EFFICIENT SPARSE APPROXIMATION METHODS FOR MEDICAL IMAGING

by

Ray Maleh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in The University of Michigan
2009

Doctoral Committee:

      Professor Jeffrey A. Fessler, Co-Chair
      Associate Professor Anna C. Gilbert, Co-Chair
      Associate Professor Selim Esedoglu
      Associate Professor Martin J. Strauss

To my Soulmate Christy

ACKNOWLEDGEMENTS

I would like to thank my adviser Anna Gilbert for constantly keeping me on track, for all her help in writing and editing this dissertation, and for supporting me during the preparation of this work. I would like to thank my external adviser Jeffrey Fessler for all his technical suggestions and honest assessments of the performance of my methods, for providing me with corrections related to this work, and for his prompt attention to bureaucratic matters related to my defense. I would like to thank Martin Strauss and Selim Esedoglu for serving on my committee, for reading this dissertation, and for their insightful discussions. I would like to thank Daehyun Yoon for teaching me just about everything I know about MRI scanners and for help with the chapter pertaining to Parallel Excitation in MRI. I would like to thank everyone else who has ever made suggestions related to my work: This list includes, but is not limited to: Doug Noll, Mark Iwen, Paul Shearer, Remi Gribonval, Albert Hero, Raviv Raich, Peter Smereka, Richard Baraniuk, Justin Romberg, Holger Rauhut, Yin Zhang, and many others. I would like to thank my mother and father for their love and motivation. And last but not least, I would like to thank my Soulmate Christy for all her love and for believing in me.

TABLE OF CONTENTS

LIST OF FIGURES

viii

CHAPTER I

Introduction

1.1 The Big Picture

For thousands of years ever since the days of Hippocrates, doctors had to face the daunting task of diagnosing and treating all sorts of medical ailments without the ability to view the "insides" of their patients. It was not until the last one-hundred twenty years that great advances were made in the field of medical imaging. In the mid 1890's, x-rays were first used to image two-dimensional projections of bones and other objects within the human body. While this in itself was a remarkable discovery that aided doctors tremendously, these projections did not paint an entire three-dimensional picture and training was required to accurately interpret these images. The more modern imaging modalities such as CT scans (computer tomography) and MRI (magnetic resonance imaging) were developed in the 1970's in order to generate cross-sectional images of slices of the human body. These slices can be assembled together to paint an overall picture of a person, inside and out. These images are generally much more powerful in terms of revealing anatomical structures, such as tumors or arterial blockages, that may otherwise not be visible in a poorly produced x-ray image.

Physically speaking, CT and MRI work based on two different principles. CT

scanners generate images by shooting x-rays through a patient from all angles and interpolating between the resulting projections to obtain a final cross-sectional image. This work will consider only the simple case of parallel-beam CT where the paths of the x-rays are parallel and the projections are formed on "flat screens". More modernized fan-beam CT systems where x-rays are shot from a motorized source traversing a circle and forming projections onto an arc will not be considered here. On the other hand, MRI works by subjecting a patient's body to varying (gradient) magnetic fields in order to force protons within the subject to precess magnetically at different frequencies depending on their position. These precessing protons will induce a signal through a receiver coil, which can be interpreted to generate an image. A more thorough discussion regarding the operation of these devices is presented in Section 1.3.

While these scientific principles are somewhat different, parallel-beam CT and MRI share one important mathematical property: they both obtain an image by measuring its Fourier spectrum. Since an image is spatially limited, an imaging machine must sample the frequencies in Fourier Space at a rate greater than the Nyquist rate, i.e. the minimum sampling rate that will guarantee no aliasing effects [47]. Now, assuming it is possible to sample at the Nyquist rate, then it is a trivial matter to reconstruct the image by utilizing an Inverse Fast Fourier Transform (IFFT). Unfortunately, obtaining all these samples isn't always very practical: CT scanners and MRI devices tend to be slow, loud, claustrophobic, and generally uncomfortable for patients. Thus, it behooves the medical community to find ways of obtaining high quality medical images utilizing the fewest Fourier Transform coefficients as possible. Luckily, the images of interest tend to enjoy the property of being sparse in gradient. In other words, they consist of large, nearly constant-valued regions separated by

edge-discontinuities. The redundancies in these images suggest a possible dimensionality reduction that should allow for Fourier sampling at sub-Nyquist rates. An example of a sparse-gradient image along with an image of its edges is shown below in Figure 1.1.
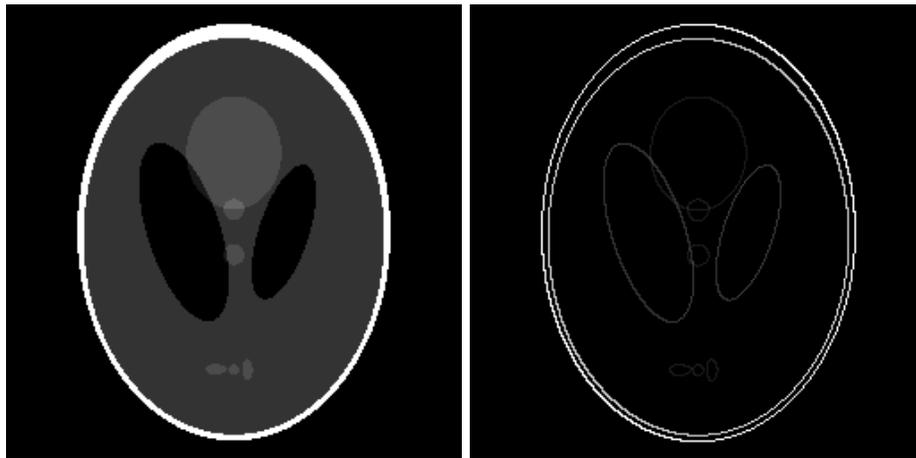
Figure 1.1: The famous Shepp-Logan Phantom image: a classical example of a sparse gradient image.

Mathematically speaking, one has some sparse-gradient image X and some subset Ω of all possible frequency pairs. The objective is to reconstruct X with knowledge only of Ω and

$$(1.1) \qquad\qquad y := \mathcal{F}_\Omega X$$

where $\mathcal{F}_\Omega$ is a 2D Fourier Transform operator restricted to the frequencies specified by Ω. When performing this reconstruction, it is important to keep the following three goals in mind:

1. Make the size of Ω as small as possible to minimize the data that must be collected. In MRI, this translates into faster scan times.

2. Convert the raw Fourier data into a final image using a fast, efficient algorithm so that patients do not need to wait long for their results.

3. Ensure that the final image is an accurate representation of the original image to avoid misdiagnoses.

As a side note, the reason that fan-beam CT is not considered in this work is because a high angular sampling rate is required to adequately interpolate the raw data into parallel-beam coordinates over which a Fourier Transform of the image can be computed. This means that for this particular imaging modality, it is detrimental to reduce the size of Ω.

To reach the above objectives, this work will appeal to the theory of sparse approximation, an area mathematics which has enjoyed significant progress during the last decade or so. In this field, there are several results that state that if some signal (or image) is sparse, say in the Euclidean sense, then it can be recovered exactly with exponentially fewer linear (e.g. Fourier) measurements than the dimension of the actual signal. It should be noted that sparse approximation is closely related to the relatively new field of compressive sensing (see [9]). The basic linear problem from both fields can be modeled by the simple-looking equation:

$$(1.2) \qquad\qquad\qquad\qquad y = \Phi x$$

where in this particular case (which happens to be a compressive sensing application), x is a sparse signal/image, Φ is a measurement matrix, and y is a set of linear measurements of x. The objective is to recover x with knowledge of only Φ and y. Observe the similarity between this equation and Equation 1.1. For the purposes of this overview, the technical difference between sparse approximation and compressive sensing is not terribly important, and therefore, a discussion of this issue is deferred to Section 1.2

Several algorithms utilizing convex optimization and/or greedy basis selection

techniques have been developed mainly for this purpose. Basis Pursuit (BP), discussed in detail by Candes et. al in [5], Donoho in [13], and others, is an example of an optimization principle that, given Fourier observations of a sparse signal, will attempt to seek the sparsest signal (in the ℓ_1 sense) that satisfies the imposed Fourier constraints. In general, Basis Pursuit signal reconstructions are of very high quality; however, algorithms that implement Basis Pursuit are typically very slow. As a result, this methodology may not be appropriate for large signals or time-sensitive procedures. Fortunately, there are faster alternatives. A simpler optimization problem would be to find the signal with the least energy that satisfies the Fourier constraints. While this turns out to be mathematically equivalent to an inverse Fourier Transform with all unknown Fourier coefficients set to zero, reconstruction quality is often very poor. However, this procedure, also known as Fourier Back-Projection, is excellent at identifying the few Euclidean basis elements that contribute the most to the representation of the signal. The most simple compressive sensing algorithm of all, Thresholding (see [52]), works entirely based on this principle. Based on previous results in statistical theory, Mallat (see [41] and [42]) introduces an algorithm known as Matching Pursuit (MP) which attempts to use the Fourier data to iteratively identify the few Euclidean basis elements that make up a sparse signal. While every iteration is extremely fast, there is no known bound on the number of iterations required for convergence. To address this problem, one can use a slightly augmented version of the algorithm known as Orthogonal Matching Pursuit (OMP) (see [58] and [61]), which uses an orthogonalization step to ensure that no single basis element gets selected more than once in subsequent iterations. Thus, the number of iterations must be bounded. However, this comes at the expense of a quadratic runtime with respect to the signal sparsity thanks to the least squares procedure that must

now be carried out. One can make OMP faster by allowing more than one basis element to be selected at any given iteration. This is the primary motivation of the K-fold Orthogonal Matching Pursuit (KOMP) and Stagewise Orthogonal Matching Pursuit (StOMP) procedures described in [22] and [15] respectively. By selecting more than one item per iteration, one typically reduces the number of total iterations needed for recovery, and therefore, the total number of least squares problems that must be solved.

Up until now, these algorithms have not been known to satisfy any sort of performance guarantees based on restricted isometry conditions. In contrast, Basis Pursuit has been shown to possess such properties (e.g. [7] and [6]). A significant contribution of this work is the development of restricted-isometry-property-based sufficient conditions that guarantee the correctness of Orthogonal Matching Pursuit as well as several of its variants. While the restricted isometry conditions pertaining to OMP are not as powerful as its Basis Pursuit counterpart, it will be shown that as one increases the number of elements that may be selected per iteration, one will obtain results that approach the Basis Pursuit golden standard. This analysis will provide some intuition as to why more recently developed algorithms, such as Regularized Orthogonal Matching Pursuit (ROMP), Compressive Sampling Matching Pursuit (CoSAMP), and Iterative Thresholding (IT) (see [46], [45], and [3] respectively), do satisfy these conditions. These three algorithms share the property that they all attempt to simultaneously select all the correct basis elements during every iteration.

It is possible to generalize the theory of sparse signal recovery to cover the case of reproducing sparse-gradient images from a small number of Fourier measurements. In works such as [5], [33], and others, convex optimization methods similar to Basis Pursuit are used to recover such images. These methods seeks the image of smallest

total variation (TV) that satisfies (or approximates) the given Fourier constraints produced by the imaging device. While such constrained TV-minimization procedures greatly reduce the number of Fourier coefficients that must be sampled and are highly effective at producing high quality results, they are extremely slow. One approach to making them faster would be to relax the setup into an unconstrained optimization problem where deviations from the Fourier constraints are penalized [10, 64]. Such methods are faster; however, their theoretical convergence rates can be slow.

This work presents an entirely different approach: It is possible to embed any one of the various orthogonalized matching pursuit algorithms as a black box device within an encompassing algorithm, called Gradient Matching Pursuit (GRADIENTMP), that works specifically to solve the same sparse-gradient image reconstruction problem as shown in (1.1). From a high level point of view, GRADIENTMP has three basic steps.

1. It first modifies the original Fourier measurements to obtain Fourier measurements of the corresponding vertical and horizontal edge images.

2. It then utilizes some algorithm from the suite of matching pursuit routines (e.g. OMP, KOMP, COSAMP, etc.) to recover the edge images.

3. Finally, it recovers the original image from the estimates of its edges using one of several specialized integration techniques.

This procedure will be discussed in much more detail later.

The layout of this work is as follows: The remainder of this chapter will be devoted to presenting background material related to sparse approximation and compressive sensing theory as well as a mathematical description of computer tomography and

magnetic resonance imaging. This is followed by a detailed discussion of the sparse gradient image approximation problem and a survey of previous work. After this thorough introduction, the main contributions of this work are presented:

1.1.1 Contributions of this Work

Chapter II features a detailed theoretical analysis of orthogonal matching pursuit and its variants. Several performance guarantees based on restricted isometry conditions will be proven for these algorithms. This analysis will theoretically justify the use of OMP-like algorithms for use in applications. In other words, this chapter will explain why orthogonal matching pursuit should not be thought of as a mere heuristic.

Chapter III introduces the Gradient Matching Pursuit algorithm and proves performance guarantees pertaining to the various edge integration schemes. These guarantees, combined with the results of Chapter II, establish the theoretical correctness of Gradient Matching Pursuit under certain conditions. However, even in cases where these conditions are not met (such as in natural images), it will be shown that the output of Gradient Matching Pursuit compares well with that of Total Variation Minimization in term of accuracy; however, with respect to runtime, Gradient Matching Pursuit is much faster.

Chapter IV takes on a new direction by attempting to solve the MRI excitation problem: In MRIs with non-ideal RF transmission coils, the strength of a transmitted pulse is not uniform over the imaging region of interest. The result of this is unwanted contrast in the final image reconstruction. Regardless of whether an MRI machine consists of a single coil or multiple transmission coils, this lack of homogeneity can be corrected by transmitting a "short" sequence of weighted RF subpulses modulated by specially selected two-dimensional waveforms induced by linear gra-

dient magnetic coils (see [69], [27], and [68]). The word "short" provides a hint as to where sparsity arises in this setup. In a single coil MRI, selecting this sequence of subpulses translates into a sparse approximation problem easily solvable by either convex optimization or a greedy algorithm like OMP. The parallel coil case is harder. To address this scenario, this chapter will provide an introduction to parallel sparse approximation and compressive sensing theory, which has appeared in the literature in various mathematically equivalent forms (see [54] and [67]). This is followed by a presentation of this work's contribution to the world of parallel approximation which is the development of an efficient generalization of Orthogonal Matching Pursuit known as Parallel Orthogonal Matching Pursuit (POMP). POMP is effectively an interpolation between OMP and its relative Simultaneous Orthogonal Matching Pursuit (SOMP) [62]. This chapter proves several compressive sensing and sparse approximation related theoretical results pertaining the performance of POMP and its many variations. Then it will be shown how POMP can be used effectively to solve the MRI Parallel Excitation (i.e. many transmission coils) problem as accurately as conventional convex optimization techniques, but in much less time.

After presenting three major contributions to the fields of sparse approximation theory, compressive sensing, and medical imaging, this work will conclude by outlining several related questions that are still open for investigation.

1.2   Sparse Approximation and Compressive Sensing Background

To get an understanding of how compressive sensing can be used to recover medical images from Fourier data, it is important to understand the basics of this field. For the purposes of this chapter, it will be assumed that $x \in \mathbb{C}^N$ is a one-dimensional signal of length N. All results presented here are easily adaptable to the two-

dimensional case. The signal (or vector) x is said to be T-sparse if the number of non-zero entries in x is at most T. The matrix $\Phi \in \mathbb{C}^{M \times N}$ denotes a measurement matrix and

(1.3) $$y = \Phi x \in \mathbb{C}^M$$

is a set of M linear measurements of x where M is assumed to be significantly smaller than N. A graphical schematic of this problem is shown in Figure 1.2.
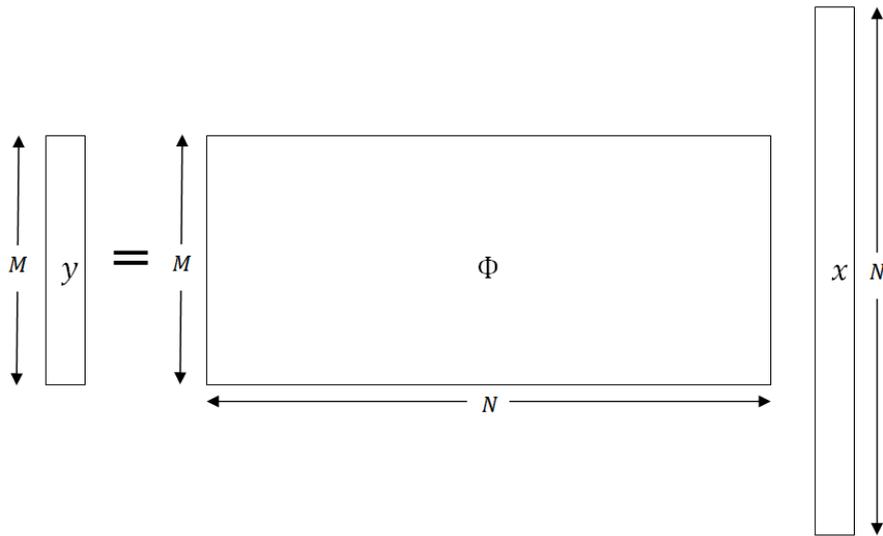
Figure 1.2: A schematic of the basic compressive sensing problem

As a notational convention, let ϕ_i, $1 \leq i \leq N$ denote the columns of Φ. Also, given $A \subseteq \{1, \cdots, N\}$, Φ_A denotes the $M \times |A|$ sub-matrix of Φ corresponding to the indices in A. The basic compressive sensing problem is to solve for x having knowledge of only y and Φ.

This problem is superficially equivalent to the basic linear sparse approximation problem where we are given a signal $y \in \mathbb{C}^M$, an over-complete dictionary Φ, and the task of trying to represent y as a sparse linear combination of the ϕ_is. In both sparse approximation and compressive sensing, the columns ϕ_i are often referred to

as "atoms."

To the outsider and occasionally to even the most seasoned mathematician, the difference between compressive sensing and sparse approximation can be confusing and seem somewhat insignificant. After all, they both involve solving the same underdetermined problem $\Phi x = y$. In addition, the same algorithms can be used in both situations. So what exactly is the difference and why is it a big deal? In simple terms, the main difference between compressive sensing and sparse approximation is that x is the signal of interest in compressive sensing whereas y is the signal of interest in sparse approximation theory. The important distinction lies in the theoretical analyses of compressive sensing and sparse approximation algorithms. Suppose one has an algorithm A that solves both problems. In compressive sensing, one would like to know if A satisfies a performance guarantee of the form $||x - \tilde{x}|| \leq C||x - x_{\text{opt}}||$ where x is some given signal (not necessarily sparse), $\tilde{x}$ is the T-term approximation of x returned by A, x_{opt} is the optimal T-term representation of x, and $|| \cdot ||$ is some suitably selected norm. On the other hand, in sparse approximation theory, one wants to know if A will return an approximation $\tilde{y}$ of y satisfying $||y - \tilde{y}|| \leq C||y - y_{\text{opt}}||$. The methods of proof for both styles of guarantees as well as the utilized norms can be different. Thus, to a theorist, compressive sensing and sparse approximation are far from being one and the same.

One frequently runs into application of both compressive sensing and sparse approximation in medical imaging and therefore this work will consider results pertaining to both situations.

1.2.1 Convex Optimization Approaches

Regardless of whether one is working in the realm of compressive sensing or sparse approximation, solving $y = \Phi x$ is a highly under-determined problem and several

solutions may exist. However, if a sparsity constraint is enforced on x, the problem becomes more interesting. For example, if it is known that x is significantly sparse, one could attempt to solve (1.3) by attempting to find the sparsest signal $z \in \mathbb{C}^N$ that satisfies $\Phi x = \Phi z$. This equates to solving the non-convex optimization problem:

(P0)
$$\tilde{x} = \operatorname*{argmin}_{z} ||z||_0 \quad \text{s.t.} \quad \Phi z = \Phi x.$$

Here, $||\cdot||_0$ denote the ℓ_0 "norm", which does nothing but count the number of non-zero entries in its argument. Unless $P = NP$, there is no polynomial-time algorithm that will, in general, solve (P0). As a result, mathematicians typically relax (P0) into the convex optimization problem shown below:

(P1)
$$\tilde{x} = \operatorname*{argmin}_{z} ||z||_1 \quad \text{s.t.} \quad \Phi z = \Phi x.$$

Now $||\cdot||_p$ for $p > 0$ denotes the usual ℓ_p norm, i.e. $||x||_p = \left(\sum_i |x_i|^p\right)^{1/p}$ for $0 < p < \infty$ and $||x||_\infty = \max_i |x_i|$. There are several algorithms that can solve $(P1)$ in runtime polynomial in N. These include simplex methods, interior point methods, and second order cone programs for complex-valued signals and measurements. These algorithms can be collectively thought of as a generalized algorithmic entity known as Basis Pursuit (BP) [11].

The next logical question to consider is for what types of signals x and measurement matrices Φ is it possible solve (1.3) using (P0) or (P1)? To answer this question, it is essential to introduce the notion of a restricted isometry property (RIP) [4].

Definition I.1. It is said that an $M \times N$ measurement matrix Φ satisfies a restricted isometry property if there exists some number $\delta_T \in [0, 1)$ such that for every T-sparse signal x, it is the case that

$$(1 - \delta_T)||x||_2^2 \leq ||\Phi x||_2^2 \leq (1 + \delta_T)||x||_2^2.$$

The number δ_T is called a restricted isometry number and is a function of the sparsity level T.

Based on the kind of restricted isometry property that a certain measurement matrix Φ enjoys, one can determine a sparsity level T such that all T-sparse signals are recoverable by (P0) or (P1). This is illustrated by the following two propositions.

