (\Psi^* h)_{T_2}, \ldots$, where $T_0$ corresponds to the largest (in absolute value) $s$ coefficients of $\Psi^* f$, $T_1$ corresponds to the largest (in absolute value) $s$ coefficients of $\Psi^* h$ in $T_0^c$, $T_2$ corresponds to the next largest (in absolute value) $s$ coefficients of $\Psi^* h$ in $T_0^c$, etc. Set $T_{01} = T_0 \cup T_1$.

In the proof of the theorem we will twice make use of the following lemma.

**Lemma 5.4.1.** *We have the following*

$$\sum_{j \geq 2} \|(\Psi^* h)_{T_j}\|_{\ell_2} \leq \|(\Psi^* h)_{T_{01}}\|_{\ell_2} + 2s^{-1/2} \|(\Psi^* f)_{T_0^c}\|_{\ell_1}.$$

*Proof.* Because $\hat{f}$ is a minimum of $(P_1)$ and $f$ is feasible, we must have

$$\begin{aligned}
\|\Psi^* f\|_{\ell_1} &\geq \|\Psi^* \hat{f}\|_{\ell_1} \\
&= \|\Psi^*(f+h)\|_{\ell_1} = \sum_{i \in T_0} |(\Psi^* f)_i + (\Psi^* h)_i| + \sum_{i \in T_0^c} |(\Psi^* f)_i + (\Psi^* h)_i| \\
&\geq \|(\Psi^* f)_{T_0}\|_{\ell_1} - \|(\Psi^* h)_{T_0}\|_{\ell_1} + \|(\Psi^* h)_{T_0^c}\|_{\ell_1} - \|(\Psi^* f)_{T_0^c}\|_{\ell_1}
\end{aligned}$$

This gives

$$\|(\Psi^* h)_{T_0^c}\|_{\ell_1} \leq \|(\Psi^* h)_{T_0}\|_{\ell_1} + 2\|(\Psi^* f)_{T_0^c}\|_{\ell_1}. \tag{5.5}$$

Also, because of how the $(\Psi^* h)_{T_j}$ are defined, we have the following for $j \geq 2$,

$$\|(\Psi^* h)_{T_j}\|_{\ell_2} \leq s^{-1/2}\|(\Psi^* h)_{T_{j-1}}\|_{\ell_1}.$$

Thus we have

$$\sum_{j\geq 2} \|(\Psi^* h)_{T_j}\|_{\ell_2} \leq s^{-1/2} \sum_{j\geq 2} \|(\Psi^* h)_{T_{j-1}}\|_{\ell_1} = s^{-1/2}\|(\Psi^* h)_{T_0^c}\|_{\ell_1}. \tag{5.6}$$

Combining (5.5) and (5.6) and using

$$\|(\Psi^* h)_{T_0}\|_{\ell_1} \leq \sqrt{s}\|(\Psi^* h)_{T_0}\|_{\ell_2} \leq \sqrt{s}\|(\Psi^* h)_{T_{01}}\|_{\ell_2}$$

gives the lemma.

□

Now to prove the theorem. Note the following

$$\|f - \hat{f}\|_{\ell_2} = \|h\|_{\ell_2} = \|\Psi^* h\|_{\ell_2} \leq \|(\Psi^* h)_{T_{01}}\|_{\ell_2} + \|(\Psi^* h)_{T_{01^c}}\|_{\ell_2}, \tag{5.7}$$

where we have used the fact that $\Psi\Psi^* = I$.

We will prove the theorem by first bounding $\|(\Psi^* h)_{T_{01}^c}\|_{\ell_2}$ by $\|(\Psi^* h)_{T_{01}}\|_{\ell_2}$ and then bounding $\|(\Psi^* h)_{T_{01}}\|_{\ell_2}$.

The bound on $\|(\Psi^* h)_{T_{01}^c}\|_{\ell_2}$ basically follows from Lemma 5.4.1:

$$\begin{aligned} \|(\Psi^* h)_{T_{01^c}}\|_{\ell_2} &= \|\sum_{j\geq 2} (\Psi^* h)_{T_j}\|_{\ell_2} \\ &\leq \sum_{j\geq 2} \|(\Psi^* h)_{T_j}\|_{\ell_2} \\ &\leq \|(\Psi^* h)_{T_{01}}\|_{\ell_2} + 2s^{-1/2}\|(\Psi^* f)_{T_0^c}\|_{\ell_1}. \end{aligned} \tag{5.8}$$

Now we turn to bounding $\|(\Psi^* h)_{T_{01}}\|_{\ell_2}$. We will use the following relation

$$\|\Phi h\|_{\ell_2} = \|\Phi(\hat{f} - f)\|_{\ell_2} \leq \|\Phi\hat{f} - y\|_{\ell_2} + \|y - \Phi f\|_{\ell_2} \leq 2\epsilon, \tag{5.9}$$

which follows from the triangle inequality and the fact that $f$ and $\hat{f}$ are feasible, as well as the following lemma which is proved in [19].

**Lemma 5.4.2 (Lemma 2.1 of [19]).** *We have*

$$|\langle Xx, Xx'\rangle| \leq \delta_{s+s'}(X)\|x\|_{\ell_2}\|x'\|_{\ell_2}$$

*for all $x, x'$ supported on disjoint subsets $T, T'$ with $|T| \leq s, |T'| \leq s'$.*

We have

$$\begin{aligned}(1-\delta_{2s}(\Phi\Psi))\|(\Psi^*h)_{T_{01}}\|_{\ell_2}^2 &\leq \|\Phi\Psi(\Psi^*h)_{T_{01}}\|_{\ell_2}^2 \\ &= \langle \Phi\Psi(\Psi^*h)_{T_{01}}, \Phi h\rangle - \langle \Phi\Psi(\Psi^*h)_{T_{01}}, \sum_{j\geq 2}\Phi\Psi(\Psi^*h)_{T_j}\rangle \\ &\leq |\langle \Phi\Psi(\Psi^*h)_{T_{01}}, \Phi h\rangle| + \sum_{j\geq 2}(|\langle \Phi\Psi(\Psi^*h)_{T_0}, \Phi\Psi(\Psi^*h)_{T_j}\rangle| \\ &\quad + |\langle \Phi\Psi(\Psi^*h)_{T_1}, \Phi\Psi(\Psi^*h)_{T_j}\rangle|).\end{aligned} \tag{5.10}$$

Using Cauchy-Schwarz, the restricted isometry property, and (5.9) we get the following bound on the first term

$$\begin{aligned}|\langle \Phi\Psi(\Psi^*h)_{T_{01}}, \Phi h\rangle| &\leq \|\Phi\Psi(\Psi^*h)_{T_{01}}\|_{\ell_2}\|\Phi h\|_{\ell_2} \\ &\leq 2\epsilon\sqrt{1+\delta_{2s}(\Phi\Psi)}\|(\Psi^*h)_{T_{01}}\|_{\ell_2}.\end{aligned} \tag{5.11}$$

By Lemma 5.4.2 we have

$$|\langle \Phi\Psi(\Psi^*h)_{T_0}, \Phi\Psi(\Psi^*h)_{T_j}\rangle| \leq \delta_{2s}(\Phi\Psi)\|(\Psi^*h)_{T_0}\|_{\ell_2}\|(\Psi^*h)_{T_j}\|_{\ell_2}, \tag{5.12}$$

and a similar bound replacing $T_0$ with $T_1$. Using (5.12), $\|(\Psi^*h)_{T_0}\|_{\ell_2} + \|(\Psi^*h)_{T_1}\|_{\ell_2} \leq \sqrt{2}\|(\Psi^*h)_{T_{01}}\|_{\ell_2}$, and Lemma 2.1 we get the following bound on the second term

$$\begin{aligned}&\sum_{j\geq 2}(|\langle \Phi\Psi(\Psi^*h)_{T_0}, \Phi\Psi(\Psi^*h)_{T_j}\rangle| + |\langle \Phi\Psi(\Psi^*h)_{T_1}, \Phi\Psi(\Psi^*h)_{T_j}\rangle|) \\ &\quad \leq \delta_{2s}(\Phi\Psi)\sqrt{2}\|(\Psi^*h)_{T_{01}}\|_{\ell_2}(\|(\Psi^*h)_{T_{01}}\|_{\ell_2} + 2s^{-1/2}\|(\Psi^*f)_{T_0^c}\|_{\ell_1}).\end{aligned} \tag{5.13}$$