Proposition I.2. *Suppose one is given a measurement matrix Φ with restricted isometry numbers $\delta_{2T} < 1$. Then the non-convex optimization problem (P0) will recover uniquely any T-sparse signal x given only Φ and Φx.*

Proof. Suppose not. Assume there are two different T-sparse (or less) signals x_1 and x_2 such that $\Phi x_1 = \Phi x_2$. Let $h = x_1 - x_2$. Then $\Phi h = 0$ by linearity. But by the RIP, $(1 - \delta_{2T})||h||_2 \le ||\Phi h||_2 = 0$. Since $\delta_{2T} < 1$, it follows that $h = 0$, which is a contradiction. $\square$

In regards to (P1), one can also prove:

Proposition I.3. *Suppose one is given a measurement matrix Φ with restricted isometry numbers satisfying $\delta_T + \delta_{2T} + \delta_{3T} < 1$, then the convex problem (P1) will recover uniquely any T-sparse signal x given only Φ and Φx.*

The proof of the latter proposition is significantly more involved and can be found in [8]. Candes et. al [6] improve this result to only require that $\delta_{2T} < \sqrt{2} - 1$. Of course, most interesting signals are not truly T-sparse, but only T-sparse up to some additive noise. In addition, there may be some additional measurement noise. To that end, suppose that x is any signal and x_T is the best T-term representation of x, i.e., the largest T terms of x in magnitude. Furthermore, suppose that the given measurements are noisy and take the form $y = \Phi x + e$ where e represents the

measurement noise. Then one may attempt to recover x by solving the following optimization problem:

(PN) $$\tilde{x} = \operatorname*{argmin}_{z} ||z||_1 \quad \text{s.t.} \quad ||\Phi z - y||_2 \leq ||e||_2,$$

which yields the mixed-norm recovery guarantee shown in the following result:

Proposition I.4. *If Φ is a measurement matrix such that the restricted isometry numbers satisfy*

(1.4) $$\delta_{3T} + 3\delta_{4T} < 2,$$

then the solution of (PN) satisfies:

(1.5) $$||\tilde{x} - x||_2 \leq O\left(||e||_2\right) + O\left(\frac{1}{\sqrt{T}} ||x - x_T||_1\right).$$

The proof of this result can also be found in [6]. In addition, one can show that the T-term truncation $\tilde{x}_T$ of $\tilde{x}$ satisfies:

(1.6) $$||\tilde{x}_T - x||_2 \leq O\left(||x - x_T||_2\right) + O\left(\frac{1}{\sqrt{T}} ||x - x_T||_1\right) + O\left(||e||_2\right).$$

An important question to the compressive sensing community how to construct measurement matrices that satisfy the restricted isometry property (1.4). Certainly, if Φ is a square matrix with orthonormal vectors, then the RIP is trivially satisfied. Of course, this case is not very interesting. In an attempt to construct useful examples, mathematicians began experimenting with random matrices consisting of appropriately normalized Gaussian measurements (see [8], [7], and [49]). By selecting an $M \times N$ matrix Φ with Gaussian measurements where M is large enough, i.e.

(1.7) $$M = \Omega\left(T \log\left(\frac{N}{T}\right)\right),$$

then Φ will satisfy restricted isometry properties such as (1.4) with high probability. This result was extended to measurement matrices whose entries follow any Sub-Gaussian distribution [44]. A special case of these distributions is the Bernoulli (0 or 1) distribution which arises in digital compressive sensing applications such as the Rice University Single-Pixel Camera project [28].

This work's primary concern is partial Fourier measurements. In other words, it is of great importance to be able to recover a sparse signal x from a small subset of its Fourier coefficients. To do this, one can define a special measurement matrix Φ_F that is generated by selecting M rows of an $N \times N$ Discrete Fourier Transform (DFT) matrix uniformly at random. The question is: given a sparsity level T and signal size N, how many Fourier measurements M are needed so that Φ satisfies the restricted isometry property in (1.4). This question was answered in 2008 by Rudelson and Vershynin [50]. They prove the following proposition:

Proposition I.5. *If $M = O(T \log(N) \log^2(N) \log(T \log(N)))$, then Φ_F will have restricted isometry numbers that satisfy $\delta_{3T} + 3\delta_{4T} < 2$ with exponentially high probability.*

By combining this result with Proposition I.4 and using the fact that $T \leq N$, one obtains the following corollary.

Corollary I.6. *If*

$$M = \Omega(T \log^4(N)),$$

then, with high probability, solving (PN) will yield a solution $\tilde{x}$ that satisfies the condition in Equation 1.5. In particular, if x is precisely T-sparse and there is no measurement noise, then (PN) will reconstruct x from its measurements exactly.

The quantity $\Omega(T \log^4(N)) \ll N$ for sufficiently large N and small enough T. This

means that it is possible to recover a sparse signal x from far fewer Fourier coefficients than the signal size. In many cases, one can recover x from fewer measurements than what would otherwise be dictated by the Nyquist rate. As an additional note, many mathematicians conjecture that one only really needs $\Omega(T \log(N))$ Fourier measurements to guarantee the RIP condition shown in Proposition I.4. As an illustration of the benefit one would obtain from proving such a conjecture, assume x is a 10-sparse signal of length 2^20. The established result $M = \Omega(T \log^4(N))$ implies that on the order of 1,600,000 Fourier measurements would be required to recover the signal. The conjecture reduces this number to something on the order of 200.

1.2.2   Orthogonal Matching Pursuit

As seen in the previous section, ℓ_1 minimization is a tractable method of solving compressive sensing problems. Unfortunately, algorithms that utilize this methodology to solve this problem tend to be extremely slow. To overcome this issue, other "greedy" techniques have been designed to solve the same problems but with more rapid localized iterative schemes. The most classical of these are Matching Pursuit (MP) (see [42]) and its more powerful brother Orthogonal Matching Pursuit (OMP) (see [22] and [60]). A pseudocode listing of the Orthogonal Matching Pursuit algorithm is shown in Figure 1.3:

OMP works by taking the measurement vector y and expressing it as a sparse linear combination of the ϕ_is, i.e. the columns of Φ. In every iteration, the algorithm selects the column that most closely resembles the current residual in the representation of y. An orthogonalization step is used to ensure that no column is selected more than once. Ideally, by iterating this procedure T times, it should be possible to express y as the exact linear combination of T columns of Φ which form

```
               Algorithm:  Orthogonal Matching Pursuit

Inputs: Φ,  y = Φ,  Sparsity T
Outputs: T term approximation x̃ to x
Initialize r₀ = y,  Λ = ∅
For t from 1 to T {
     Set
```
$$\lambda_t = \operatorname*{argmax}_{i} |(\Phi^* r_{t-1})_i|$$
```
     Set  Λ_t = Λ_{t-1} ∪ {λ_t}
     Set
```
$$a_t = \operatorname*{argmin}_{a} \|\Phi_{\Lambda_t} a - y\|_2 .$$
```
     r_t = y - Φ_{Λ_t} a_t
     }
Set the entries of x̃ corresponding to indices Λ_T equal to a_T.  Set all
other entries equal to zero.
```

Figure 1.3:   Pseudocode for OMP

its representation. If y is reconstructed exactly, then x will be recovered exactly as well.

Previously derived sufficient conditions that guarantee the correctness of the OMP algorithm utilize some additional machinery different from the restricted isometry property presented earlier. To that end, it is necessary to define an alternate notion of how close the columns of Φ resemble each other.

Definition I.7. Given a dictionary Φ with atoms ϕ_i, $1 \leq i \leq N$, the cumulative coherence $\mu(T)$ is defined, as a function of the sparsity parameter T, to be:

$$\mu(T) = \max_{i} \max_{\substack{|\Lambda|=T \\ i \notin \Lambda}} \sum_{j \in \Lambda} |\phi_i^* \phi_j| .$$

The special case $\mu(1)$ is simply referred to as the coherence of the dictionary.

In [58], Tropp proves a sufficient condition that guarantees that OMP will recover x from Φ and Φx.

Proposition I.8. *Suppose that Φ is a dictionary and T is a sparsity parameter such*

that

(1.8) $$\mu(T) + \mu(T-1) < 1,$$

Then OMP will correctly recover any T-sparse signal x from its measurements.

With respect to sparse approximation, Tropp also proves an ℓ_2-ℓ_2 performance guarantee pertaining to OMP's ability to reconstruct some signal y as a T-linear combination of atoms of Φ.

Proposition I.9. *Suppose that Φ is a dictionary and T is a sparsity parameter such that $\mu(T) < 0.5$. Then,* OMP *will produce a T-term approximation $\tilde{y}$ of some signal y such that*

(1.9) $$\|\tilde{y} - y\|_2 \leq \sqrt{1 + \frac{T(1 - \mu(T))}{(1 - 2\mu(T))^2}} \, \|y - y_T\|_2$$

where y_T is the optimal T-term representation of y in Φ.

1.2.3 Iterative Thresholding and Compressive Sampling Matching Pursuit

While Basis Pursuit is slow, its performance is governed by powerful restricted isometry conditions. On the other hand, OMP works very well in practice, is easily implementable, and runs much faster than convex optimization routines. However, its performance guarantees, including those based on the restricted isometry conditions that will be presented in Chapter II, are not as powerful as its slower counterpart. As a result, one may wonder if there is some happy medium, i.e. an algorithm that is as fast and easily implementable as OMP, but also satisfies RIP-based performance guarantees similar to those of convex optimization. The answer turns out to be yes. The first such algorithm is called Iterative Thresholding (IT) [3].

The main idea behind iterative thresholding is quite simple. Given measurements $y = \Phi x$ of some T-sparse signal x, then assuming Φ is nearly unitary, then $x^1 =$

$(\Phi^*y)_T = (\Phi^*\Phi x)_T$, where $(\cdot)_T$ is the non-linear operator that prunes its argument to the top T entries, should be a somewhat good approximation to x. Recall that the restricted isometry property provides a measure of how close a measurement matrix Φ is to being unitary. Now to improve upon this estimate, one may write $x = x^1 + e^1$ where e^1 is the error in the initial estimate. Of course, it is impossible to have a priori knowledge of e_1; however, since Φ is nearly unitary, it follows that

$$e^1 \approx \Phi^*\Phi e^1 = \Phi^*\Phi(x - x^1) = \Phi^*y - \Phi^*\Phi x^1.$$

With this in mind, it is possible to form a second better approximation of x as $x^2 = (x^1 + \Phi^*\Phi(x - x^1))_T$. This procedure can now be iterated to obtain the recurrence relation

$$x^t = \left(x^{t-1} + \Phi^*\Phi(x - x^{t-1})\right)_T.$$

This is the basis of the iterative thresholding (IT) algorithm, which is shown in Figure 1.4.

```
            Algorithm:   Iterative Thresholding

Inputs: Φ,  y = Φx + e where e is noise,  T
Outputs: T term approximation x̃ to x
Initialize x⁰ = 0,  t = 0
While (stop condition not true) {
      xᵗ = (xᵗ⁻¹ + Φ*y − Φ*Φxᵗ⁻¹)_T          }
```

Figure 1.4:   Pseudocode for Iterative Thresholding

Blumensath et. al. (see [3]) prove the following proposition regarding iterative thresholding.

Proposition I.10. *Let Φ be a matrix satisfying an RIP property with $\delta_{3T} < 1/\sqrt{32}$. Let $y = \Phi x + e$ be a set of noisy measurements. Let*

$$\epsilon_T := 5\|x - x_T\|_2 + \frac{1}{\sqrt{T}}\|x - x_T\|_1 + \|e\|_2$$

where x_T is the best T-term approximation of x. Then, after t iterations, x^t will

satisfy

$$||x - x^t||_2 \le 2^{-t}||x_T||_2 + \epsilon_T.$$

Furthermore, after $t^* = \lceil \log(||x_T||_2/\epsilon_T) \rceil$ iterations, one gets that

$$||x - x^{t^*}||_2 \le O\left[||x - x_T||_2 + \frac{1}{\sqrt{T}}||x - x_T||_1 + ||e||_2\right].$$

In particular, if there is no measurement noise and x is exactly T-sparse, then x^t will converge exactly to x.

In many cases, if the Iterative Thresholding algorithm is directly applied as shown above, then algorithm may not converge. In these situations, this can be corrected by using a smaller step-size μ, i.e. at every iteration, set

$$x^t = (x^{t-1} + \mu(\Phi^*y - \Phi^*\Phi x^{t-1}))_T$$

This will yield better results at the expense of runtime. Blumensath offers a well-written MATLAB implementation of this algorithm that will automatically determine an optimal step-size and number of iterations [2].

Along the same lines, Needell and Vershynin [46] developed an algorithm known as Regularized Orthogonal Matching Pursuit ROMP that also allows for the selection of "good" atoms and the removal of "bad" atoms at every iteration. Due to a rather involved combinatorial step (which does actually have a somewhat complicated polynomial time solution with respect to the sparsity level), Needell and Tropp [45] introduced a simpler version of this algorithm known as Compressive Sampling Matching Pursuit (CoSAMP). Pseudocode for this algorithm can be found in Figure 1.5.

It is possible to generalize CoSAMP by allowing it to select αT new atoms instead of $2T$ atoms shown above, i.e. let $\Lambda = \text{supp}(z_{\alpha T})$. This will be denoted as

```
           Algorithm:  Compressive Sampling Matching Pursuit

Inputs: Φ, y = Φx + e, Sparsity T
Outputs: T term approximation x̃ to x
Initialize a₀ = 0, r₀ = y, t = 0
While (stopping condition not met) {
     Set t = t + 1.          Set z = Φ*r_{t-1},
     Set Λ = supp(z_{2T}),    i.e., identify entries of largest 2T entries
        of z in magnitude.
     Set Λ = Λ ∪ supp(a_{t-1}).
     Set b|_Ω = (Φ*_Ω Φ_Ω)^{-1} Φ*_Ω y.
     Set b|_{Ω^C} = 0.
     Set a_t = b_T,            i.e., prune b to the top T entries.
     Set r_t = y - Φa_t.
     }
```

Figure 1.5:   Pseudocode for CoSAMP

CoSAMP_α. With this in mind, the algorithm shown above is actually CoSAMP_2. Later in this work, we will also consider CoSAMP_1 and show empirically that it performs better than CoSAMP_2 in terms of stability.

On close inspection, CoSAMP is very similar to iterative thresholding. The main difference is that at every iteration, the algorithm selects αT new atoms and combines those with the T atoms representing the signal estimate from the previous iteration. Then after performing a least squares projection of the residual onto these (at most $(\alpha + 1)T$) atoms, it prunes the projection coefficients down to the top T entries over which another least squares projection is performed in order obtain a new signal estimate. Needell and Tropp [45] prove a mixed ℓ_2-ℓ_1 performance guarantee (i.e., a bound on the ℓ_2 error that depends on the ℓ_1 norm of the optimal representation error) similar to that of iterative thresholding.

Proposition I.11. *Suppose that Φ is a measurement matrix satisfying $\delta_T < c$ for some constant c. Let $y = \Phi x + e$ be measurements of x contaminated with noise given by e. For a given precision parameter η, CoSAMP_2 will produce a T-sparse*

approximation $\tilde{x}$ of x that satisfies

$$\|\tilde{x} - x\|_2 \leq O\left(max\left\{\eta, \frac{1}{\sqrt{T}}\left\|x - x_{T/2}\right\|_1 + \|e\|_2\right\|\right\}\right).$$

where $x_{T/2}$ is the optimal $(T/2)$-sparse representation of x.

Up to this point, very little motivation has been provided as to why iterative thresholding and CoSAMP closely resemble OMP but possess the theoretical guarantees of Basis Pursuit. The "non-technical" explanation is that much like convex optimization, both Iterative Thresholding and CoSAMP work globablly by attempting to select all the correct non-zero entires of a signal during each iteration. Chapter II will bridge the gap between OMP and these two hybrid algorithms by exploring what happens when OMP is allowed to select more than one item per iteration with the possibility of some of these items being chosen erroneously. It will turn out that this relaxation will actually allow OMP to enjoy stronger performance guarantees based on restricted isometry conditions that more closely resemble those of convex optimization.

1.2.4 Simultaneous Signal Recovery

Another useful extension of the orthogonal matching pursuit concept is the simultaneous recovery of jointly sparse signals. Suppose that $x_1, \cdots, x_K$ are K T-sparse signals with a common support set of size T. As before, Φ is a measurement matrix and $y_1, \cdots, y_K$ are measurement vectors where $y_i = \Phi x_i$. For convenience, $\mathbf{x}$ will denote the $N \times K$ matrix whose columns are the x_is. $\mathbf{y}$ is defined similarly. The goal is to recover all K signals $x_1, \cdots, x_K$ or their common support set. A schematic of this simultaneous compressive sensing problem is shown in Figure 1.6

One way to solve this problem is to utilize an algorithm such as OMP independently K times to recover each x_i from its respective y_i and Φ. This method has two
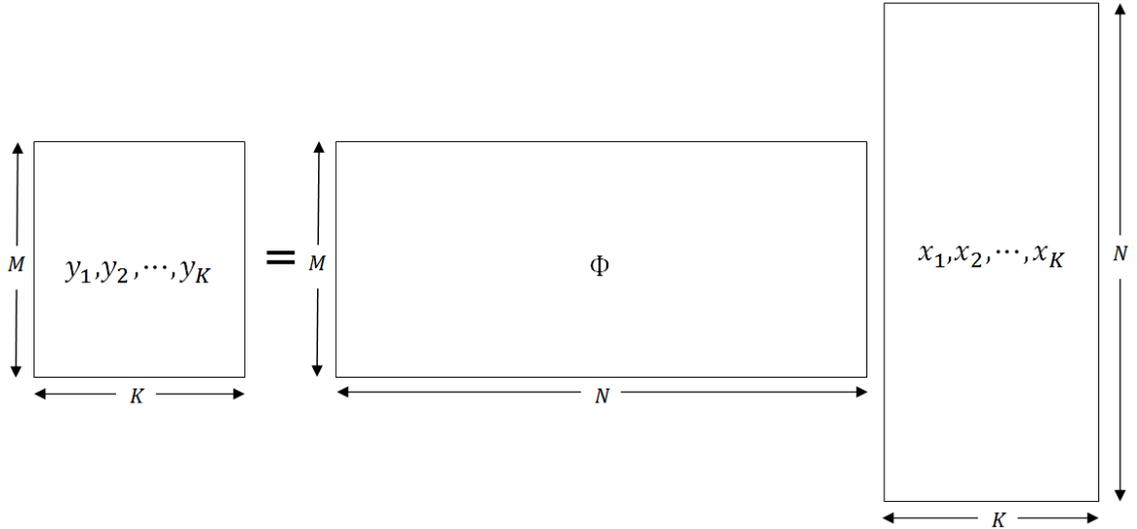
Figure 1.6: A schematic of the basic simultaneous compressive sensing problem

problems associated with it: First, it is not time-efficient as one is required to solve essentially the same problem K times. Secondly, this method is not powerful because it does not exploit the fact that the x_is have a common support set. To address both problems, one may attempt to recover the x_is simultaneously. This can be done with an algorithm known as Simultaneous Orthogonal Matching Pursuit (SOMP_p)) (see [60]) whose pseudocode listing is shown in Figure 1.7:

Observe that the parameter p is used to specify which norm to use in the atom selection criterion. Since this problem setup involves the processing of multiple signals, a norm is required to tie them together. In [60], Tropp proves that Equation 1.8 is also a sufficient condition to ensure that SOMP_1 will recover any jointly T-sparse signals $\mathbf{x}$ from Φ and $\Phi\mathbf{x}$.

By refocusing our interest onto $\mathbf{y}$ instead of $\mathbf{x}$, this setup can also be thought of as a sparse approximation problem. In that light, Tropp also proves the following result regarding signals $\mathbf{y} \in \mathbb{C}^{M \times K}$ that are jointly T-sparse in Φ.

Proposition I.12. *Let Φ be a dictionary and T be a sparsity parameter such that*

```
           Algorithm:  Simultaneous Orthogonal Matching Pursuit (p)

Inputs: Φ, y = Φx, Sparsity T
Outputs: T term approximations x̃ᵢs to the xᵢs
Initialize r₀ = y, Λ = ∅
For t from 1 to T {
     Set
```

$$\lambda_t = \operatorname*{argmax}_i \sum_{k=1}^{K} |(\Phi^* \mathbf{r}_{t-1})_{i,k}|^p$$

```
     Set Λₜ = Λₜ₋₁ ∪ {λₜ}
     Set
```

$$\mathbf{a}_t = \operatorname*{argmin}_{\mathbf{a}} \|\Phi_{\Lambda_t}\mathbf{a} - \mathbf{y}\|_F, \quad \text{(Frobenius Norm)}$$

```
     rₜ = y − Φ_{Λₜ}aₜ
     }
Set the rows of x̃ corresponding to indices Λ_T equal to a_T.  Set all
other entries equal to zero.
```
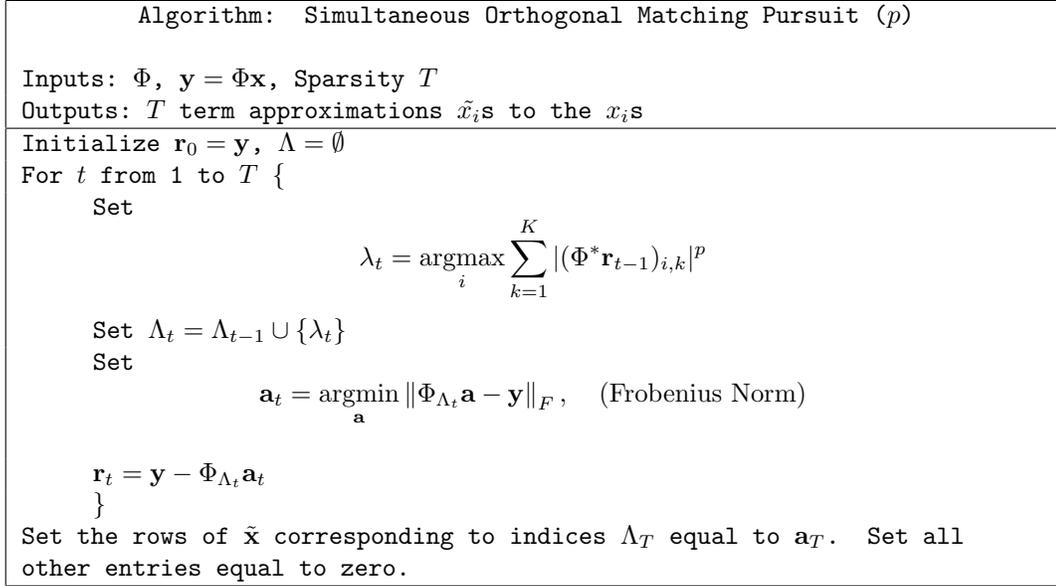
Figure 1.7:   Pseudocode for SOMP(p)

$\mu(T) < 1/2$. *Let* $\mathbf{y}$ *be any set of* K *jointly* T*-sparse signals in* Φ. *Then* SOMP_1 *will return an approximation* $\tilde{\mathbf{y}}$ *satisfying:*

$$(1.10) \qquad \|\tilde{\mathbf{y}} - \mathbf{y}\|_F \leq \left[1 + KT\frac{1 - \mu(T)}{[1 - 2\mu(T)]^2}\right]^{1/2} \|\mathbf{y} - \mathbf{y}_T\|_F$$

where $\mathbf{y}_T$ *is the optimal set of* K *jointly* T*-sparse signals in* Φ *approximating* $\mathbf{y}$ *and* $\|\cdot\|_F$ *represents the Frobenius norm.*

Observe that this result is actually very similar to the one presented earlier for regular OMP. Gribonval et. al. later generalized Tropp's work to cover other values of p (see [26]). It is also possible to prove restricted isometry properties that will generate performance guarantees for SOMP; however, that will not be discussed here.