Combining (5.10), (5.11), and (5.13), and rearranging terms we get the following bound on $\|(\Psi^* h)_{T_{01}}\|_{\ell_2}$

$$\|(\Psi^* h)_{T_{01}}\|_{\ell_2} \leq (1-\rho)^{-1}(\alpha\epsilon + 2\rho s^{-1/2}\|(\Psi^* f)_{T_0^c}\|_{\ell_1}) \tag{5.14}$$

where

$$\alpha \equiv \frac{2\sqrt{1+\delta_{2s}(\Phi\Psi)}}{1-\delta_{2s}(\Phi\Psi)}, \quad \rho \equiv \frac{\sqrt{2}\delta_{2s}(\Phi\Psi)}{1-\delta_{2s}(\Phi\Psi)}.$$

Note that when rearranging terms we divide by $1-\rho$ and thus require that $1-\rho$ be positive, which it is since $\delta_{2s}(\Phi\Psi) < \sqrt{2}-1$ by the hypotheses of the theorem.

Finally, the conclusion of the theorem follows from (5.7), (5.8) and (5.14) and we have

$$\|f - \hat{f}\|_{\ell_2} \leq \frac{2}{1-\rho} s^{-1/2}\|(\Psi^* f)_{T_0^c}\|_{\ell_1} + \frac{2\alpha}{1-\rho}\epsilon. \tag{5.15}$$

□

### 5.4.2 Proof of Theorem 5.2.6

The proof of 5.2.6 is very similar to the proof of 5.2.3 and we only outline here the main differences. Instead of (5.10) we have

$$\begin{aligned}
(1-\delta_{2s}(A))\|(\Psi^* h)_{T_{01}}\|_{\ell_2}^2 &\leq \|A(\Psi^* h)_{T_{01}}\|_{\ell_2}^2 \\
&= \langle A(\Psi^* h)_{T_{01}}, A\Psi^* h\rangle - \langle A(\Psi^* h)_{T_{01}}, \sum_{j\geq 2} A(\Psi^* h)_{T_j}\rangle \\
&\leq |\langle A(\Psi^* h)_{T_{01}}, \Phi h\rangle| \\
&\quad + \sum_{j\geq 2}(|\langle A(\Psi^* h)_{T_0}, A(\Psi^* h)_{T_j}\rangle| + |\langle A(\Psi^* h)_{T_1}, A(\Psi^* h)_{T_j}\rangle|)
\end{aligned}$$

where we have used the fact that $A\Psi^* = \Phi$ in going from the equality to the second inequality. This is exactly the same form as (5.10) with every instance of $\Phi\Psi$ replaced by $A$. Thus the rest of the proof goes through exactly the same as the proof of Theorem 5.2.3, with $\Phi\Psi$ replaced by $A$. In particular, we now have

$$\alpha \equiv \frac{2\sqrt{1+\delta_{2s}(A)}}{1-\delta_{2s}(A)}, \quad \rho \equiv \frac{\sqrt{2}\delta_{2s}(A)}{1-\delta_{2s}(A)}$$

and need only $\delta_{2s}(A) < \sqrt{2}-1$ so that $1-\rho$ is positive. □

# Bibliography

[1] $\ell_1$ Magic. `http://www.acm.caltech.edu/l1magic`.

[2] The USC-SIPI Image Database. `http://sipi.usc.edu/database/index.html`.

[3] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. System Sci.*, 66(4):671–687, 2003. Special issue on PODS 2001 (Santa Barbara, CA).

[4] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 557–563. ACM, New York, 2006.

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

[6] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[7] A. Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[8] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. Accepted, *Annals of Statistics*, 2007.

[9] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[10] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

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

[12] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007.

[13] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[14] E. Candès and J. Romberg. Practical signal recovery from random projections. In *SPIE International Sumposium on Electoric Imaging: Computational Imaging III*, 2005.

[15] E. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[16] E. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 295–308, 2005.

[17] E. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l 1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[18] E. J. Candès. Compressive sampling. In *International Congress of Mathematicians. Vol. III*, pages 1433–1452. Eur. Math. Soc., Zürich, 2006.

[19] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

[20] E. J. Candès and D. L. Donoho. Curvelets-a surprisingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, 1999.

[21] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise $C^2$ singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.

[22] E. J. Candès and Y. Plan. Near-ideal model selection by $\ell_1$ minimization. Accepted, *Annals of Statistics*, 2007.

[23] E. J. Candès and P. A. Randall. Error correction and convex programming. *Proc. Appl. Math. Mech.*, 7(1):2070001–2070002, 2007.

[24] E. J. Candès and P. A. Randall. Highly robust error correction by convex programming. *IEEE Trans. Inform. Theory*, 54(7):2829–2840, 2008.

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

[26] E. J. 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.

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

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

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

[30] E. J. Candès and T. Tao. The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *Ann. Statist.*, 35(6):2313–2351, 2007.

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

[32] J. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826, 1973.

[33] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and the best $k$-term approximation. manuscript, 2006.

[34] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001.

[35] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[36] G. M. Davis, S. G. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Engineering*, 33(7):2183–2191, 1994.

[37] R. DeVore and V. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996.

[38] D. L. Donoho. Orthonormal ridgelets and linear singularities. *SIAM J. Math. Anal.*, 31(5):1062–1099, 2000.

[39] D. L. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report, Stanford University, 2005.

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

[41] D. L. Donoho. For most large underdetermined systems of equations, the minimal $l_1$-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006.

[42] D. L. 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.

[43] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

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

[45] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

[46] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

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

[48] D. L. Donoho and P. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, pages 906–931, 1989.

[49] C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of lp decoding. In *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 85–94, New York, NY, USA, 2007. ACM.

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

[51] J. Feldman and D. R. Karger. Decoding turbo-like codes via linear programming. *J. Comput. System Sci.*, 68(4):733–752, 2004.

[52] J. Feldman, T. Malkin, R. A. Servedio, C. Stein, and M. J. Wainwright. LP decoding corrects a constant fraction of errors. *IEEE Trans. Inform. Theory*, 53(1):82–89, 2007.

[53] J. Feldman and C. Stein. LP decoding achieves capacity. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 460–469 (electronic), New York, 2005. ACM.

[54] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proc. of SODA*, pages 243–252, 2003.

[55] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *Information Theory, IEEE Transactions on*, 49(12):3320–3325, Dec. 2003.

[56] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[57] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[58] M. Jansen. *Noise reduction by wavelet thresholding*, volume 161 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2001.

[59] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

[60] I. M. Johnstone. Chi-square oracle inequalities. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 399–418. Inst. Math. Statist., Beachwood, OH, 2001.

[61] V. Koltchinskii. Dantzig selector and sparsity oracle inequalities. *Bernoulli, to appear*, 2009.

[62] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[63] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[64] M. Lustig, D. Donoho, J. Santos, and J. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008.

[65] S. Mallat. *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA, 1998.

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

[67] P. Marziliano, M. Vetterli, and T. Blu. Sampling and exact reconstruction of bandlimited signals with additive shot noise. *IEEE Trans. Inform. Theory*, 52(5):2230–2233, 2006.

[68] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[69] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[70] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24:227, 1995.

[71] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, In Press, Corrected Proof:–, 2008.

[72] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, pages 1–18.

[73] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44 vol.1, Nov 1993.

[74] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[75] M. Rabbani and R. Joshi. An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication*, 17(1):3–48, 2002.

[76] M. Rudelson and R. Vershynin. Geometric approach to error-correcting codes and reconstruction of signals. *Int. Math. Res. Not.*, (64):4019–4041, 2005.

[77] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. *Information Sciences and Systems, 2006 40th Annual Conference on*, pages 207–212, March 2006.

[78] F. Santosa and W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7:1307, 1986.

[79] J. Starck, M. Elad, and D. Donoho. Redundant multiscale transforms and their application for morphological component separation. *Advances in Imaging and Electron Physics*, 132:288–348, 2004.

[80] J. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE transactions on image processing*, 14(10):1570–1582, 2005.

[81] G. Strang and T. Nguyen. *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley, MA, 1996.

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

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

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

[85] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.

[86] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Math. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.

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

[88] S. van de Geer. Oracle inequalities and regularization. In *Lectures on empirical processes*, EMS Ser. Lect. Math., pages 191–252. Eur. Math. Soc., Zürich, 2007.

[89] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

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

[91] G. Wallace. The JPEG still picture compression standard. *Consumer Electronics, IEEE Transactions on*, 38(1), 1992.

[92] G. Wright. Magnetic resonance imaging. *IEEE Signal Processing Magazine*, 14(1):56–66, 1997.

[93] C.-H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

[94] T. Zhang. Some sharp performance bounds for least squares regression with $l_1$ regularization. Accepted, *Annals of Statistics*, 2007.