Simultaneous Orthogonal Matching Pursuit has many applications. It can be used in instances where one has several noisy measurements of a single signal. It is also a powerful tool when processing multi-spectral signals like colored images and

hyperspectral data [53]. In all these examples, one is given each of the measurements $y_i = \Phi x_i$ individually. However, there are more complicated situations where one only has access to the sum of a collection of measurements and not the individual y_is. This will be the subject of Chapter IV.

1.3 A Mathematical Description of Medical Imaging Devices

1.3.1 Parallel-beam Computer Tomography (CT)

The main idea behind parallel-beam computer tomography is as follows: The technique involves shooting x-rays through an object of interest at different angles to generate projections of its cross-section. Then using Fourier Transforms, these projections can be pieced together to generate an image of this cross-section. The basic setup is shown in Figure 1.8.
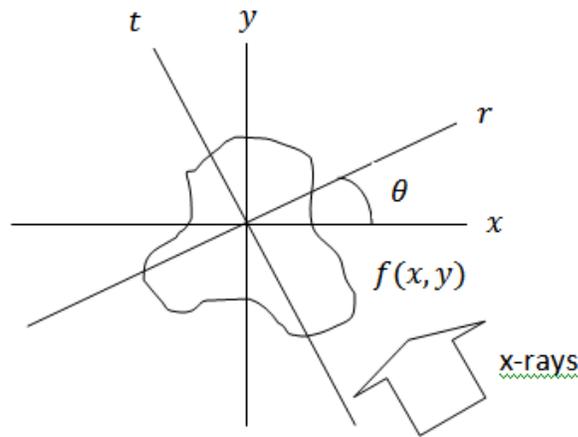
Figure 1.8: Basic Parallel-beam CT Setup: x-rays are shot at an angle in order to generate a projection

Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is an attenuation function representing the cross-section of some object or person that is being imaged. If one shoots an x-ray through an object in a straight line denoted C, then the initial x-ray intensity I_0 will partially

be attenuated by the object of interest and the resulting weaker intensity exiting the object I is given by the following line integral:

$$I = I_0 \exp\left\{-\int_C f(x,y)d\ell\right\}$$

Based on this result, it is apparent that if x-rays are shot in the direction of the t-axis, then the projection $p_\theta(r)$ of $f(x,y)$ onto the r-axis takes the following form:

$$p_\theta(r) = -\ln(I/I_0) = \int_{C_{\theta,r}} f(x,y)d\ell$$

where $C_{\theta,r}$ is the line parallel to the t-axis that passes through the point $(r\cos(\theta), r\sin(\theta))$.

This line-integral can be parameterized using the change of variables $x = -t\sin(\theta) + r\cos(\theta)$ and $y = t\cos(\theta) + r\sin(\theta)$. This yields:

$$(1.11) \qquad p_\theta(r) = \int_{-\infty}^{\infty} f(-t\sin(\theta) + r\cos(\theta), t\cos(\theta) + r\sin(\theta))dt.$$

Now going in a different direction, observe that the 2D-Fourier Transform of $f(x,y)$ is given by

$$\mathcal{F}_{2D}\{f\}(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i(\omega_x x + \omega_y y)}dxdy.$$

Now let $s_\theta(\omega_r)$ be the projection of this Fourier Transform onto the same r-axis (but in frequency space). Then $\omega_x = \omega_r\cos(\theta)$, $\omega_y = \omega_r\sin(\theta)$, and

$$s_\theta(\omega_r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i\omega_r(x\cos(\theta) + y\sin(\theta))}dxdy$$

Now one can perform the same change of variables as before and do some simple algebra to obtain

$$s_\theta(\omega_r) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(-t\sin(\theta) + r\cos(\theta), t\cos(\theta) + r\sin(\theta))e^{-i\omega_r r}dtdr$$

The next step is to exchange the order of integration and recognize from Equation 1.11 the resulting inner integral as $p_\theta(r)$. This yields

$$s_\theta(\omega_r) = \int_{-\infty}^{\infty} p_\theta(r)e^{-i\omega_r r}dr = \mathcal{F}_{1D}\{p_\theta\}(\omega_r)$$

where $\mathcal{F}_{1D}$ represents the continuous one-dimensional Fourier Transform operator.

This observation is the key to understanding parallel-beam computer tomography. By shooting x-rays at an object through different angles, one obtains projections whose Fourier Transforms are equivalent to the two-dimensional Fourier Transform of the cross-sectional image $f(x, y)$ restricted to radial lines. By spacing these angles equally and sampling sufficiently according to the Nyquist criterion, it is possible to recover f from its projections. Let Ω represent a set of two-dimensional frequencies representing the evenly spaced radial lines in frequency described above . Then Ω can be visualized as shown in Figure 1.9.
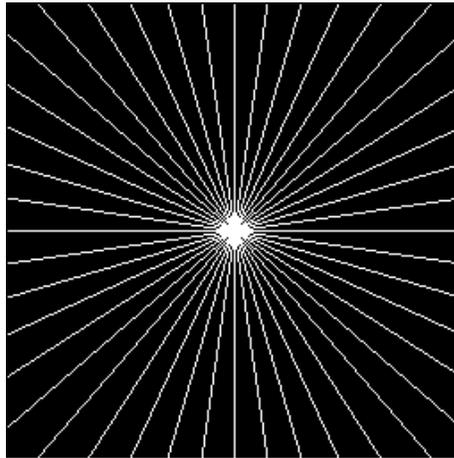
Figure 1.9: Visualization of frequencies over which one can obtain the Fourier Transform of an image via computer tomography.

1.3.2   Magnetic Resonance Imaging (MRI)

MRIs work based on principles somewhat different than that of CT. They form images by measuring the proton (H^+) density inside an object on interest via magnetization and then they return raw data in the form of Fourier Transforms. Suppose one wishes to scan an object with a proton density function given by $\rho(x, y, z)$ and to obtain either a 3D image of ρ or a 2D image of a "slice" of ρ. Let $\mathbf{M}(x, y, z, t)$

denote the magnetization experienced by protons at position (x, y, z) at time t. In a non-time-varying external magnetic field $\mathbf{B}_0$, it is true that $\mathbf{M} \propto \rho \mathbf{B}_0$. In other situations, one can calculate $\mathbf{M}$ from a time-varying magnetic field $\mathbf{B}$ by solving the Bloch equation (see [1]) which, assuming any relaxation terms are negligible, is given by

$$(1.12) \qquad \frac{d}{dt}\mathbf{M}(x, y, z, t) = \gamma \mathbf{M}(x, y, z, t) \times \mathbf{B}(x, y, z, t)$$

where γ is the gyro-magnetic ratio. This is the physical framework that can be used describe the inner workings of MRI.

The core of any MRI machine is a strong magnet producing a constant uniform magnetic field $\mathbf{B}_0$ in the direction of the z-axis. The strength of this magnetic field is typically on the order of 1.5-7 Teslas (T). The first step of the MRI process is called RF-excitation. This involves using current inside a transmission coil to induce a spatially-independent magnetic field $\mathbf{B}_1$ orthogonal to $\mathbf{B}_0$. This will cause $\mathbf{M}$ to point away from the direction of $\mathbf{B}_0$ and have its transverse component precess around the z-axis at the Larmor frequency given by $\omega_0 = \gamma |\mathbf{B}_0|$ as shown in Figure 1.10.
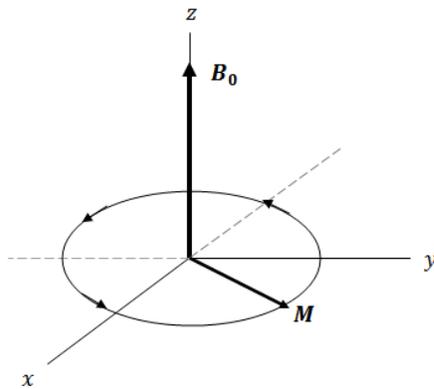
Figure 1.10: Using an RF-pulse to introduce a magnetic field orthogonal to $\mathbf{B}_0$ results in the precession of $\mathbf{M}$ around the z-axis at frequency $\omega_0 = \gamma |\mathbf{B}_0|$.

If the RF pulse is designed correctly (see [18] for more details), it is possible to "excite" only the protons within a thin slice (assumed here to be parallel to the x-y plane) to precess as described above. The angle at which $\mathbf{M}$ is projected out of the longitudinal direction is referred to as the "tip angle". After transmitting the RF pulse, the Bloch equation dictates that the proton magnetization will behave according to the following equation:

$$(1.13) \qquad \mathbf{M}(x,y,z,t) \propto \rho(x,y,z) \left[\sin(\theta) e^{-t/T_2} e^{i(\omega_0 t + \phi)}, \cos(\theta)(1 - e^{-t/T_1}) \right]$$

where θ is the tip angle, T_1 and T_2 are decay parameters (possibly depending on spatial position (x,y,z) that dictate how fast it takes $\mathbf{M}$ to realign with $\mathbf{B}_0$, and t is the time since the RF pulse was turned off. For convenience, the term in brackets is written in the form $[M_x + iM_y, M_z]$. Outside the thin slice, $\mathbf{M}$ is aligned with $\mathbf{B}_0$. After the RF pulse is transmitted, there is often a delay t_e before the image acquisition takes place known as the "echo time." This delay allows for the development of image contrast that will distinguish matter of similar proton density but different relaxation rates. However, for the purposes of this discussion, it will be assumed that only an image reflecting proton density ρ is desired; thus, $t_e = 0$. The precession of the transverse component of $\mathbf{M}$ will induce an electrical current through strategically placed coils. From Maxwell's equations, one can derive that the measured signal in this coil satisfies the following proportionality.

$$S_0(t) \propto e^{i(\omega_0 t + \phi)} \int_{\text{slice}} e^{-t/T_2} \rho(x,y,z) dx dy dz$$

Observe that if T_2 is assumed to be a constant, then this quantity is proportional to the Fourier Transform of ρ at the DC frequency $(0,0)$. Of course, this equation assumes that the receiver coil is perfect and not subject to any non-uniform sensitivity. Otherwise, a weighting function would need to be introduced into the integrand.

Now suppose, on the other hand, that one introduces a new spatially-dependent magnetic field of the form

$$\mathbf{G} = (0, 0, k_x x + k_y y)$$

in addition to $\mathbf{B}_0$ after the RF pulse is done being transmitted. Observe that this new field defines a linear gradient in terms of magnetization. This will ensure that the frequency of precession of the protons now depends on their positions. It is this observation that will allow for the extraction of Fourier Coefficients of our desired image $\rho(x, y, z)$ in the plane $z = z_0$. More specifically, at position (x, y, z), the protons' net magnetization will be precessing at an angular velocity given by $\omega = \omega_0 + \gamma(k_x x + k_y y)$. But this therefore implies that the measured signal $S(t)$ must now take the form

$$S_0(t) \propto e^{i(\omega_0 + \phi)} \int_{\text{slice}} \rho(x, y, z_0) e^{-t/T_2} e^{it\gamma(k_x x + k_y y)} dx dy.$$

Observe that for some fixed t, this is proportional to the 2D Fourier Transform of $\rho(x, y, z_0)$ at the frequency $(\omega_x, \omega_y) = (\gamma t k_x, \gamma t k_y)$. If k_x and k_y are allowed to be time varying functions of t, then they can parameterize any trajectory in the Fourier spectrum. It is possible to sample an entire discrete grid of Fourier coefficients by selecting t, $k_x(t)$, and $k_y(t)$ appropriately. However, as will be shown in this work, one can get away with far fewer frequencies.

1.4 Sparse Gradient Images and Total Variation Minimization Techniques

1.4.1 Notation and Problem Statement

When dealing with images, this work will utilize the following notational conventions. Let $X \in \mathbb{C}^{N \times N}$ denote an image. Any particular pixel of X will be written as $X_{n,m}$ or $X(n, m)$ depending on whichever is more convenient in any given setting.

$X^{(K)}$ is defined to be the image consisting of only the top K largest pixels of X in magnitude and takes the following form:

(1.14)

$$X_{n,m}^{(K)} = \begin{cases} X_{n,m} & \text{if } (n,m) \text{ corresponds to one of the largest } K \text{ pixels of } X \text{ in magnitude} \\ 0 & \text{otherwise} \end{cases}$$

As a convention, any ties in selecting the largest K pixels will be broken lexicographically.

When dealing with individual pixels, it may be convenient to use the delta function image δ_{n_0,m_0}, which is given by

(1.15)
$$\delta_{n_0,m_0}(n,m) = \begin{cases} 1 & \text{if } (n,m) = (n_0, m_0) \\ 0 & \text{otherwise} \end{cases}$$

The discrete directional derivative operators ∂_x and ∂_y on X pixelwise are defined pixel-wise as

(1.16)
$$(\partial_x X)_{n,m} = X_{n,m} - X_{n-1,m}$$

(1.17)
$$(\partial_y X)_{n,m} = X_{n,m} - X_{n,m-1}$$

Based on these, the discrete gradient operator ∇ where $\nabla X \in \mathbb{C}^{N \times N \times 2}$ is defined as

(1.18)
$$(\nabla X)_{n,m} = ((\partial_x X)_{n,m}, (\partial_y X)_{n,m}).$$

From these operators, one can define the discrete total-variational operator TV or $|\nabla|$ on X as

(1.19)
$$(TV[X])_{n,m} = (|\nabla|(X))_{n,m} = \sqrt{|(\partial_x X)_{n,m}|^2 + |(\partial_y X)_{n,m}|^2},$$

from which one can also define the total-variation seminorm of X as

(1.20)
$$||X||_{TV} = ||TV[X]||_1,$$

where $|| \cdot ||_p$ for $0 < p < \infty$ is the ℓ_p norm defined as

$$(1.21) \qquad ||X||_p = \left(\sum_{n=1}^{N} \sum_{m=1}^{N} |X_{n,m}|^p \right)^{\frac{1}{p}}.$$

For $p = 0$, we have the quasi-norm

$$(1.22) \qquad ||X||_0 = \{\#(n,m) : X_{n,m} \neq 0\},$$

i.e., $||X||_0$ is the number of non-zero pixels in X. It is said that an image X is T-sparse if $||X||_0 = T$. It is also said that X is T-sparse in gradient (or in the total-variational sense) if $|||\nabla|(X)||_0 = T$.

As discussed earlier, the goal is to recover an image X that is T-sparse in gradient from a set of $M \ll N^2$ Fourier measurements. To that end, define a set Ω of M two-dimensional frequencies $\underline{\omega}_k = (\omega_{x,k}, \omega_{y,k})$, $1 \le k \le M$ chosen uniformly at random from $\{0, 1, \cdots, N-1\}^2$. Let $\mathcal{F}$ denote the two-dimensional DFT of X and $\mathcal{F}^{-1}$ its inverse. Next define the operator $\mathcal{F}_\Omega : \mathbb{C}^{N \times N} \to \mathbb{C}^M$ as

$$(1.23) \qquad (\mathcal{F}_\Omega X)_k = (\mathcal{F}X)_{\underline{\omega}_k}.$$

$\mathcal{F}_\Omega^*$ will represent its conjugate adjoint.

Equipped with the above notation, the main problem can be stated formally as follows:

Problem I.13. *Given a set Ω of $M \ll N^2$ frequencies and Fourier observations of a T-sparse in gradient image X given by $\mathcal{F}_\Omega X$, how can one recover X accurately and efficiently?*

This work will consider two techniques: total variation minimization which is summarized in this chapter and the proposed algorithm Gradient Matching Pursuit (GRADIENTMP), which will be shown in Chapter III to be a much faster alternative.

1.4.2 Total Variation Minimization

The most popular method of attacking this problem, which will accurately, but not necessarily efficiently, solve it (provided M is sufficiently large) is to find the image of least total variation that satisfies the given Fourier constraints. This corresponds to solving the following convex optimization problem

$$\text{(TV)} \qquad \tilde{X} = \underset{Y}{\operatorname{argmin}} \, ||Y||_{TV} \text{ s.t. } \mathcal{F}_\Omega Y = \mathcal{F}_\Omega X$$

In [5], the following proposition is proven:

Proposition I.14. *Let X be a real-valued T-sparse in gradient image. If $M = O(T \log^4 N)$, then the solution $\tilde{X}$ of (TV) is unique and equal to X with probability at least $1 - O(N^{-M})$.*

In the case of an image corrupted by noise, the measurements take the form

$$\text{(1.24)} \qquad b = \mathcal{F}_\Omega X + w_m$$

where w_m is the measurement noise.

This problem can be solved by a similar convex optimization problem, which can be written as:

$$\text{(TVN)} \qquad \tilde{X} = \underset{Y}{\operatorname{argmin}} \, ||Y||_{TV} \text{ s.t. } ||\mathcal{F}_\Omega Y - b||_2^2 \leq ||w_m||_2^2$$

It is possible to use interior point methods to solve both (TV) and (TVN). However, such algorithms run in time asymptotically polynomial in N^2. For moderate sized images (e.g. 256 by 256), this performance is very slow, especially if any reasonable amount of accuracy is desired.

Fortunately for the proponents of TV minimization, it is possible to speed things up by rewriting (TV) and (TVN) both as an unconstrained optimization problem of

the form

$$\text{(TVUC)} \qquad \tilde{X} = \underset{Y}{\text{argmin}} \, ||Y||_{TV} + \lambda ||\mathcal{F}_\Omega Y - b||_2^2$$

where λ is a penalty parameter. If Y is strictly T-sparse in gradient, then one can pick λ to be very large. In the case of a noisy image or noisy measurements, one can reduce the value of λ to ensure an optimal reconstruction. (TVUC) can be solved by deriving the corresponding Euler-Lagrange equation and then using an iterative artificial time marching PDE scheme or a fixed point algorithm to solve it [10]. Both of these methods have a run time of $O(N^2 \log(N))$ per iteration. The PDE based approach takes too many iterations to achieve sufficient accuracy due to the required CFL condition necessary for numerical stability. While preconditioning techniques have been proposed (see [43]) to relax this CFL condition, the fixed point solvers alleviate this problem altogether.

A new and exciting approach to TV-minimization known as Fast Total Variation Deconvolution, or FTVD for short, was introduced by Wang, Yin, and Zhang in [64]. This procedure works using the principle of alternating minimization. It sets $\tilde{X}$ to be the image Y that minimizes (over all $Y \in \mathbb{C}^{N \times N}$ and $W \in \mathbb{C}^{N \times N \times 2}$) the following expression:

$$\text{(TVd)} \qquad |||W|||_1 + \lambda ||\mathcal{F}_\Omega Y - b||_2 + \beta \, |||W - \nabla Y|||_2$$

Here, λ is the penalty parameter as described before and β is a parameter that controls how much ∇Y must look like W. The procedure works by solving two easy optimization problems in every iteration: First, it assumes that Y is fixed and finds a W that minimizes the above expression. Then it fixes this W and minimizes again to find a new Y. Both minimization subproblems can be solved in closed form: the first one in linear time and the second one in time $O(N^2 \log(N))$ thanks

to a few required FFT computations. Thus, the overall runtime per iteration is $O(N^2 \log(N))$. It can be proven that this procedure does converge (see [63]). If β is small, the convergence is fast; however, the reconstruction error will be large since W, the variable over which the TV minimization is being performed, will not closely resemble the gradient of the reconstructed image Y. On the other hand, if β is large, one will obtain more accuracy at the expense of an increased number of iterations required for convergence. One solution to this tradeoff issue is to begin with a small value of β and then increment it to obtain more accuracy. Figure 1.11 below shows a sketch of the overall FTVD algorithm:

```
                        Algorithm:  FTVD

Inputs: -Set Ω of M frequencies
            -Fourier observations b = F_Ω X  over  Ω
            -Penalty parameter λ
            -Tolerance δ
            -Set of increasing β values:   β_1, β_2, ... β_L
Outputs:  -Estimate X̃ of X

Set  Y = F*_Ω b,  Y_p = 0
For each  ℓ = 1, ··· , L  {
      Set  β = β_ℓ
      While ||Y − Y_p||_2 ≥ δ||Y_p||_2  {
            Set  Y_p = Y.
            Obtain W by minimizing (TVd) with Y fixed.
            Now fix W and minimize (TVd) to obtain a new Y.
      }
}
Set  X̃ = Y.
```

Figure 1.11:   Algorithmic sketch of Fast Total Variation Deconvolution.

The default values for β presented in [64] range from $\beta = 4$ to $\beta = 2^{20}$ where the parameter is incremented by powers of 2. These were also the values that were employed in the implementation of the algorithm used in this work. As will be seen, FTVD is indeed much faster than TV-Minimization with strict Fourier constraints. Also, interestingly enough, the number of iterations required for FTVD to converge does not depend on the image size, making the overall asymptotics with respect

to N seem incredible: only $O(N^2 \log(N))$. However, there is a drawback. Given an error parameter ϵ and assuming that the iterates of FTVD will converge to X, experiments show that $O(1/\epsilon^2)$ iterations are required so that $||\tilde{X} - X||_2 < \epsilon$. This is rather expensive and makes the overall runtime $O(N^2 \log(N)/\epsilon^2)$. In Chapter III, this work will improve upon this result by presenting the Gradient Matching Pursuit algorithm. This algorithm takes a greedy approach to solving the sparse-gradient image reconstruction problem and requires only $O(T)$ iterations, each of which takes time $O(N^2 \log(N))$, resulting in an overall runtime of $O(TN^2 \log(N))$.

Restricted Isometry Conditions For Orthogonal Matching Pursuit

2.1 The Basic Orthogonal Matching Pursuit Algorithm

One of the major contributions of this work is the introduction of a restricted isometry condition that results in a mixed ℓ_2-ℓ_1 norm (e.g., Equation 1.6) compressive sensing performance guarantee for Orthogonal Matching Pursuit. While not as powerful as the result proven for ℓ_1-minimization, this result does provides evidence that the heuristic power of OMP can be analytically verified under certain conditions. Now one may question the importance of this analysis given the fact that other similar algorithms such as Iterative Thresholding and CoSAMP already enjoy such guarantees. The problem with these latter two methods is that they sometimes process too much data too quickly: at every iteration, they attempt to select all T correct entries of a signal at once. Further iterations only try to fine-tune this selection. As will be seen later in this chapter, both of these algorithms become highly unstable if the sparsity parameter T is chosen to be too large. On the other hand, OMP sequentially builds a list of entries without any provision for deletion. Ultimately, there may be a point where this list becomes too long for stable signal recovery; however, one can simply choose to stop running any further iterations when this point is believed to have been reached. With Iterative Thresholding

and CoSAMP, one is doomed in the first iteration. As a result, it is important to consider the development of RIP-based performance guarantees for OMP and its variations discussed in Section 2.2.

This chapter begins with a few important lemmas, the first two of which are proven in [45].

Lemma II.1. *Let x be a T-sparse signal with support set Λ. Let A be any subset of $\{1, \cdots, N\}$ such that $A \cap \Lambda = \emptyset$. Let Φ be a measurement matrix with restricted isometry numbers δ_ℓ. Then the following two properties are true:*

$$\|\Phi_\Lambda^* \Phi x\|_2 \geq (1 - \delta_T) \|x\|_2$$

$$\|\Phi_A^* \Phi x\|_2 \leq \delta_{T+|A|} \|x\|_2.$$

The following lemma describes how Φ blows up the energy contained in a general signal x that does not enjoy any sparsity.

Lemma II.2. *Let Φ be a measurement matrix with restricted isometry number δ_T. Let x be any general signal. Then*

$$(2.1) \qquad \|\Phi x\|_2 \leq \sqrt{1 + \delta_T} \left(\|x\|_2 + \frac{1}{\sqrt{T}} \|x\|_1 \right).$$

It is now possible to use Lemma II.1 and the fact that the restricted isometry numbers δ_T form an increasing sequence in T to prove the following regarding OMP's ability to recover exactly T-sparse signals from their measurements.

Lemma II.3. *Suppose Φ is a dictionary whose RIP constant satisfies*

$$(2.2) \qquad \delta_{T+1} < \frac{1}{1 + \sqrt{T}}.$$

Then OMP will recover any T sparse signal x from Φx.

Proof. Let x be any T-sparse signal supported on a set Λ. At iteration t, suppose that OMP has only selected correct atoms. Let r_t be the current residual. Then $r_t = \Phi c_t$ where c_t is also supported on Λ. Now observe that, by the last lemma,

$$\left\|\Phi_\Lambda^* \Phi c_t\right\|_2 \geq (1 - \delta_T) \left\|c_t\right\|_2 \geq (1 - \delta_{T+1}) \left\|c_t\right\|_2.$$

This implies that

$$\left\|\Phi_\Lambda^* \Phi c_t\right\|_\infty \geq \frac{(1 - \delta_{T+1})}{\sqrt{T}} \left\|c_t\right\|_2.$$

Now let i be any element not in Λ. Then observe that for each such i,

$$\left\|\Phi_{\{i\}}^* \Phi c_t\right\|_\infty = \left\|\Phi_{\{i\}}^* \Phi_\Lambda c_t\right\|_2 \leq \delta_{T+1} \left\|c_t\right\|_2.$$

This implies that

$$\left\|\Phi_{\Lambda^C}^* \Phi c_t\right\|_\infty = \max_{i \in \Lambda^C} \left\|\Phi_{\{i\}}^* \Phi c_t\right\|_\infty \leq \delta_{T+1} \left\|c_t\right\|_2.$$

OMP will recover the next atom correctly provided that

$$||\Phi_\Lambda^* \Phi c_t||_\infty \geq ||\Phi_{\Lambda^C}^* \Phi c_t||_\infty.$$

This will happen if

$$\frac{(1 - \delta_{T+1})}{\sqrt{T}} \left\|c_t\right\|_2 \geq \delta_{T+1} \left\|c_t\right\|_2,$$

which is equivalent to Condition 2.2. $\qquad\qquad\square$

Observe that this result is very similar in nature to that of Proposition I.8. In order to derive a performance guarantee for other less trivial signals that are not exactly T-sparse, one needs to investigate the conditions under which OMP will recover a correct atom at a particular iteration given such a signal. This is done in the next proposition.

Proposition II.4. Φ *be a measurement matrix that satisfies an RIP of the form shown in Condition 2.2. Let* $x \in \mathbb{C}^N$ *be any signal and let* x_T *be its best T-term approximation. Let* $\Lambda = \text{supp}(x_T)$ *and let* $x_{T^C} = x - x_T$*. Suppose OMP has noisy measurements of the form* $y = \Phi x + w = \Phi x_T + e$ *where* $e = \Phi x_{T^C} + w$*. Then after T iterations,* OMP *will recover an estimate* $\tilde{x}$ *of* x *that satisfies:*

$$\|x - \tilde{x}\|_2 \leq (1 + C_1(T)) \|x - x_T\|_2 + \frac{C_1(T)}{\sqrt{T}} \|x - x_T\|_1 + C_1(T) \|w\|_2$$

where, for reasonable restricted isometry numbers, $C_1(T)$ *grows asymptotically like* $\sqrt{T}$*.*

Proof. First suppose that at iteration t, OMP has selected only correct atoms indexed within Λ. At iteration $t + 1$, OMP will select another correct atom provided the greedy selection condition below is satisfied:

$$(2.3) \qquad \frac{\left\|\Phi_{\Lambda^C}^* r_t\right\|_\infty}{\left\|\Phi_\Lambda^* r_t\right\|_\infty} < 1$$

Now rewrite the residual as $r_t = \Phi_\Lambda (x_T - a_t) + e$. Here, a_t is the coefficient vector of the projection of y onto the currently selected atoms. Then one can bound the numerator of (2.3) by:

$$\left\|\Phi_{\Lambda^C}^* r_t\right\|_\infty \leq \left\|\Phi_{\Lambda^C}^* \Phi_\Lambda (x_T - a_t) + \Phi_{\Lambda^C}^* e\right\|_\infty$$

$$\leq \left\|\Phi_{\Lambda^C}^* \Phi_\Lambda (x_T - a_t)\right\|_\infty + \left\|\Phi_{\Lambda^C}^* e\right\|_\infty$$

$$\leq \delta_{T+1} \left\|x_T - a_t\right\|_2 + \left\|e\right\|_2.$$

On the other hand, the denominator can be bounded from below by:

$$\|\Phi_\Lambda^* r_t\|_\infty \geq \frac{1}{\sqrt{T}} \|\Phi_\Lambda^* r_t\|_2$$

$$= \frac{1}{\sqrt{T}} \|\Phi_\Lambda^* \Phi_\Lambda (x_T - a_t) + \Phi_\Lambda^* e\|_2$$

$$\geq \frac{\|\Phi_\Lambda^* \Phi_\Lambda (x_T - a_t)\|_2 - \|\Phi_\Lambda^* e\|_2}{\sqrt{T}}$$

$$\geq \frac{(1 - \delta_{T+1})}{\sqrt{T}} \|x_T - a_t\|_2 - \sqrt{\frac{1 + \delta_{T+1}}{T}} \|e\|_2.$$

A sufficient condition for the next atom to be selected correctly is that the numerator is less than the denominator. This is guaranteed if

$$\delta_{T+1} \|x_T - a_t\|_2 + \|e\|_2 < \frac{(1 - \delta_{T+1})}{\sqrt{T}} \|x_T - a_t\|_2 - \sqrt{\frac{1 + \delta T + 1}{T}} \|e\|_2.$$

Rearrange terms to obtain:

$$\|x_T - a_t\|_2 > \frac{\sqrt{T} + \sqrt{1 + \delta_{T+1}}}{1 - \delta_{T+1}(1 + \sqrt{T})} \|e\|_2.$$

Let t^* denote the first iteration where

$$\|x_T - a_{t^*}\|_2 \leq \frac{\sqrt{T} + \sqrt{1 + \delta_{T+1}}}{1 - \delta_{T+1}(1 + \sqrt{T})} \|e\|_2.$$

Now observe that, by the definition of OMP, $\tilde{x} = a_T$. Then

$$\|x - \tilde{x}\|_2 = \|x_T - \tilde{x} + x_{T^C}\|_2 \leq \|x_T - \tilde{x}\|_2 + \|x_{T^C}\|_2$$

$$\leq \frac{1}{\sqrt{1 - \delta_{2T}}} \|\Phi_{\Lambda'} (x_T - \tilde{x})\|_2 + \|x_{T^C}\|_2$$

where $\Lambda' = \Lambda \cup \text{supp}(\tilde{x})$ which has cardinality at most $2T$. It is further possible to bound the left hand side by:

$$\|x - \tilde{x}\|_2 \leq \frac{1}{\sqrt{1 - \delta_{2T}}} \left(\|\Phi_{\Lambda'} (x_T - \tilde{x}) + e\|_2 + \|e\|_2 \right) + \|x_{T^C}\|_2$$

$$\leq \frac{1}{\sqrt{1 - \delta_{2T}}} \left(\|\Phi_\Lambda (x_T - a_{t^*}) + e\|_2 + \|e\|_2 \right) + \|x_{T^C}\|_2$$

$$\leq \frac{1}{\sqrt{1 - \delta_{2T}}} \left(\|\Phi_\Lambda (x_T - a_{t^*})\|_2 + 2 \|e\|_2 \right) + \|x_{T^C}\|_2$$

$$\leq \frac{\sqrt{1 + \delta_T}}{\sqrt{1 - \delta_{2T}}} \|x_T - a_{t^*}\|_2 + \frac{2}{1 - \delta_{2T}} \|e\|_2 + \|x_{T^C}\|_2$$

where the second inequality comes from the fact that in OMP, the residual is always decreasing in magnitude regardless whether correct atoms are being chosen or not. Now combining this with the bound on $||x_T - a_{t^*}||_2$ yields:

$$\|x - \tilde{x}\|_2 \leq \left[\left(\frac{\sqrt{1 + \delta_T}}{\sqrt{1 - \delta_{2T}}} \right) \left(\frac{\sqrt{T} + \sqrt{1 + \delta_{T+1}}}{1 - \delta_{T+1}(1 + \sqrt{T})} \right) + \frac{2}{1 - \delta_{2T}} \right] \|e\|_2 + \|x_{T^C}\|_2 .$$

For convenience, let $C_1'(T)$ denote the expression within the brackets. Now, one can bound $||e||_2$ as follows:

$$\|e\|_2 = \|\Phi x_{T^C} + w\|_2$$

$$\leq \|\Phi x_{T^C}\|_2 + \|w\|_2$$

$$\leq \sqrt{1 + \delta_T} \left(\|x_{T^C}\|_2 + \frac{1}{\sqrt{T}} \|x_{T^C}\|_1 \right) + \|w\|_2 .$$

Finally, let $C_1(T) = \sqrt{1 + \delta_T} C_1'(T)$ to obtain that

$$\|x - \tilde{x}\|_2 \leq (1 + C_1(T)) \|x_{T^C}\|_2 + \frac{C_1(T)}{\sqrt{T}} \|x_{T^C}\|_1 + C_1(T) \|w\|_2$$

$$= (1 + C_1(T)) \|x - x_T\|_2 + \frac{C_1(T)}{\sqrt{T}} \|x - x_T\|_1 + C_1(T) \|w\|_2$$

as was to be shown. $\square$

This is the first known mixed ℓ_2-ℓ_1 performance guarantee proven for OMP. While not a near-optimal guarantee, it does bound by how much worse than the optimal T-term approximation error that the reconstruction error can be. In addition, it shows the OMP is fairly robust with respect to measurement noise. Unfortunately, this result is not as strong as that provable for Basis Pursuit because of the $\sqrt{T}$ term in C_1. In addition, because of the $\sqrt{T}$ factor in the denominator of the required RIP, designing typical Gaussian or Fourier measurement matrices that satisfy (2.2) becomes more expensive in terms of the number of rows required. It can be shown that:

Proposition II.5. *Suppose Φ is an $M \times N$ Gaussian (or Sub-Gaussian) measure-ment matrix. If*

$$M = \Omega \left(T^2 \log \left(\frac{N}{T} \right) \right),$$

then (2.2) will be satisfied with high probability. On the other hand, if Φ is an $M \times N$ partial Fourier measurement matrix with

$$M = \Omega \left(T^2 \log^4(N) \right),$$

then (2.2) will be satisfied.

The proof simply follows from the fact that the constants specified by the big-O terms are proportional to $1/\epsilon^2$ where ϵ is related to the probability that a measure-ment matrix of size $M \times N$ will satisfy a RIP.

As an example, for the same T sparse signal x, one would need $\Omega(T^2 \log(N/T))$ Gaussian measurements to guarantee that OMP will recover x whereas convex opti-mization would only need $\Omega(T \log(N/T))$. While these are asymptotically equivalent in N, there is still an extra factor of T that should otherwise not be needed. Fortu-nately, there is a way around this problem. In the next section, it will be shown that by modifying the basic OMP algorithm so that it can select more than one atom per iteration and allowing for the possibility of a controlled number of incorrectly chosen atoms, one can improve upon these theoretical performance bounds. In fact, it will be demonstrated that by selecting a sufficiently large number (say αT for some fraction $0 < \alpha \leq 1$) of atoms per iteration, one can obtain results similar to those of Basis Pursuit.

2.2 Variations of Orthogonal Matching Pursuit

This section will explore the result if Orthogonal Matching Pursuit is allowed to select more than one atom per iteration. First, we discuss variations of OMP that

have such a provision. Then the focus will be narrowed upon one of these algorithms, namely 2-Orthogonal Matching Pursuit, where exactly two atoms are selected per iteration. We prove a performance guarantee for this algorithm based on a restricted isometry property. Unfortunately, this guarantee depends upon the iteration at which a certain condition fails to hold. Assuming the restricted isometry numbers of a measurement matrix do not grow too quickly, the performance guarantee can be uniformly bounded with respect to iteration numbers. This new bound will be used to show that 2-OMP performs slightly better than regular OMP in terms of reconstruction error. This will lay the framework that will be used to justify how the Hybrid Orthogonal Matching Pursuit algorithm of the next section possesses error bounds asymptotically equivalent to that enjoyed by Basis Pursuit.

There is a multitude of variations of the basic OMP algorithm. For now, this work will concern itself with three of these: K-fold Orthogonal Matching Pursuit (KOMP), Stagewise Orthogonal Matching Pursuit (StOMP) (see [15]), and Orthogonal Matching Pursuit with Thresholding (OMP-THRESH). The only difference between OMP, KOMP, StOMP, and OMP-THRESH is the number of atoms selected at each iteration and the criteria used for selecting these atoms. All other steps in the algorithm remain the same. These different criteria are outlined in Figure 2.1.

OMP	Select the one i that maximizes $|(\Phi^* r_t)_i|$
KOMP	Select the is corresponding to the K largest values of $|(\Phi^* r_t)_i|$
StOMP	Select the is such that $|\phi_i^* r_t| \geq \tau_t ||\Phi^* r_t||_\infty$ for some $\{\tau_1, \cdots, \tau_T\} \subseteq [0,1]$
OMP-THRESH	Select the is such that $|\phi_i^* r_t| \geq \tau ||\Phi^* r_t||_\infty$ for some $0 < \tau < 1$.

Figure 2.1: Atom selection critera for OMP, KOMP, StOMP, and OMP-THRESH.

As a side note, observe that OMP-THRESH is a special case of StOMP with all thresholds being set as a constant value $\tau_t = \tau$. Since these three modified algorithms can select more than one atom per iteration, these algorithms typically

converge faster than regular OMP.

In many cases, the three modified algorithms actually perform more accurately than ordinary OMP. This typically occurs in cases where one is given a measurement or dictionary matrix that is not extremely incoherent, i.e. has columns that are highly correlated. As an example of this, consider Figure 2.2.
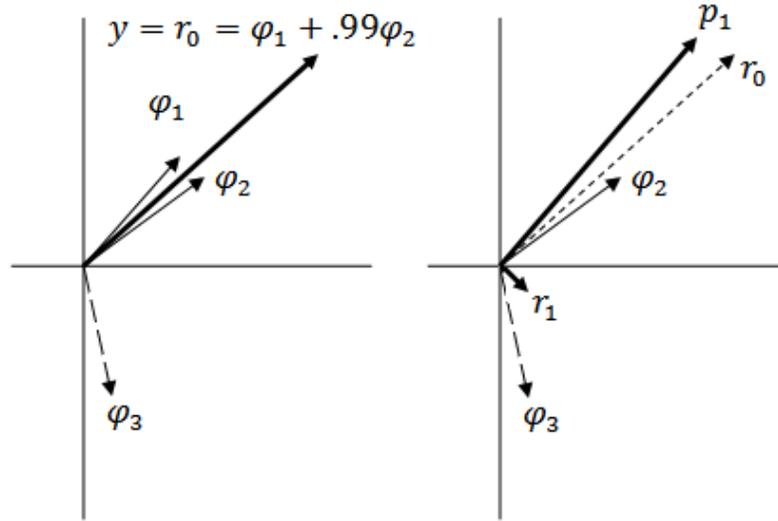
Figure 2.2: Graphical illustration of when thresholding is useful.

Suppose one has a signal (or measurement vector) $y \in \mathbb{R}^3$ that is a linear combination of two highly coherent dictionary vectors ϕ_1 and ϕ_2. Let ϕ_3 be a third dictionary vector that sits slightly outside of the plane spanned by ϕ_1 and ϕ_2. If one ran ordinary OMP on this problem setup, then OMP would select ϕ_1 during the first iteration. Then OMP would project y onto ϕ_1 and then subtract this projection from $r_0 = y$ to form the next residual r_1. Since ϕ_2 is highly correlated to ϕ_1, it follows that a large portion of the contribution of ϕ_2 would be subtracted from the initial residual, which would result in r_1 being more correlated to ϕ_3. Thus, OMP would select an incorrect atom in the second and final iteration. If, on the other hand, one

ran SтOMP or OMP-тнreѕн with an appropriately chosen threshold τ, then the algorithm would correctly select both ϕ_1 and ϕ_2 in the first iteration and stop. This would yield a more accurate reconstruction than regular OMP in approximately half the time.

The above analysis provides an intuitive idea of why one actually gains more accuracy by allowing a matching pursuit algorithm to select more than one atom per iteration. However, in the case of SтOMP and OMP-тнreѕн, selecting an appropriate threshold is a highly non-trivial issue. If a threshold is set too low, these is a good chance that these algorithms will select too many incorrect atoms, which will significantly affect the quality of the final signal reconstruction. However, if a threshold is set too high, then it may not be possible to realize the full benefit of utilizing a threshold-based atom selection scheme. Along these lines, [15] introduces two threshold selection schemes based on the ideas of false alarm and false discovery control from statistical decision theory. This work, on the the hand, will present an RIP-based result that will determine how to appropriately select the thresholds τ_t in order to guarantee that SтOMP will perform correctly. It will be shown that given a set of good thresholds and any T-sparse signal x, SтOMP will choose all the correct non-zero entries of x and will keep the number of incorrectly chosen entries under control (so that one can still obtain an exact reconstruction of x).

Proposition II.6. *Let x be any T-sparse signal and let Φ be a measurement matrix whose restricted isometry numbers satisfy:*

$$(2.4) \qquad \delta_{T+t+1} \leq \frac{1}{1 + 2^{-1/4}\sqrt{T-t+1}}$$

for each $t = 1, \cdots, T$. Suppose at each iteration we set the threshold τ_t as follows:

$$(2.5) \qquad \tau_t \in \left(\frac{\delta_{T+t+1}\sqrt{T-t+1}}{\sqrt{2}\left(1 - \delta_{T+t+1}\right)}, \frac{1 - \delta_{T+t+1}}{\delta_{T+t+1}\sqrt{T-t+1}} \right).$$

Then, at every iteration, STOMP *will select at least one correct non-zero entry x and no more than one incorrect entry. In particular, after T iterations,* STOMP *will have identified all T non-zero entries of x and no more than T incorrect entries.*

Proof. Proceed by induction. Suppose that after t iterations, STOMP has selected at least t correct atoms indexed by the set A_t and no more than t incorrect atoms indexed by B_t. Let r_t denote the current residual and c_t be the corresponding coefficient vector, i.e. $r_t = \Phi c_t$. First, make the observation that

$$\left\| \Phi^*_{\Lambda \cup B_t} \Phi c_t \right\|_2 \geq (1 - \delta_{T+t+1}) \left\| c_t \right\|_2$$

since c_t is supported on $\Lambda \cup B_t$ and $|\Lambda \cup B_t| \leq T + t + 1$. By orthogonality, it is known that $\Phi^*_{\Lambda \cup B_t} \Phi c_t$ is zero on A_t and B_t. Therefore,

$$\left\| \Phi^*_\Lambda \Phi c_t \right\|_\infty \geq \frac{(1 - \delta_{T+t+1})}{\sqrt{T - t + 1}} \left\| c_t \right\|_2 .$$

The next step is to show that only one entry of $\Phi^*_{\Lambda^C} \Phi c_t$ can be greater than $\delta_{T+t+1} \|c_t\|_2 / \sqrt{2}$ in magnitude. To do this, suppose the contrary. Let i and j be two indices in Λ^C such that $|(\Phi^* \Phi c_t)_i| > \delta_{T+t+1} \|c_t\|_2 / \sqrt{2}$ and $|(\Phi^* \Phi c_t)_j| > \delta_{T+t+1} \|c_t\|_2 / \sqrt{2}$. Then it would follow that

$$\left\| \Phi^*_{\{i,j\}} \Phi c_t \right\|_2 > \delta_{T+t+1} \left\| c_t \right\|_2 ,$$

which would violate the restricted isometry condition that is being enforced. Now, it must be shown that STOMP will select at least one correct atom. Well, certainly if the maximum value of $|(\Phi^* \Phi c_t)_i|$ takes place for some $i \in \Lambda$, then the result is trivial. So suppose $i \notin \Lambda$. Then for a correct atom to be selected, it must be the case that

$$\left\| \Phi^*_{\Lambda \cup B_t} \Phi c_t \right\|_\infty \geq \tau \left\| \Phi^*_{\Lambda^C} \Phi c_t \right\|_\infty ,$$

which will definitely occur if

$$\tau \delta_{T+t+1} \|c_t\|_2 \leq \frac{(1 - \delta_{T+t+1})}{\sqrt{T - t + 1}} \|c_t\|_2 \,.$$

One can rearrange terms to obtain that $\tau \leq (1 - \delta_{T+t+1})/(\delta_{T+t+1}\sqrt{T - t + 1})$, which is equivalent to the right end-point constraint in Expression 2.5. The final step is to show that no more than one incorrect atom will be chosen. Because there can only be one index $i \notin \Lambda$ such that $|(\Phi^*\Phi c_t)_i| > (\delta_{T+t+1}/\sqrt{2})\|c_t\|_2$, it follows that no more than one incorrect atom will be chosen if

$$\tau \frac{(1 - \delta_{T+t+1})}{\sqrt{T - t + 1}} \|c_t\|_2 \geq \frac{\delta_{T+t+1}}{\sqrt{2}} \|c_t\|_2 \,.$$

Rearranging terms shows that this is equivalent to the left end point of Expression 2.5. Finally observe that for iteration t, Expression 2.5 defines a non-empty interval if, and only if, Condition 2.4 holds. $\qquad\square$

From this proof, one can draw an immediate corollary.

Corollary II.7. *Let (a_t, b_t) be the intervals defined by (2.5) for $t = 1, \cdots, T$. If*

$$\bigcap_{t=1}^{T}(a_t, b_t) \neq \emptyset,$$

then there exists some threshold τ so that for any T-sparse signal x, after T iterations, OMP-THRESH will correctly recover all T non-zero entires of x and no more than T incorrect entries.

Now using essentially a very similar (and somewhat easier) argument as before, one can prove the following proposition regarding the KOMP algorithm with $K = 2$.

Proposition II.8. *Suppose that Φ is a measurement matrix with RIP constants that, for each $t = 1, \cdots T$, satisfy*

(2.6)
$$\delta_{T+t+1} < \frac{1}{1 + \sqrt{\frac{T-t+1}{2}}},$$

then, after T iterations, KOMP *with* $K = 2$ *will recover all the correct entries of any T-sparse signal x as well as no more than T incorrect entries.*

The last three results provide sufficient condition that guarantee that StOMP, OMP-THRESH, and 2-OMP will recover the correct non-zero entries of any T-sparse signal x along with no more than T incorrect entries. A consequence of this is that, if $\delta_{2T} < 1$, then performing a least squares projection of $y = \Phi x$ onto the at-most-$2T$ selected atoms will result in a coefficient vector $\tilde{x}$ that is exactly equal to x.

It is possible to also prove a more general result about 2-OMP's ability to recover signals x that do not necessarily enjoy any sparsity properties. To do this, it is necessary to define the following special norm:

Definition II.9. Let x be a signal with sorted entries $x_{(1)}, x_{(2)}, \cdots, x_{(N)}$ where $|x_{(1)}| \geq |x_{(2)}| \geq \cdots \geq |x_{(N)}|$. Then the top-$K$ norm is defined to be

$$\|x\|_{\mathrm{Top}K} := \big\|(x_{(1)}, x_{(2)}, \cdots, x_{(K)})\big\|_2 = \sqrt{\sum_{k=1}^{K} |x_{(k)}|^2}.$$

It is easy to verify that $\|\cdot\|_{\mathrm{Top}K}$ does indeed satisfy the three fundamental norm properties. As special cases, observe that $\|x\|_{\mathrm{Top}1} = \|x\|_{\infty}$ and $\|x\|_{\mathrm{Top}N} = \|x\|_2$.

In this particular setting, the top-2 norm will be useful for producing a "greedy selection ratio" that can be used to determine if, at any given iteration, 2-OMP will select at least one correct atom. The following proposition's proof parallels that of Proposition II.4; however, it is presented in full detail because of the additional intricacies required to handle the processing of incorrectly chosen atoms.

Proposition II.10. *Let* Φ *be a measurement matrix that satisfies the RIP shown in Equation 2.6. Let* $x \in \mathbb{C}^N$ *be any signal with best T-term approximation* x_T, *the latter being supported on the set* Λ. *Suppose, as before, that* 2-OMP *is given noisy*

measurements of the form $y = \Phi x + w = \Phi x_T + e$ *where* $e = \Phi x_{T^C} + w$. *Then* 2-OMP *will recover a* $2T$-*sparse approximation* $\tilde{x}$ *of* x *that satisfies:*

$$\|x - \tilde{x}\|_2 \leq (1 + C_2(T, t^*)) \|x - x_T\|_2 + \frac{C_2(T, t^*)}{\sqrt{T}} \|x - x_T\|_1 + C_2(T, t^*) \|w\|_2$$

where t^* *is the stopping time of "good iterations," i.e. the first iteration in which no correct entries from* Λ *will be selected in iteration* $t^* + 1$ *and* C_2 *grows roughly like* $\sqrt{T/2}$.

Proof. Suppose that at every iteration up to iteration t, 2-OMP has selected at least one correct atom and at most one incorrect atom. Let B_t represent the set of incorrectly selected atoms and let $\Lambda_t = \Lambda \cup B_t$. Observe that, by assumption, the cardinality of Λ_t does not exceed $T + t$. Now at iteration $t + 1$, 2-OMP will select at least one more correct atom if the greedy selection ratio satisfies:

$$\frac{\left\|\Phi_{\Lambda_t^C}^* r_t\right\|_{\text{Top2}}}{\left\|\Phi_{\Lambda_t}^* r_t\right\|_{\text{Top2}}}.$$

Now the numerator of this expression can be bounded from above as follows:

$$\left\|\Phi_{\Lambda_t^c}^* r_t\right\|_{\text{Top2}} \leq \left\|\Phi_{\Lambda_t^C}^* \Phi_{\Lambda_t} (x_T - a_t) + \Phi_{\Lambda_t^C}^* e\right\|_{\text{Top2}}$$

$$\leq \delta_{T+t+1} \|x_T - a_t\|_2 + \sqrt{1 + \delta_2} \|e\|_2.$$

Using similar arguments, the denominator can be bounded from below via:

$$\left\|\Phi_{\Lambda_t}^* r_t\right\|_{\text{Top2}} \geq \frac{\left\|\Phi_{\Lambda_t}^* r_t\right\|_2}{\sqrt{(T - t + 1)/2}}$$

$$\geq \frac{\left\|\Phi_{\Lambda_t}^* \Phi_{\Lambda_t} (x_T - a_t) + \Phi_{\Lambda_t}^* e\right\|_2}{\sqrt{(T - t + 1)/2}}$$

$$\geq \frac{\left\|\Phi_{\Lambda_t}^* \Phi_{\Lambda_t} (x_T - a_t)\right\|_2}{\sqrt{(T - t + 1)/2}} - \frac{\left\|\Phi_{\Lambda_t}^* e\right\|_2}{\sqrt{(T - t + 1)/2}}$$

$$\geq \left(\frac{1 - \delta_{T+t+1}}{\sqrt{(T - t + 1)/2}}\right) \|x_T - a_t\|_2 - \frac{\sqrt{1 + \delta_{T+t+1}}}{\sqrt{(T - t + 1)/2}} \|e\|_2$$

The $T - t + 1$ factor in the square root term comes from the fact that $\Phi^*_{\Lambda_t} r_t$ will be zero on the entries within Λ that have already been selected. A sufficient condition for 2-OMP to chose at least one correct atom in iteration $t + 1$ is that the numerator is less than the denominator. Using the same argument as before, this is guaranteed if:

$$\|x_T - a_t\|_2 > \frac{\sqrt{1 + \delta_2}\sqrt{(T - t + 1)/2} + \sqrt{1 + \delta_{T+t+1}}}{1 - \delta_{T+t+1}(1 + \sqrt{(T - t + 1)/2})} \|e\|_2 .$$

Again let t^* be the first iteration at which the above inequality does not occur. In other words, one has:

$$\|x_T - a_{t^*}\|_2 \leq \frac{\sqrt{1 + \delta_2}\sqrt{(T - t^* + 1)/2} + \sqrt{1 + \delta_{T+t^*+1}}}{1 - \delta_{T+t^*+1}(1 + \sqrt{(T - t^* + 1)/2})} \|e\|_2 .$$

The next step is to obtain a bound on $||x - \tilde{x}||$ as follows:

$$\|x - \tilde{x}\|_2 = \|x_T - \tilde{x} + x_{T^C}\|_2 \leq \|x_t - \tilde{x}\|_2 + \|x_{T^C}\|_2$$

$$\leq \frac{1}{\sqrt{1 - \delta_{3T}}} \left(\|\Phi_{\Lambda'}(x_T - \tilde{x}) + e\|_2 + \|e\|_2\right) + \|x_{T^C}\|_2$$

where the last inequality comes from the fact that the support Λ' of $x_T = \tilde{x}$ is at most $3T$. Now recognizing $\Phi_{\Lambda'}(x_T - \tilde{x}) + e$ as the final residual and using the fact that the residual in 2-OMP is non-increasing to bound this by:

$$\|x - \tilde{x}\|_2 \leq \frac{1}{\sqrt{1 - \delta_{3T}}} \left(\|\Phi_{\Lambda^{t^*}}(x_T - a_{t^*}) + e\|_2 + \|e\|_2\right) + \|x_{T^C}\|_2$$

$$\leq \frac{1}{\sqrt{1 - \delta_{3T}}} \left(\|\Phi_{\Lambda^{t^*}}(x_T - a_{t^*})\|_2 + 2\|e\|_2\right) + \|x_{T^C}\|_2$$

$$\leq \frac{\sqrt{1 + \delta_{T+t^*+1}}}{\sqrt{1 - \delta_{3T}}} \|x_T - a_{t^*}\|_2 + \frac{2}{\sqrt{1 - \delta_{3T}}} \|e\|_2 + \|x_{T^C}\|_2 .$$

Putting the two big pieces together yields:

$$\|x - \tilde{x}\|_2$$

$$\leq \left[\left(\frac{\sqrt{1 + \delta_{T+t^*+1}}}{\sqrt{1 - \delta_{3T}}}\right)\left(\frac{\sqrt{1 + \delta_2}\sqrt{\frac{T-t^*+1}{2}} + \sqrt{1 + \delta_{T+t^*+1}}}{1 - \delta_{T+t^*+1}\left(\sqrt{\frac{T-t^*+1}{2}} + 1\right)}\right) + \frac{2}{\sqrt{1 - \delta_{3T}}}\right] \|e\|_2 .$$

For convenience, let $C_2'(T, t^*)$ denote the expression within the brackets. Bounding $\|e\|_2$ as before by

$$\|e\|_2 \leq \sqrt{1 + \delta_T} \left(\|x_{T^C}\|_2 + \frac{1}{\sqrt{T}} \|x_{T^C}\|_1 \right) + \|w\|_2$$

and setting $C_2(T, t^*) = \sqrt{1 + \delta_T} C_2'(T, t^*)$ gives the final result

$$\|x - \tilde{x}\|_2 \leq (1 + C_2(T, t^*)) \|x_{T^C}\|_2 + \frac{C_2(T, t^*)}{\sqrt{T}} \|x_{T^C}\|_1 + C_2(T, t^*) \|w\|_2$$

$$= (1 + C_2(T, t^*)) \|x - x_T\|_2 + \frac{C_2(T, t^*)}{\sqrt{T}} \|x - x_T\|_1 + C_2(T, t^*) \|w\|_2.$$

which was to be shown. □

The drawback of the last several propositions is that they all require a "per iteration" restriction on the RIP numbers. Ideally, one would have an all-encompassing result such as that in Propositions II.3 or II.4 that do not depend on t or t^*. Certainly, it is possible to naively define one by nelecting to use the fact that at iteration t, $\Phi^* r_t$ is zero on all t correct indices already identified by the algorithm. In addition, one could use the monotonicity of restricted isometry numbers to apply constant-valued upper bounds, e.g., bound δ_{T+t+1} by δ_{2T+1}. In the case of 2-OMP, this would convert Equation 2.6 into

$$\delta_{2T} < \frac{1}{1 + \sqrt{\frac{T}{2}}}.$$

Unfortunately, this simplification will actually makes regular OMP appear to be better, i.e., require a less-stringent RIP property, than the three variations. Instead, this work will take a different route. A consequence of one of the results in [45] is that the restricted isometry numbers of a dictionary matrix satisfy a sublinear constraint of the form $\delta_\ell \leq \ell \delta_2$. In other words, the growth of the RIP numbers cannot be faster than linear. For the sake of this analysis, it will be assumed that the restricted

isometry numbers take the form

$$\delta_\ell \leq C\delta_2 \ell^\beta$$

where C is a constant $\beta \in [0, 1]$ is a power growth parameter. With that in mind, it can be shown that that the sufficient condition for 2-OMP, i.e. (2.6), is satisfied if

(2.7)
$$\delta_2 < \frac{1}{(T+t+1)^\beta} \cdot \frac{1}{1 + \sqrt{(T-t+1)/2}}$$

where it is assumed without loss of generality that $C = 1$. As a side note, for a given measurement matrix, δ_2 can be determined by simply computing the $N(N-1)$ inner products of its columns. Calculating δ_T exactly for larger T is a problem that is not known to have a polynomial time solution. Thus, restricted isometry conditions on δ_2 can be argued to be useful in practice. Figure 2.2 shows plots of the above expression as a function of the iteration number t for different values of β when $T = 50$.
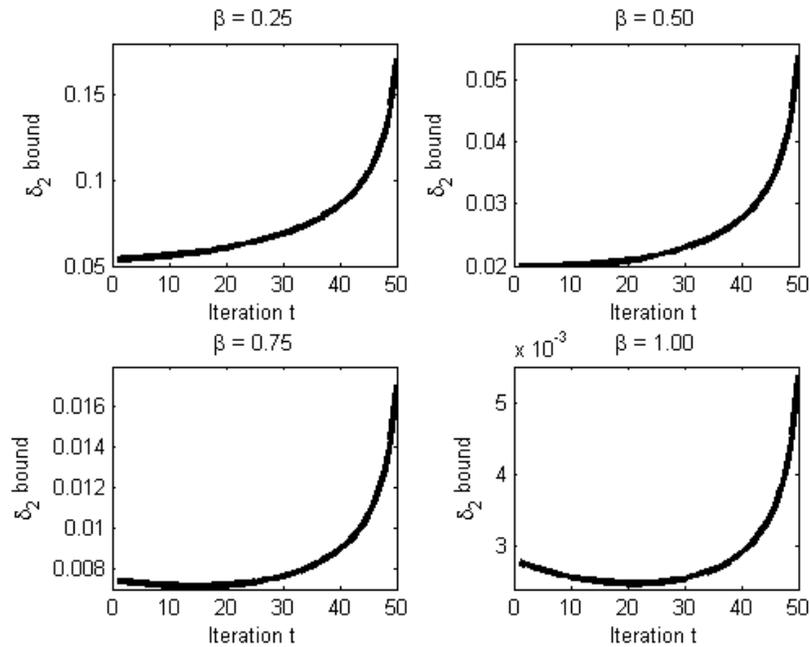
Figure 2.3: Restricted isometry bounds for 2-OMP with sparsity parameter $T = 50$.

These curves looks very similar if one varies the value of T. Based on the plots,

it is clear that the first few iterations of 2-OMP have a harder sufficient restricted isometry condition to satisfy than the final few iterations. In fact, if β is small enough, the very first iteration is the toughest of all. And, for larger values of β, the restricted isometry constraint on δ_2 for the first iteration is still a very good approximation of the restricted isometry condition required for all iterations to successfully identify at least one correct atom and no more than one incorrect atom. For the sake of mathematical rigor, it will be shown that the expression in (2.11) is monotone increasing with respect to t if $\beta < 0.5$.

Proposition II.11. *Suppose that $T \geq 2$. Then the expression*

$$(2.8) \qquad \frac{1}{(T+t+1)^\beta \left(1 + \sqrt{\frac{T-t+1}{2}}\right)}$$

is a monotone increasing function of t on the interval $(0, T]$. In particular, this quantity will attain its minimum value at the first iteration, i.e. $t = 1$.

Proof. The result will be proved by deriving a condition on β so that

$$f(t) := (T+t+1)^\beta \left(1 + \sqrt{\frac{T-t+1}{2}}\right)$$

will be monotone decreasing with respect to t. First differentiate f to obtain

$$f'(t) = \beta\,(T+t+1)^{\beta-1}\left(1 + \sqrt{\frac{T-t+1}{2}}\right) - \frac{1}{4\sqrt{\frac{T-t+1}{2}}}(T+t+1)^\beta.$$

Since only the sign of f' is important for this particular application, it suffices to work with the simpler equation

$$
\begin{aligned}
g(t) \;&:=\; 4\beta\sqrt{\frac{T-t+1}{2}}\left(1 + \sqrt{\frac{T-t+1}{2}}\right) - (T+t+1) \\
&=\; 4\beta\sqrt{\frac{T-t+1}{2}}\left(1 + \sqrt{\frac{T-t+1}{2}}\right) - (T-t+1) - 2t
\end{aligned}
$$

which carries the same sign as f' for all values of $t \in (0, T]$. Now make the substitution $u = \sqrt{T - t + 1}$ to obtain the quadratic expression

$$g(t, u(t)) = (2\beta - 1)u^2 + \frac{4\beta}{\sqrt{2}}u - 2t.$$

Now observe that for any value of $\beta < 1$,

$$
\begin{aligned}
g(T, u(T)) &= (2\beta - 1) + 2\sqrt{2}\beta - 2T \\
&\leq 1 + 2\sqrt{2} - 2T
\end{aligned}
$$

since $\beta < 1$. If $T \geq 2$, then this quantity must be less than zero. In other words, f is decreasing at $t = T$ and therefore (2.8) is increasing at this point. Now if it can be shown that g does not obtain any zeros on $(0, T]$, then it will be possible to deduce that f is decreasing everywhere on this interval. Since g is quadratic in u, its roots can be expressed as

$$
\begin{aligned}
u &= \frac{\frac{-4\beta}{\sqrt{2}} \pm \sqrt{8\beta^2 + 8(2\beta - 1)t}}{2(2\beta - 1)} \\
&= \frac{-\sqrt{2}\beta \pm \sqrt{2}\sqrt{\beta^2 + (2\beta - 1)t}}{(2\beta - 1)}.
\end{aligned}
$$

Observe that there will be a positive root in $(0, T]$ if $\sqrt{\beta^2 + (2\beta - 1)t} > \beta$. But this is equivalent to the statement $\beta > 0.5$. Thus, (2.8) will be an increasing function when $\beta < 0.5$. $\qquad\square$

The same phenomenon occurs for StOMP, and utilizing the same method, the required condition can also be shown to be $\beta < 1/2$.

To see if it is possible to gain any improvement over regular OMP, one can convert (2.2), (2.4), and (2.6) into restrictions on δ_2. Assuming β is sufficiently small ($\beta < 0.5$) so that the RIP property on the first iteration is the hardest to satisfy, it is possible to show the following:

Proposition II.12. *Suppose* Φ *is a dictionary matrix whose RIP numbers obey the growth restriction* $\delta_\ell < \delta_2 \ell^\beta$ *for* $0 \leq \beta < 0.5$. *Also suppose that* $\delta_{2T} < 1$. *Then the following statements are true after* T *iterations:*

1. OMP *will recover any* T-*sparse signal* x *exactly if*

$$(2.9) \qquad \delta_2 < \left(\frac{1}{T+1}\right)^\beta \cdot \frac{1}{1 + \sqrt{T}}.$$

2. STOMP *with the thresholds outlined earlier can recover any* T-*sparse signal* x *exactly if*

$$(2.10) \qquad \delta_2 < \left(\frac{1}{T+2}\right)^\beta \cdot \frac{1}{1 + \sqrt{T}/2^{1/4}}.$$

3. 2-OMP *will recover any* T-*sparse signal* x *exactly if*

$$(2.11) \qquad \delta_2 < \left(\frac{1}{T+2}\right)^\beta \cdot \frac{1}{1 + \sqrt{T/2}}.$$

This result provides a clue as to why in many cases, STOMP and 2-OMP will outperform OMP. To see this, consider the following table of sufficient δ_2 conditions for all three algorithms that will guarantee exact T-sparse reconstructions. For STOMP and 2-OMP, the percent improvement in the sufficient condition over ordinary OMP is presented as well.

Sparsity T	10^1	10^2	10^3	10^4	10^5
OMP	.1319	.0287	.0055	.0010	.0002
STOMP	.1468 (11%)	.0334 (17%)	.0064 (18%)	.0012 (19%)	.00021 (19%)
2-OMP	.1660 (26%)	.0390 (36%)	.0076 (40%)	.0014 (41%)	.0003 (41%)

Figure 2.4: Advantage of STOMP and 2-OMP over OMP for sparse signal recovery in terms of a restricted isometry condition on δ_2.

The percentages reflect the improvement of STOMP and 2-OMP over OMP.

In practice, it can also be observed that STOMP and 2-OMP do perform better than OMP if all three algorithms are run for T iterations. In addition, the runtimes

of STOMP and 2-OMP are typically much faster than that of OMP because, on average, one requires much fewer than T iterations for the residual to converge to zero. The stipulation of having T iterations in the above propositions is simply in place to cover worst case anomalies where a single incorrect atom is chosen at each iteration.

At this point, attention will be shifted to the case of general signal recovery. When initially discussing 2-OMP's ability to recover an arbitrary signal, a mixed ℓ_2-ℓ_1 performance guarantee was derived that depends on the stopping time t^* of 2-OMP's good iterations (where at least one correct atom is selected per iteration). Unfortunately, given a signal x and measurement matrix Φ, it is hard to calculate the value of t^* a priori. The result is that it is a challenge compare directly the performance of OMP against that of 2-OMP. This is the objective of what now follows: The general signal recovery results for OMP and 2-OMP can effectively by compared by simply comparing the expressions $C_1'(T)$ and $C_2'(T, t^*)$ respectively. In order to address the stopping time issue, assume again that the given restricted isometry numbers obey the growth restriction $\delta_\ell \leq \delta_2 \ell^\beta$ for some value $0 \leq \beta \leq 1$. Using this assumption, it is possible to find the following upper bounds on both of these quantities:

$$(2.12) \quad C_1(T) \leq \left(\frac{\sqrt{1 + \delta_2 T^\beta}}{\sqrt{1 - \delta_2 (2T)^\beta}} \right) \left(\frac{\sqrt{T} + \sqrt{1 + \delta_2 (T+1)^\beta}}{1 - \delta_2 (T+1)^\beta (1 + \sqrt{T})} \right) + \frac{2}{1 - \delta_2 (2T)^\beta}$$

and

$$
\begin{aligned}
C_2(T, t^*) \quad \leq \quad & \left(\frac{\sqrt{1 + \delta_2 (T + t^* + 1)^\beta}}{\sqrt{1 - \delta_2 (3T)^\beta}} \right) \left(\frac{\sqrt{1 + \delta_2} \sqrt{\frac{T - t^* + 1}{2}} + \sqrt{1 + \delta_2 (T + t^* + 1)^\beta}}{1 - \delta_2 (T + t^* + 1)^\beta \left(1 + \sqrt{\frac{T - t^* + 1}{2}} \right)} \right) \\
& + \frac{2}{\sqrt{1 - \delta_2 (3T)^\beta}}
\end{aligned}
$$

In order to obviate the need for calculating the stopping time, it will help to find an

upper bound on the function:

$$f(t) = \frac{1}{1 - \delta_2(T + t + 1)^\beta \left(1 + \sqrt{\frac{T-t+1}{2}}\right)} \sqrt{\frac{T - t + 1}{2}}$$

which occurs in $C_2'(T, t^*)$. This can be accomplished by finding a lower bound on the reciprocal function

$$g(t) = \left[1 - \delta_2(T + t + 1)^\beta \left(1 + \sqrt{\frac{T - t + 1}{2}}\right)\right] \sqrt{\frac{2}{T - t + 1}}.$$

Now by Proposition II.11, the bracketed term in the above expression is minimized at iteration $t = 1$ assuming $\beta < 0.5$. Furthermore, the second term is also minimized at $t = 1$. Now both instances of $\sqrt{1 + \delta_2(T + t + 1)^\beta}$ can be simply bounded by $\sqrt{1 + \delta_2(2T)^\beta}$. The result of applying these bounds is that:

$$
\begin{aligned}
C_2(T, t^*) &\leq \left(\frac{\sqrt{1 + \delta_2(2T)^\beta}}{\sqrt{1 - \delta_2(3T)^\beta}}\right) \left(\frac{\sqrt{1 + \delta_2}\sqrt{\frac{T}{2}} + \sqrt{1 + \delta_2(2T)^\beta}}{1 - \delta_2(T + 2)^\beta \left(1 + \sqrt{\frac{T}{2}}\right)}\right) \\
&+ \frac{2}{\sqrt{1 - \delta_2(3T)^\beta}}
\end{aligned}
$$

Let $\tilde{C}_2(T)$ denote the above expression and $\tilde{C}_1(T)$ refer to bound in (2.12). Then it is possible to derive the following performance guarantees related to the general performance of OMP and 2-OMP.

Proposition II.13. *Let x be any signal and Φ be a measurement matrix with restricted isometry numbers δ_ℓ that satisfy $\delta_\ell \leq \ell^\beta$ for some $\beta < 0.5$. If Condition 2.9 is met, then* OMP *will generate a T-term approximation $\tilde{x}_1$ that satisfies:*

$$(2.13) \quad \|x - \tilde{x}_1\|_2 \leq \left(1 + \tilde{C}_1(T)\right) \|x - x_T\|_2 + \frac{\tilde{C}_1(T)}{\sqrt{T}} \|x - x_T\|_1 + \tilde{C}_1(T) \|w\|_2.$$

Furthermore, if Condition 2.11 is met, then 2-OMP *will generate a $2T$ term approximation $\tilde{x}_2$ that satisfies:*

$$(2.14) \quad \|x - \tilde{x}_1\|_2 \leq \left(1 + \tilde{C}_2(T)\right) \|x - x_T\|_2 + \frac{\tilde{C}_2(T)}{\sqrt{T}} \|x - x_T\|_1 + \tilde{C}_2(T) \|w\|_2.$$

where in both expression, x_T represents the optimal T-term representation of x.

By carefully examining both error bounds, it is apparent that as long as the restricted isometry numbers do not grow too fast, then the result for 2-OMP is significantly stronger than that for OMP. Now one may argue that this is an unfair comparison because 2-OMP is allowed to select up to $2T$ atoms (with up to T of them being incorrect) whereas OMP is only allowed to choose T. However, if runtime is considered as the primary criterion for generating an approximation of x, 2-OMP is the real winner because T iterations of 2-OMP does not take significantly more time to compute than T iterations of OMP. Thus, one can obtain less error in roughly an equivalent amount of time. In addition, one can also truncate the $2T$-term approximation $\tilde{x}$ into a T-term representation $\tilde{x}_T$ without significantly affecting the error. Indeed, one can show that

$$(2.15) \qquad\qquad \|x - \tilde{x}_T\|_2 \leq 2\|x - \tilde{x}\|_2 + \|x - x_T\|_2.$$

As an illustration of the comparison performance of OMP and 2-OMP, Figure 2.5 shows a tabulation of the constant $\tilde{C}_1(T)$ and $\tilde{C}_2(T)$ for different sparsity levels T and fixed values of $\delta_2 = .00001$ and $\beta = 0.4$. The percent reduction error afforded by 2-OMP compared to OMP is also shown.

Sparsity T	10^1	10^2	10^3	10^4	10^5
$\tilde{C}_1(T)$ (OMP)	6.163	13.01	34.80	107.3	467.2
$\tilde{C}_2(T)$ (2-OMP)	5.237 (15%)	10.08 (23%)	25.45 (27%)	75.86 (29%)	292.1 (37%)

Figure 2.5: Advantage of 2-OMP over OMP for general signal recovery in terms of a restricted isometry condition on δ_2.

The objective of this section was to demonstrate that by allowing an orthogonalized matching pursuit algorithm to select two or more items per iteration, more favorable performance guarantees could be obtained. Of course, this cannot be done

in a sloppy fashion. We first relaxed the greedy selection criterion in 2-OMP by only insisting that, at every iteration, at least one correct atom and no more than one incorrect atom are chosen. Then we carefully bounded the resulting performance guarantees so that they no longer depend on any iteration numbers. In the following section, we perform a similar analysis, but for the case where a significantly larger number of atoms are selected at a time.

2.3   Increasing The Number of Atoms Per Iteration Even Further

This section addresses the question of what would happen if the number of atoms selected per iteration is increased further. The short answer is that one is more likely to select all the correct atoms; however, many incorrect atoms will be selected as well. If too many incorrect atoms are chosen, then it may not be possible to recover a T-sparse signal x even if all the correct atoms are chosen. If S represents the number of incorrectly chosen atoms by the algorithm, then it would be required that $\delta_{T+S} < 1$ to guarantee that a least squares procedure will be successful in reconstructing x.

The above question will now be more precisely answered by introducing another algorithm which will be referred to as Hybrid Orthogonal Matching Pursuit (HYBRID α-OMP). It is basically the same as regular OMP except for the fact that at iteration t, it will select $\alpha(T - t + 1)$ atoms. Assuming that one correct atom is selected per iteration, then this is equivalent to selecting the fraction α of all remaining correct atoms in the next iteration. The pseudocode for this algorithm is shown below in Figure 2.6.

This algorithm is similar to the other OMP variations in that all the algorithms maintain a growing list of atoms that they believe are correct. If α is very large, HYBRID α-OMP begins to resemble the more globalized algorithms Iteration Thresh-

```
          Algorithm:  Hybrid Orthogonal Matching Pursuit

Inputs: Φ, y = Φ, Sparsity T, Fraction α
Outputs: αT(T + 1)/2 + O(1) term approximation x̃ to x
```

Initialize $r_0 = y$, $\Lambda = \emptyset$
For t from 1 to T {
    Set

$$A_t = \{\text{Top } \alpha(T - t + 1) \text{ values of } i \text{ that maximize } |(\Phi^* r_{t-1})_i|\}$$

    Set $\Lambda_t = \Lambda_{t-1} \cup A_t$
    Set

$$a_t = \underset{a}{\arg\min} \, \|\Phi_{\Lambda_t} a - y\|_2 .$$

    $r_t = y - \Phi_{\Lambda_t} a_t$
    }
Set the entries of x̃ corresponding to indices Λ_T equal to a_T. Set all
other entries equal to zero.

Figure 2.6:   Pseudocode for HYBRID α-OMP

olding and CoSAMP except for the fact that it does not allow selected atoms to be deleted from the list.

Now consider the following proposition related to HYBRID α-OMP's ability to recover exactly T-sparse signals.

Proposition II.14. *Let Φ be a measurement matrix and let α be any positive real number. Then, given any T-sparse signal x, HYBRID α-OMP will recover at least one correct atom per iteration (and therefore recover the signal) provided that*

$$(2.16) \qquad \delta_{T+(\alpha T-1)t+1} < \frac{1}{1 + \sqrt{\frac{T-t+1}{\alpha(T-t+1)}}} = \frac{1}{1 + \sqrt{1/\alpha}}.$$

Proof. The proof uses the same methods as before. $\qquad\square$

At first glance, this result seems spectacular because it eliminates the $\sqrt{T}$ factor in the denominator of the RIP. In plain English, this means that the algorithm is much more likely to recover all the correct atoms representing a signal x. However, the downside is that it come at the expense of a restricted isometry condition (that

guarantees the correctness of the least squares procedure) given by

$$\delta_{\alpha T(T+1)/2} < 1$$

Thus, for a typical dictionary Φ, it seems that one must tradeoff the ability of a matching pursuit algorithm to select the correct non-zero entries of a signal with its ability to recover the signal by projecting its measurements onto the selected atoms.

A similar tradeoff occurs if one attempts to use the HYBRID α-OMP algorithm for general signal recovery. Using the same methods as for 2-OMP, the following can be proved:

Proposition II.15. *Let x be a general signal and let Φ be a dictionary matrix whose RIP numbers satisfy (2.16). Given noisy measurements of the form $y = \Phi x + w = \Phi x_T + e$, HYBRID α-OMP will recover a $(\hat{T} = \alpha T(T+1)/2 + O(1))$-term approximation $\tilde{x}$ satisfying*

$$\|x - \tilde{x}\|_2 \le (1 + C_\alpha) \|x - x_T\|_2 + \frac{C_\alpha}{\sqrt{T}} \|x - x_T\|_1 + C_\alpha \|w\|_2,$$

where $C_\alpha = \sqrt{1 + \delta_T} C'_\alpha$ and

$$C_\alpha = \left(\frac{\sqrt{1 + \delta_{T'}}}{\sqrt{1 - \delta_{\hat{T}+T}}} \right) \left(\frac{\sqrt{1 + \delta_{\alpha T}}\sqrt{\frac{1}{\alpha}} + \sqrt{1 + \delta_{T'}}}{1 - \delta_{T'}\left(1 + \sqrt{\frac{1}{\alpha}}\right)} \right) + \frac{2}{\sqrt{1 - \delta_{\hat{T}+T}}}.$$

Here $T' = T + (\alpha T - 1)t$ and $\tilde{x}$ is the best T-term approximation of x.

This version of orthogonal matching pursuit finally possesses the benchmark performance guarantee of the form $\|x - \tilde{x}\| \le O(\|x - x_T\|_2 + (1/\sqrt{T})\|x - x_T\|_1)$ that one typically enjoys with the convex optimization approaches discussed in Chapter I. In addition, as stated in the previous section, $\tilde{x}$ can be truncated to a T-term approximation without affecting the above guarantee by any more than a constant factor (see Equation 2.15).

The drawback to this method is that the processing of so many atoms may lead to instability. This is a problem that is also shared by Iterative Thresholding and CoSAMP. In the following section, we empirically compare the accuracy, efficiency, and stability of all the compressive sensing algorithm discussed in this work.

2.4   Empirical Evaluation of Algorithms

An observation that one will quickly make when running experiments utilizing compressive sensing algorithms is that they all work better than predicted by their respective theoretical guarantees. In other words, the restricted isometry properties of Chapter I Section 1.2 and this chapter provide very weak sufficient conditions regarding when some algorithm can exactly recover any signal with a given number T of non-zero entries. The reason for this is that RIPs provide worse case scenarios that may not appear often in practice.

In order to address this issue, much work has been done in performing "average-case" analyses on compressive sensing algorithms (see [59], [17], etc.). In these works, theoretical results are obtained regarding the various algorithms' performance in recovering commonplace sparse signals, e.g. with Gaussian or Bernoulli coefficients. For our purposes, we will empirically perform a similar analysis by designing several experiments which are shown below.

In the first experiment, for every sparsity level T from 4 to 52 in increments of 4, the following test was repeated 100 times: A T-sparse Gaussian signal of length 256 was generated and measurements of the form Φx were collected where Φ is a 100×256 Gaussian random matrix (selected differently each time). Then the algorithms OMP, 2-OMP, Hybrid 0.2-OMP, CoSAMP$_2$, CoSAMP$_1$, Iterative Thresholding, and Basis Pursuit were all used to attempt to recover x. Both versions of CoSAMP were

run using 10 iterations. For Iterative Thresholding, the Hard Thresholding routine in the Sparsify MATLAB package [2] was used with all parameters being selected optimally by the program. We used the L1-Magic package [48] for Basis Pursuit with the default settings which have been found to work well in our experiments. The two performance criteria evaluated were the probability of exact reconstruction (within a 1% tolerance for relative error) and the runtime. Plots of the results are shown below in Figures 2.7 and 2.8.
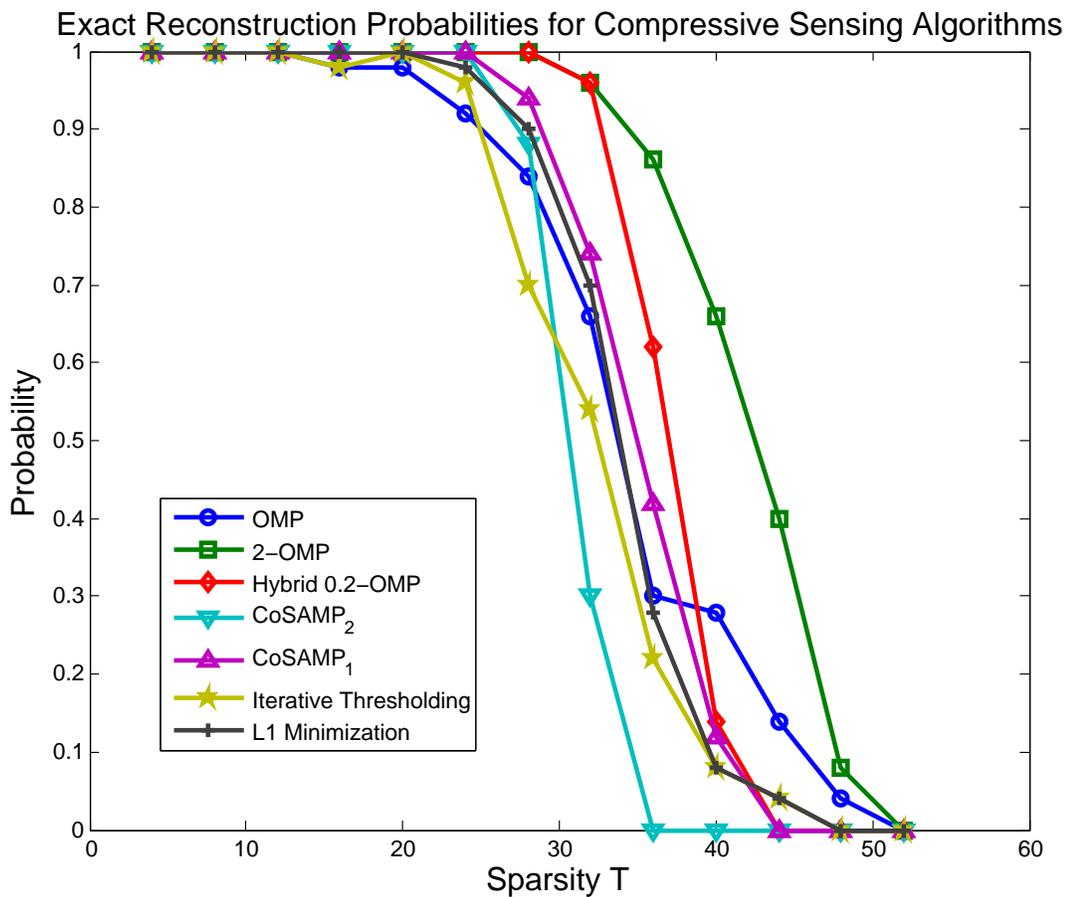
Figure 2.7: Probability of exact reconstruction of T-sparse signals using various compressive sensing algorithms.

In terms of exact reconstruction probability, Basis Pursuit did slightly better than OMP. However, the modifications proposed in Section 2.2 came in quite handy because 2-OMP and HYBRID 0.2-OMP both outperformed Basis Pursuit. Thus, the
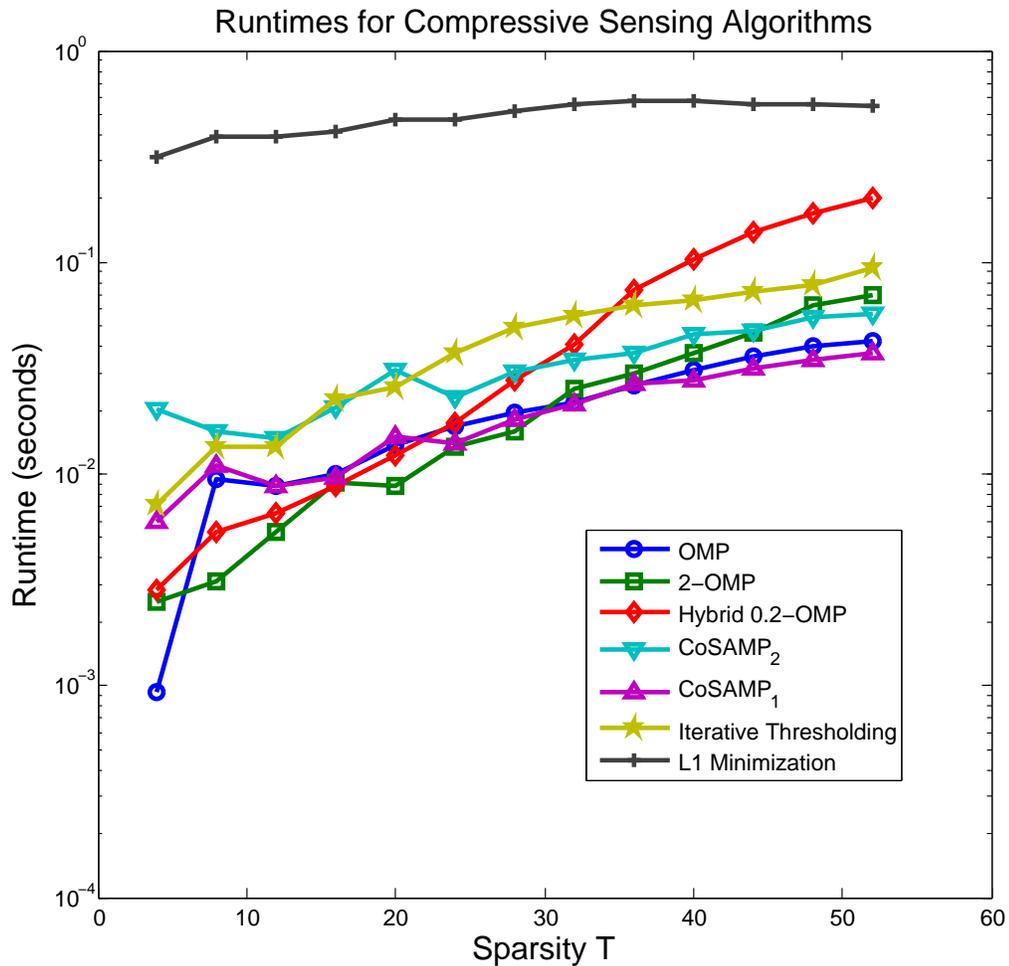
Figure 2.8: Runtimes of various compressive sensing algorithms when recovering T-sparse signals.

suggestion of allowing multiple atoms to be selected per iteration was exactly what was needed to give OMP the extra boost to put them on top. COSAMP_1 performed better than COSAMP_2 and Iterative Thresholding fell roughly in between in this particular experimental setup. When comparing the four algorithms possessing RIP sufficient conditions of the form $\delta_{O(T)} < c$, HYBRID 0.2-OMP was the clear winner with respect to accurately recovering exactly sparse signals.

With respect to runtime, all of the algorithms were very fast with the exception of convex optimization. These algorithms took no more than a tenth of a second to

run whereas ℓ_1 minimization took about a half of a second. The overall conclusion of this experiment is that the 2-OMP and HYBRID 0.2-OMP algorithms were the best overall performers.

Of course, this experiment only compared the various compressive sensing algorithms in terms of their accuracy in recovering exactly sparse signals. In the following experiment, the objective signals will not strictly possess this "nice" property. Here, 20 versions of a length 256 signal were generated with exponentially decaying coefficients in random locations that follow the decay rate $|x_{(n)}| \leq .9^n$. These 20 signals were reconstructed using the same algorithms as before with sparsity parameters varying from $T = 4$ to $T = 52$ in steps of four. Figure 2.9 below shows the various average ℓ_2 reconstruction errors produced by these algorithms for the various sparsity parameters T.

The astute reader will immediately notice that for smaller values of T, 2-OMP, HYBRID 0.2-OMP, and ℓ_1 Minimization all produce errors that are smaller than the optimal T-term error. This is not any sort of bug: the reason for this phenomenon is because these three algorithms produce reconstructions that are more than T-sparse for a particular sparsity parameter T. For those interested in comparing only strictly T-sparse approximations, we reproduce the same plot, except with all reconstructions truncated to T-terms, in Figure 2.10.

In terms of strictly T-sparse representations, observe that all the algorithms hugged the optimal error curve for $T < 24$. The interesting variations in performance seem to happen after this point. OMP's accuracy is not the greatest; however, it is the most stable. Both variations of CoSAMP, Iterative Thresholding, and HYBRID .2-OMP all perform similarly and experience a breakdown in performance in the $T = 24$ to $T = 28$ range. Out of these four, HYBRID .2-OMP was the most accu-
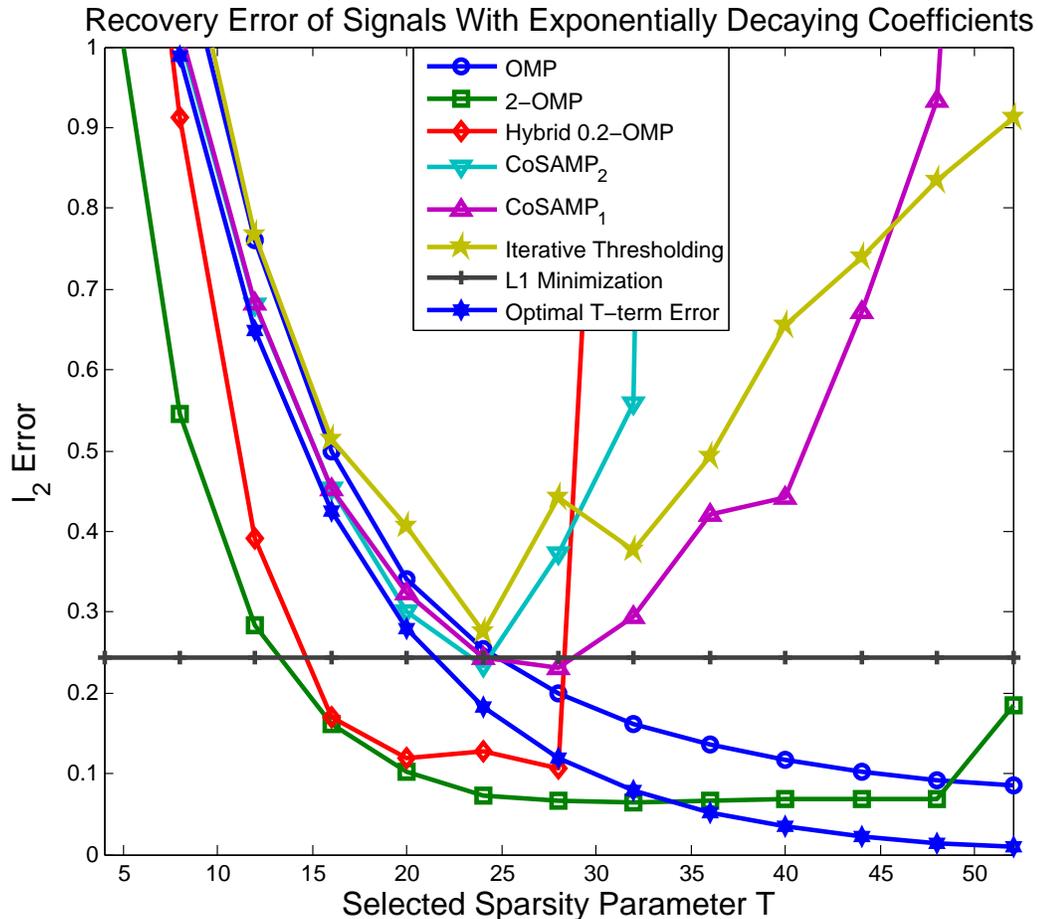
Figure 2.9: Average reconstruction errors in recovering signals with exponentially decaying coefficient generated by the various compressive sensing algorithms as a function of the sparsity parameter T.

rate. Basis Pursuit was very stable; however, it did not converge to the optimal error curve. Selecting a smaller primal-dual tolerance value within the L1-Magic routine may slightly reduce this error; however, this will come at the expense of increased runtime. 2-OMP was the best overall performer in this experiment because it was stable until $T = 48$ and returned T-term reconstructions whose average errors sat very close to the optimal T-term error curve. Thus, 2-OMP is capable of performing very accurately while being very forgiving in terms of how one selects the parameter T.

The final experiment tests the asymptotic runtimes of all the algorithms. For every
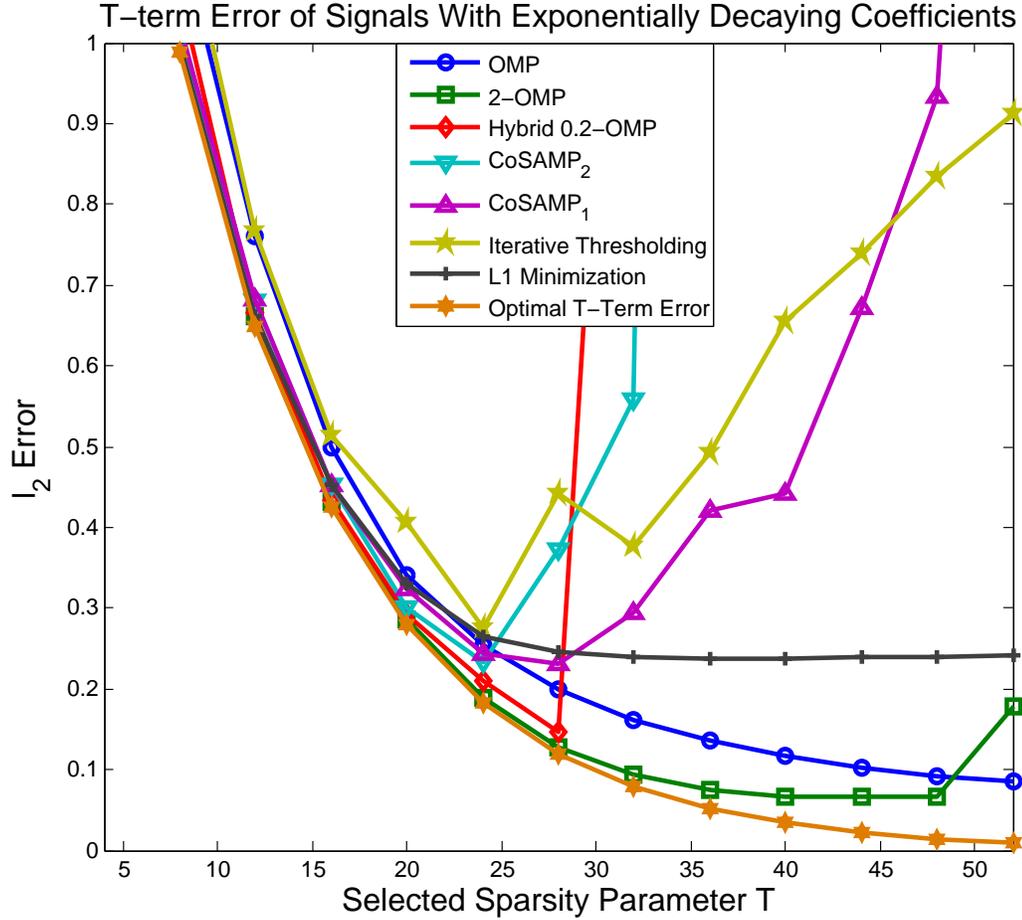
Figure 2.10: Average T-term reconstruction errors in recovering signals with exponentially decaying coefficient generated by the various compressive sensing algorithms as a function of the sparsity parameter T.

signal size $N =$round(2^n) with $n \in \{8, 8.5, 9, \cdots, 12\}$, we let $T =$round$(N/100)$ and $M =$ceil$(4T \log(N))$. For each setting of n, repeat the following five times: Let Φ be an $M \times N$ Gaussian matrix and x be a T sparse signal with Gaussian entries. The 7 algorithms mentioned above were used to recover x from Φ_x and the runtimes were measured. It was ensured that every single reconstruction in this experiment was exact. The log-log plot of the average runtimes as a function of the signal length N are shown below in Figure 2.11.

To make sense of this data, we perform a linear regression on this log-log plot in order to obtain a power law of the form $O(N^b)$ describing the runtime of each
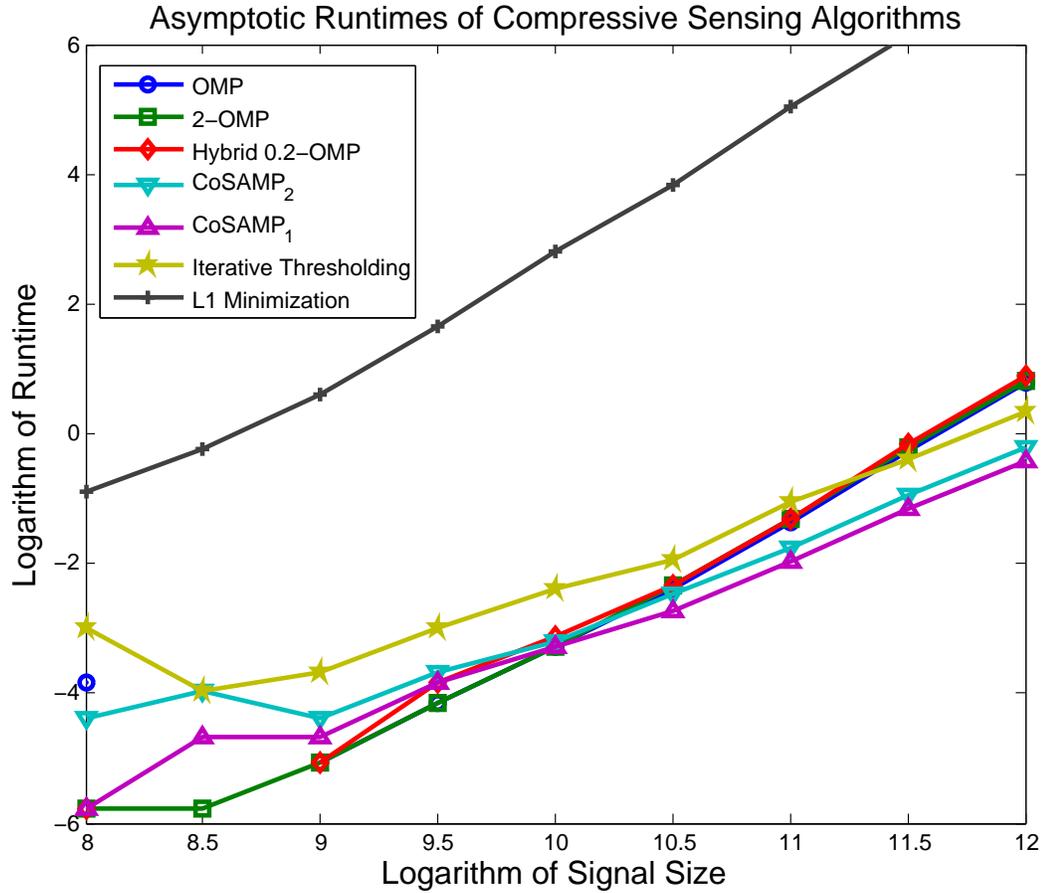
Figure 2.11: Log-log plot of the average runtimes of compressive sensing algorithms as a function of the signal length.

algorithm. The results are shown below in Figure 2.12

Algorithm	b	95% Confidence Interval
OMP	1.95	(1.84, 2.06)
2-OMP	1.97	(1.87, 2.06)
HYBRID .2-OMP	1.93	(1.77, 2.09)
CoSAMP$_2$	1.39	(1.29, 1.49)
CoSAMP$_1$	1.39	(1.28, 1.49)
Iterative Thresholding	1.33	(1.23, 1.43)
ℓ_1 Minimization	2.24	(2.19, 2.28)

Figure 2.12: Power law asymptotical analysis of compressive sensing algorithms.

The data suggests that ℓ_1 Minimization performed the worst asymptotically. OMP, 2-OMP, and HYBRID .2-OMP were slightly better, but not by much. The best performers in this particular setup were CoSAMP and Iterative Thresholding.

This is because these algorithms can converge in only a few iterations whereas OMP required $O(T) = O(.01N) = O(N)$ iterations. If T was fixed and did not depend on N, then it would be expected that OMP and its variants would perform asymptotically as well as CoSAMP and Iterative Thresholding. In all cases, this analysis provides empirical evidence that greedy algorithms perform asymptotically better than convex optimization methods in compressive sensing.

In summary, this chapter demonstrated that Orthogonal Matching Pursuit and its many variants enjoy performance guarantees based on restricted isometry properties. While the error bound for regular OMP is somewhat large and grows like $\sqrt{T}$, this growth can be stunted by allowing OMP to select more than one atom per iteration. In fact, if one selects a large fraction α of all atoms during every iteration, then these error bounds begin to look like the ideal Basis Pursuit standard. The downside is that selecting too many atoms will eventually lead to instability in the underlying least squares solver within the orthogonalized matching pursuit algorithm. This was seen empirically with a sudden, rapid rise in reconstruction error when the sparsity parameter T crossed a certain threshold. Pruning incorrect atoms, as is done in algorithms such as Iterative Thresholding and CoSAMP reduces the severity of the problem, but does not remove it entirely. Fortunately, as long as care is taken to make sure that T is not selected to be too large, orthogonal matching pursuit and its variants are fast and accurate tools for compressive sensing applications.

CHAPTER III

Medical Image Recovery using Gradient Orthogonal Matching Pursuit

3.1 Overview

In Chapter I, Section 1.4, the problem of recovering a sparse gradient image from a small set of Fourier measurements was briefly discussed. The primary motivation of this problem is to reduce the scanning time required in computer tomography or magnetic resonance imaging by allowing a reconstruction to be feasible with sub-Nyquist sampling rates. Constrained total variation minimization was the method initially suggested for solving this problem. Unfortunately, the savings in scanning time allotted by using fewer Fourier measurements do not necessarily justify the large runtime requirements of such convex optimization routines. While faster unconstrained solvers, such as FTVD (see [64]), can be adapted to solve this problem more efficiently, this chapter presents a completely different, novel approach based on the orthogonal matching pursuit paradigm. Now that theoretical compressive sensing performance guarantees have been proven for OMP and its variants (see Chapter II), its use in medical imaging can now be justified. While medical images are not inherently sparse, their edges are. Thus, this chapter will explore the use of orthogonalized matching pursuit algorithms for the purpose of edge recovery and answer questions regarding how to convert between the image domain and edge do-

main and vice versa. The various pieces of this puzzle can be put together to create an overall algorithm that is known as Gradient Matching Pursuit (GRADIENTMP). The previous works [39] and [37] chronicle the development of this algorithm from its initial conception; but here, this work will present a highly optimized version of the algorithm that competes very well with total variation minimization techniques in both CT- and MRI-based applications.

To begin, an overall outline of the Gradient Orthogonal Matching Pursuit algorithm is shown in Figure 3.1. The rest of this section will be devoted to explaining each step in much greater detail. As a note to the reader, the term xMP will refer to any greedy matching pursuit algorithm, e.g. OMP, K-OMP, CoSAMP, etc.

```
                    Algorithm:   GRADIENTMP
```

Inputs: -Set Ω of M frequencies containing the DC frequency $(0,0)$
          -Fourier observations $\mathcal{F}_\Omega X$ over Ω
          -Total variational sparsity T of X
          -Any other necessary parameters for xMP
Output: -An approximation $\tilde{X}$ of X

-Obtain Fourier observations of the vertical and horizontal edges of X
  via

$$(\mathcal{F}_\Omega \partial_x X)_k = (1 - e^{-2\pi i \omega_{x,k}/N})(\mathcal{F}_\Omega X)_k$$
$$\text{and}$$
$$(\mathcal{F}_\Omega \partial_y X)_k = (1 - e^{-2\pi i \omega_{y,k}/N})(\mathcal{F}_\Omega X)_k.$$

-Use xMP to obtain estimates $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ from $\mathcal{F}_\Omega \partial_x X$ and $\mathcal{F}_\Omega \partial_y X$.
-Use some inverse filter to obtain a final reconstruction $\tilde{X}$ of X from
  $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$.

Figure 3.1:   Outline of the GRADIENTMP algorithm.

3.2   Edges From Fourier Measurements

To motivate the technique behind GRADIENTMP, observe that the edge image of a T-sparse in gradient image is $O(T)$-sparse. Now suppose one has some operator

Ψ that acts as an edge detector. Then in theory, it should be possible to recover the edges ΨX using a fast greedy matching pursuit type algorithm. However, selecting such an appropriate edge detector and taking this approach poses some challenges.

1. It must be possible to obtain Fourier observations of the edges $\mathcal{F}_\Omega \Psi X$ from $\mathcal{F}_\Omega X$.

2. Once one recovers $\widetilde{\Psi X}$ from $\mathcal{F}_\Omega \Psi X$, it must be possible to "invert" Ψ to obtain an estimate $\tilde{X} = \Psi^{-1} \widetilde{\Psi X}$?

3. It must be true that $||\Psi X||_0 = O(||TV(X)||_0)$ in order to maintain the asymptotic runtime guarantees of the underlying greedy algorithm.

Certainly, if Ψ is the non-linear total-variational operator defined previously in Equation 1.19, it will be a daunting task to overcome these challenges. On the other hand, if Ψ is a directional derivative, e.g. $(\partial_x + \partial_y)/\sqrt{2}$, then all the above conditions can be satisfied and one can obtain an algorithm known as Differential Matching Pursuit (DIFFMP), see [36].

Assuming that the matching pursuit algorithm correctly identifies the locations of the edge pixels and estimates the directional derivative at these locations exactly, then DIFFMP will output an exact reconstruction of the original image. However, if there is mistake in the edge detection step, such as an incorrectly chosen edge pixel, then the error in the final reconstruction will look like a blur in the direction of the derivative. This is because of the integration required to invert Ψ.

To alleviate this problem, instead of working with one directional derivative, one can work with two derivatives in orthogonal directions. For convenience, these will be the vertical and horizontal directions, which yield the derivative operators ∂_x and ∂_y respectively. Before going any further, Figure 3.2 presents the famous $256 \times$

256 Shepp-Logan phantom image along with its reconstructions using one and two directional derivatives from approximately 6.6% of its Fourier Transform.

Original Phantom   DIFFMP Reconstruction   GRADIENTMP Reconstrction
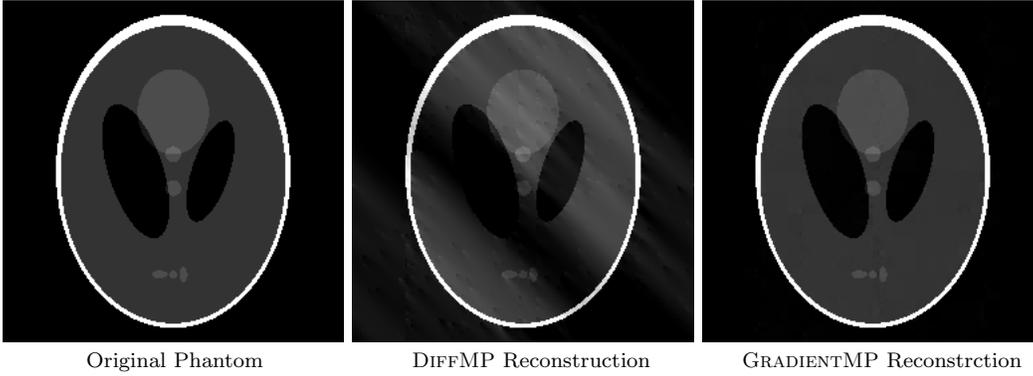
Figure 3.2: Comparison of DIFFMP and GRADIENTMP reconstructions of a Shepp-Logan phantom from 6.6% of its Fourier coefficients.

Clearly one can benefit greatly by using two partial derivatives instead of one.

Now given Fourier observations $\mathcal{F}_\Omega X$ over some set of frequencies Ω, one can obtain the Fourier observations of $\partial_x X$ and $\partial_y X$ over Ω via the equations:

$$(3.1) \qquad (\mathcal{F}_\Omega \partial_x X)_k = (1 - e^{-2\pi i \omega_{x,k}/N})(\mathcal{F}_\Omega X)_k$$

$$(3.2) \qquad (\mathcal{F}_\Omega \partial_y X)_k = (1 - e^{-2\pi i \omega_{y,k}/N})(\mathcal{F}_\Omega X)_k$$

After this is done, any one of many greedy pursuit algorithms can be used to recover $\partial_x X$ and $\partial_y Y$ from their respective Fourier observations. This work will primarily consider the use of K-fold Orthogonal Matching Pursuit (KOMP), which was discussed in Chapter II. Given a T-sparse in gradient image X, running T iterations of regular Orthogonal Matching Pursuit (OMP) can be somewhat time consuming if T is fairly large. Running KOMP for only $O(T/K) + O(1)$ iterations can produce almost identical results if K is not chosen too large. Choosing K too large will result in the selection of many incorrect atoms. The use of iterative thresholding and CoSAMP is avoided as these algorithms do not perform well in this particular application. This issue will be discussed a bit further in Section 3.5.

If the greedy algorithm utilized in GRADIENTMP recovers all of the edges correctly, then all the other steps of the overall algorithm will return a perfect reconstruction. Thus, the performance guarantees of these algorithms give us sufficient conditions to ensure that GRADIENTMP will perform accurately. What is more interesting are the cases where the edge reconstruction is not perfect. Then it is the inverse filter's duty to do everything possible not to blow up the magnitude of the already existing error any further. Designing such a nearly stable filter is a non-trivial task that will be discussed in the next section.

3.3 Integration of Edge Images

3.3.1 Naive Integration in the Vertical or Horizontal Direction

After obtaining estimates $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ of $\partial_x X$ and $\partial_y X$ respectively, some kind of integration must be performed to recover an estimate $\tilde{X}$ of X. One way to do this is to integrate against one image and to use the other image as a correction factor. To this end, suppose one integrates $\widetilde{\partial_x X}$ with respect to the vertical direction. This yields:

$$(3.3) \qquad \tilde{X} = \int \widetilde{\partial_x X} \, dx + H_y$$

where H_y is an image that consists of constant valued vertical stripes. H_y can be estimated very easily if one has a boundary value condition on the image X of the form

$$(3.4) \qquad X_{1,m} = h_m$$

where $h \in \mathbb{C}^N$ is a vector of pixels representing the top row of X. Then assuming the integration is carried out starting from $n = 1$, one gets that $(H_y)_{n,m} = h_m$. Thus, given the assumption of zero-valued boundary conditions, then $H_y = 0$ and no column-wise discrepancy in contrast will be present.

Unfortunately, it is often very impractical to assume that one has boundary condition on an image. For example, in an MRI scanning application, it would be impossible to determine $X_{1,m}$ a priori. To remedy this situation, one may assume periodic boundary conditions and work in the Fourier Domain. In this case, instead of requiring knowledge of $X_{1,m}$ for all m, it will be required that $(0, \omega_y) \in \Omega$ for all possible values of ω_y. In other words, this requirement states that the Fourier Transform of X is known on all frequencies that make up the horizontal axis of the frequency space. This yields the following (Fourier Domain) representation of Equation 3.3:

$$(3.5) \qquad \mathcal{F}\tilde{X} = \begin{cases} \mathcal{F}\widetilde{\partial_x X}/(1 - e^{-2\pi i \omega_1/N}) & \text{if } \omega_1 \neq 0 \\ (\mathcal{F}_\Omega X)_{\omega_1, \omega_2} & \text{if } \omega_1 = 0 \end{cases}$$

A similar formulation can be derived for integration in the horizontal direction. Unfortunately, this method only utilizes information from one set of the edge estimates (vertical and horizontal), and therefore, produces poor results when the edge estimates are not exact. More sophisticated techniques of inverse filtering the edge images are discussed in the next few sections.

3.3.2 Haar Wavelet Post Processing Technique

Of course, if $\widetilde{\partial_x X} = \partial_x X$, then one will have that $\tilde{X} = X$. However, because integration is an unstable operation, any errors in the edge recovery step will blow up in magnitude. This is why it is important to also utilize a horizontal estimate of the edges and do something a bit more intelligent. From here on, $\tilde{X}_v$ will represent the result of integrating $\widetilde{\partial_x X}$ in the vertical direction and $\tilde{X}_h$ will be the analogue reconstruction in the horizontal direction. Assuming each edge estimates are flawed in some way, the both $\tilde{X}_v$ and $\tilde{X}_h$ will exhibit streakiness. Examples of such vertically and horizontally streaked images are shown in Figure 3.3. A 256 × 256 Shepp-

Logan phantom was recovered from a small subset of its Fourier Coefficients utilizing the naive integration discussed in the previous section. The images with minor streaking were generated from 4.5% of all available Fourier Coefficients while the images with severe streaking utilized only 2.5% of all Fourier data. The edges were recovered utilizing 22 iterations of KOMP with $K = 100$. In situations where the
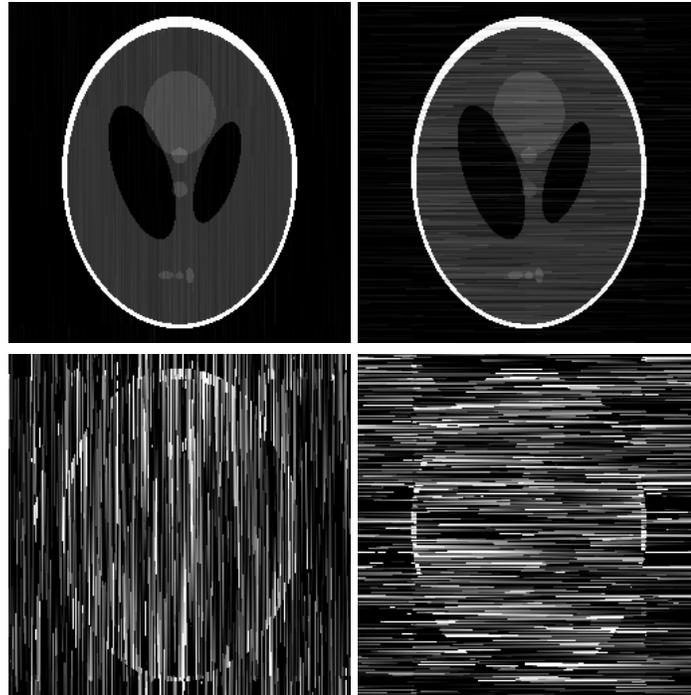
Figure 3.3: Examples of minor and severe streakiness in both the vertical and horizontal directions.

streaking effect is not too severe, it is apparent that $\tilde{X}_v$ retains some sharpness in the vertical direction whereas $\tilde{X}_h$ retains some sharpness in the horizontal direction. This suggests that one should be able to combine these two estimates to form a better overall estimate of the original image. To motivate a technique for doing this, consider the Haar Wavelet Transforms of $\tilde{X}_v$ and $\tilde{X}_h$ (in the minor streakiness case) shown in Figure 3.4 Every pixel in the representation of the full Haar Wavelet Transform of an image is ultimately based on a difference taken in the vertical, horizontal, or diagonal direction. In the case of the Haar Wavelet Transform of the vertically
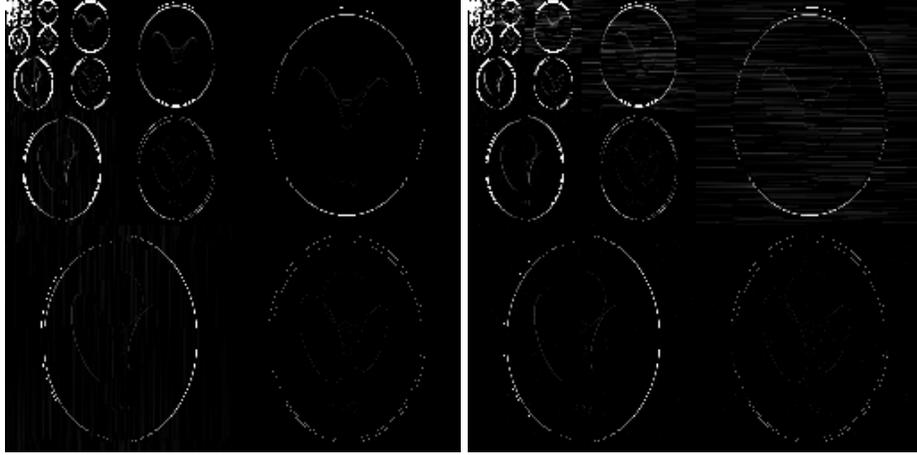
Figure 3.4: Haar Wavelet Transforms of Streaky Images.

streaked image, observe that only pixels corresponding to horizontal differences show any streakiness. In other words, it is possible to isolate the streakiness in the wavelet domain. The same principle applies to the horizontally streaked image. Thus, one can form the Haar Wavelet Transform of a streak-less image by defining

(3.6)

$$(\mathcal{H}\tilde{X})_{n,m} = \begin{cases} (\mathcal{H}\tilde{X}_v)_{n,m} & \text{if } (n,m) \text{ represents a vertical difference} \\ (\mathcal{H}\tilde{X}_h)_{n,m} & \text{if } (n,m) \text{ represents a horizontal difference} \\ ((\mathcal{H}\tilde{X}_v)_{n,m} + (\mathcal{H}\tilde{X}_h)_{n,m})/2 & \text{if } (n,m) \text{ represents a diagonal difference.} \end{cases}$$

where $\mathcal{H}$ represents the discrete Haar Wavelet Transform operator and $(\mathcal{H}\tilde{X})_{1,1}$ is simply taken to be the average of $(\mathcal{H}\tilde{X}_v)_{1,1}$ and $(\mathcal{H}\tilde{X}_h)_{1,1}$.

By applying this method to the first two streaky images in Figure 3.3, one obtains the reconstruction shown in Figure 3.5. Clearly, there is a significant improvement. The streakiness is gone. However, a little bit of blockiness is introduced.

It will be shown later in Section 3.4 that synthesizing the vertical and horizontal integrals of the edge estimates using this Haar Wavelet method is very nearly stable and only multiplies the edge reconstruction error by a small factor of $O(\sqrt{T \log N})$. This is one of two methods we utilized in the experiments that follow in this work.

Figure 3.5: Shepp Logan phantom with streakiness corrected by the Haar Wavelet method vs. the original phantom.

There is another similar method, however, which is also useful in situations where an image is strictly and highly sparse in the total-variational sense. It is presented here for completeness.

3.3.3 Cross-sectional Post Processing

Suppose that one runs GRADIENTMPon a 256×256 Shepp-Logan phantom X and obtains vertically and horizontally streaked reconstructions $\tilde{X}_v$ and $\tilde{X}_h$ respectively. Consider the 90th horizontal cross-sections of X, $\tilde{X}_v$, and $\tilde{X}_h$, which are shown in Figure 3.6: The cross-section of $\tilde{X}_v$ appears to be somewhat noisy, but on average, the values seem to approximate the cross-section of X well. However, because of the fluctuation, it is difficult to determine where the jump discontinuities occur in X just by examining $\tilde{X}_v$ alone. On the other hand, $\tilde{X}_h$ approximates the values of X poorly, but does a much better job identifying the jumps in X. Thus, for each cross-section, one can 'fix' the horizontally streaked image by estimating the jumps in X from $\tilde{X}_h$ and then estimating the value of X over each constant-valued interval by averaging the values of $\tilde{X}_v$ on those intervals. The same process can be carried out for vertical cross-sections. Results of this technique are shown in Figure 3.7; Clearly, this method
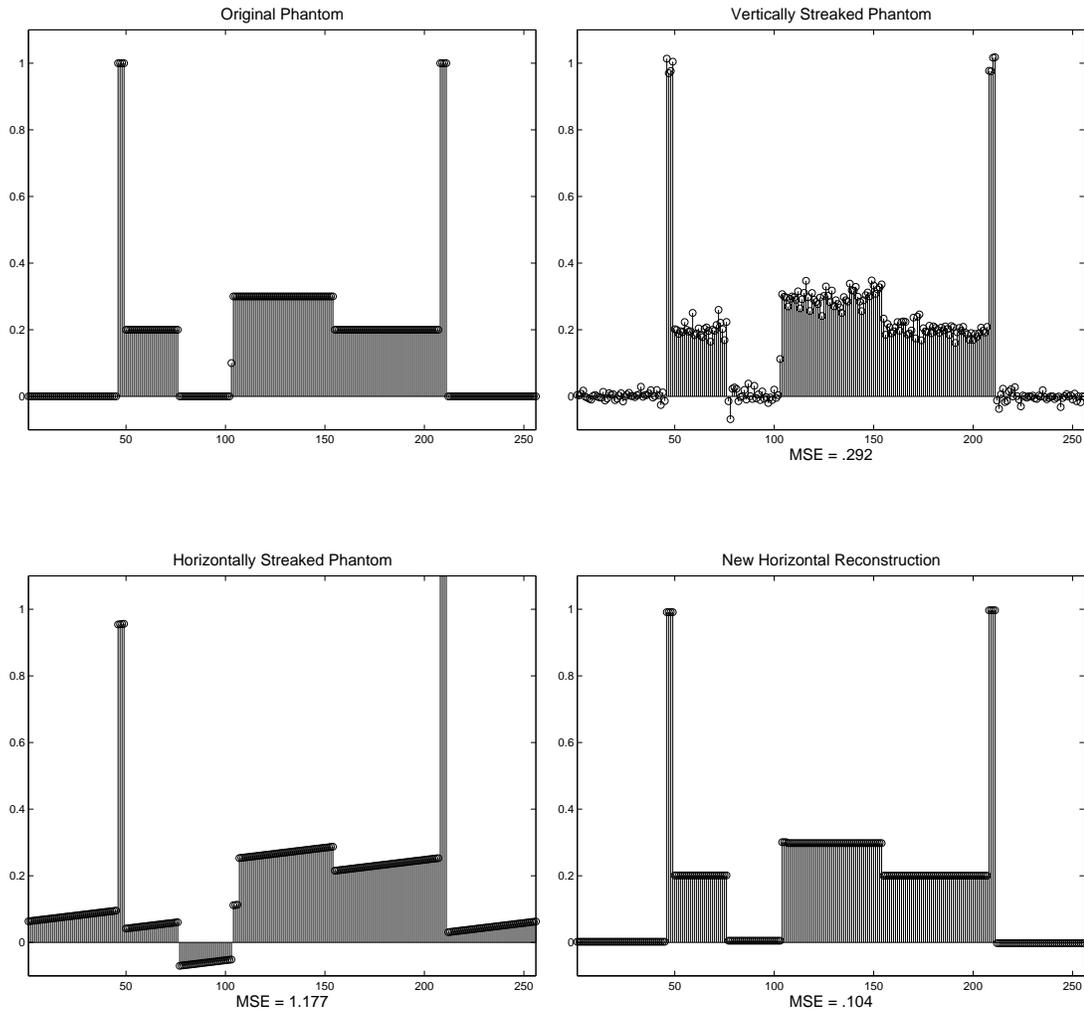
Figure 3.6: Cross-sections of orginal phantom, vertically streaked phantom, horizontally streaked phantom, and finally the reconstruction of the horizontally streaked image based on the cross-section method.

has greatly reduced the streakiness that was seen in Figure 3.3. Unfortunately, as suggested earlier at the end of Section 3.3.2, this method only works on images that are very sparse in gradient. Otherwise, the distances between jump discontinuities will be much smaller and so one would have fewer values over which to calculate an average. This will result in a loss of accuracy. Furthermore, these short distances will induce a lack of stationarity, which will cause the final reconstructed images to look very blurry. As a result, this method will not be considered any further in this
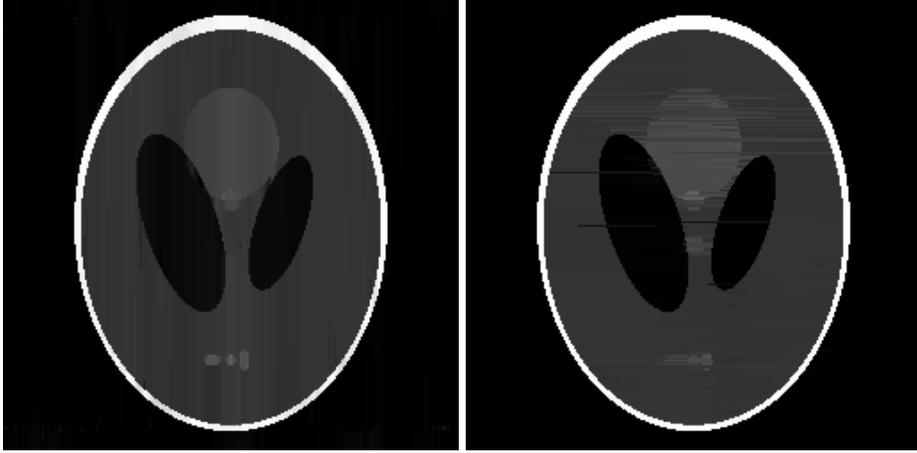
Figure 3.7: Reduction of streakiness using cross-section method.

work.

Both the Haar Wavelet and Cross-sectional methods perform post processing the vertically and horizontally integrated edge images in order to eliminate or reduce streakiness. The following section will illustrate a different technique that will eliminate the need to compute the integrated estimates $\tilde{X}_v$ and $\tilde{X}_h$ all together.

3.3.4   Least Squares Integration Method

The section will outline one more sophisticated way of generating a reconstruction $\tilde{X}$ of X given approximations $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ of $\partial_x X$ and $\partial_y X$ respectively. As initially motivated by [19], one can solve the convex optimization problem

$$(3.7) \quad \tilde{X} = \operatorname*{argmin}_{Y} \left\| \partial_x Y - \widetilde{\partial_x X} \right\|_2^2 + \left\| \partial_y Y - \widetilde{\partial_y X} \right\|_2^2 + \beta \left\| Y \right\|_{\mathrm{TV}} + \lambda \left\| \mathcal{F}_\Omega Y - \mathcal{F}_\Omega X \right\|_2^2$$

where β and λ are penalty parameters that determine the degrees to which the TV-minimization and Fourier constraints are enforced. While this approach will yield good results, solving optimization problems like (3.7) can be time consuming and difficult. This methodology also defeats the purpose of GRADIENTMP because the primary motivation for the algorithm was to eliminate the need for non-least squares optimization. To that end, one can replace the intrinsic ℓ_1 norm associated with the

TV-norm in Equation 1.20 with an ℓ_2 norm. This changes (3.7) into the following least squares optimization problem:

$$
\begin{aligned}
\tilde{X} \quad &= \quad \underset{Y}{\operatorname{argmin}} \left\| \partial_x Y - \widetilde{\partial_x X} \right\|_2^2 + \left\| \partial_y Y - \widetilde{\partial_y X} \right\|_2^2 \\
(3.8) \qquad &+ \quad \beta \left\| \partial_x Y \right\|_2^2 + \beta \left\| \partial_y Y \right\|_2^2 + \lambda \left\| \mathcal{F}_\Omega Y - \mathcal{F}_\Omega X \right\|_2^2 .
\end{aligned}
$$

While transitioning from an ℓ_1 norm to an ℓ_2 norm may seem to be a step backwards to both the compressive sensing (e.g. [9] and [8]) and noise reduction (e.g. [51]) communities, it can be shown that the solution of this least squares problem is typically better than pure integration in a single direction and at least empirically better than utilizing the Haar Wavelet method. Now observe that if hats are used to denote the Fourier Transform operator, it is possible to use Parseval's Theorem to rewrite (3.8) as the following equivalent problem in the Fourier domain:

$$
\begin{aligned}
\hat{\tilde{X}} \quad &= \quad \underset{\hat{Y}}{\operatorname{argmin}} \left\| \left(1 - e^{-2\pi i \omega_1 / N} \right) \hat{Y} - \widehat{\widetilde{\partial_x X}} \right\|_2^2 + \left\| \left(1 - e^{-2\pi i \omega_2 / N} \right) \hat{Y} - \widehat{\widetilde{\partial_y X}} \right\|_2^2 \\
(3.9) \qquad &+ \quad \beta \left(\left\| \left(1 - e^{-2\pi i \omega_1 / N} \right) \hat{Y} \right\|_2^2 + \left\| \left(1 - e^{-2\pi i \omega_2 / N} \right) \hat{Y} \right\|_2^2 \right) \\
&+ \quad \lambda \left\| \left(\hat{Y} - \hat{X} \right) \mathbf{1}_\Omega \right\|_2^2 .
\end{aligned}
$$

Here $\mathbf{1}_\Omega$ denotes an indicator function which is 1 on Ω and 0 otherwise. Based on this convenient alternative formulation of the problem, one can derive the following result:

Proposition III.1. *The least squares problem (3.9) can be solved element-wise by the following formula:*

$$
(3.10) \quad \hat{\tilde{X}}_{\omega_1, \omega_2} = \frac{\left(1 - e^{2\pi i \omega_1 / N} \right) \widehat{\widetilde{\partial_x X}}_{\omega_1, \omega_2} + \left(1 - e^{2\pi i \omega_2 / N} \right) \widehat{\widetilde{\partial_y X}}_{\omega_1, \omega_2} + \lambda \hat{X}_{\omega_1, \omega_2} \mathbf{1}_\Omega}{(1 + \beta) \left(\left| 1 - e^{-2\pi i \omega_1 / N} \right|^2 + \left| 1 - e^{-2\pi i \omega_2 / N} \right|^2 \right) + \lambda \mathbf{1}_\Omega} .
$$

Furthermore, if one lets $\lambda \to \infty$, then this solution will take the piecewise form

$$(3.11) \qquad \hat{\tilde{X}}_{\omega_1,\omega_2} = \begin{cases} \hat{X}_{\omega_1,\omega_2} & \text{if } (\omega_1,\omega_2) \in \Omega \\ \dfrac{\left(1-e^{2\pi i\omega_1/N}\right)\widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} + \left(1-e^{2\pi i\omega_2/N}\right)\widehat{\widetilde{\partial_y X}}_{\omega_1,\omega_2}}{(1+\beta)\left(\left|1-e^{-2\pi i\omega_1/N}\right|^2 + \left|1-e^{-2\pi i\omega_2/N}\right|^2\right)} & \text{otherwise} \end{cases}.$$

One can obtain $\tilde{X}$ by simply inverting the Fourier Transform. Now observe that if $\lambda \to \infty$, $\beta = 0$, and the edge approximations are exact, i.e. $\widetilde{\partial_x X} = \partial_x X$ and $\widetilde{\partial_y X} = \partial_y X$, then it follows that $\tilde{X} = X$. In general, selecting $\beta > 0$ will only attenuate the magnitude of any Fourier coefficients outside the set Ω. If one lets $\beta \to \infty$ (with $\lambda = \infty$, then the solution becomes equivalent to that obtained by naive Fourier back-projection, i.e. selecting $\tilde{X} = \mathcal{F}_\Omega^* \mathcal{F}_\Omega X$. This produces poor results. As a result, it is prudent to simply leave $\beta = 0$.

The next section will show that asymptotically speaking, the least squares method returns a reconstruction that is $O(N/r)$ times worse than the maximum edge error. Here, r is the radius of the largest circle centered at $(0,0)$ whose entire area is contained within the frequency set Ω. While this is worse than the result that will be derived for the Haar Wavelet Method; in practice, the least squares method actually performs better.

Figure 3.8 compares a reconstruction of the Shepp Logan phantom generated using the least squares method against the original image. This comparison uses the same frequency set Ω (with approximately 4.5% of all possibly frequencies selected) that was used previously to generate the streaky images and the Haar Wavelet corrected reconstruction.

If one compares this reconstruction to that in (3.5), it is apparent that the least squares generated image is better than the Haar Wavelet based image. In fact, the ℓ_2 reconstruction error is 1.27 utilizing least squares versus the 1.70 obtained using the wavelet method.

Figure 3.8: Shepp Logan phantom reconstruction using the least squares integration technique vs. the original phantom.

Of course, everything presented thus far is only empirical evidence regarding the capabilities of these methods. The following section will analyze each method more thoroughly and derive performance guarantees that will bound the factor by which any edge detection error from xMP can blow up.

3.4   Theoretical Performance

Suppose one uses xMP to obtain estimates $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ of $\partial_x X$ and $\partial_y X$ respectively utilizing some matching pursuit algorithm. Let $\tilde{X}_v$ and $\tilde{X}_h$ represent the respective antiderivatives of $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ and let $\tilde{X}$ represent the final reconstruction by applying one of the methods described in Section 3.3. In this section, it will be shown how the overall reconstruction error of GRADIENTMP is related to the edge reconstruction errors $||\widetilde{\partial_x X} - \partial_x X||_2$ and $||\widetilde{\partial_y X} - \partial_y X||_2$. More specifically, this section will demonstrate how integration alone is highly unstable and will blow up the edge error by a factor of $O(N)$, i.e. the image size. However, it will also be shown that by applying the proposed Haar wavelet post-processing technique, one can reduce this 'blow-up' factor to $O(\sqrt{T \log N})$. Furthermore, the Least squares method will be shown not to increase the edge reconstruction error by more than

$O(N/r)$ where r can be a large constant if Ω is chosen correctly. As a result, by using either the Haar Wavelet method or the Least Squares method, the overall algorithm becomes very nearly stable with respect to perturbations in the edge estimates.

Now to begin the analysis, the following proposition is related to the estimate $\tilde{X}_v$.

Proposition III.2. *The estimate $\tilde{X}_v$ possesses the following error guarantee:*

$$(3.12) \qquad ||\tilde{X}_v - X||_2 \leq \frac{N+1}{\sqrt{2}}||\widetilde{\partial_x X} - \partial_x X||_2$$

Proof. First observe that

$$(3.13) \qquad ||\tilde{X}_v - X||_2^2 = \sum_{n=1}^{N}\sum_{m=1}^{N}\left((\tilde{X}_v)_{n,m} - X_{n,m}\right)^2.$$

Rewriting this as an 'antiderivative' gives:

$$(3.14) \qquad = \sum_{n=1}^{N}\sum_{m=1}^{N}\left[\sum_{k=1}^{n}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)\right]^2.$$

Expanding the square yields:

$$(3.15) \qquad = \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)\left((\widetilde{\partial_x X})_{\ell,m} - (\partial_x X)_{\ell,m}\right).$$

Now apply the Cauchy-Schwartz inequality to obtain:

$$
\begin{aligned}
||\tilde{X}_v - X||_2^2 &\leq \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)^2. \\
&= \sum_{n=1}^{N}\sum_{m=1}^{N}\sum_{k=1}^{n}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)^2 n \\
&= \sum_{m=1}^{N}\sum_{k=1}^{N}\sum_{n=k}^{N}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)^2 n \\
&\leq \sum_{m=1}^{N}\sum_{k=1}^{N}\sum_{n=1}^{N}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)^2 n \\
&= \sum_{m=1}^{N}\sum_{k=1}^{N}\left((\widetilde{\partial_x X})_{k,m} - (\partial_x X)_{k,m}\right)^2 \frac{N(N+1)}{2} \\
&\leq \frac{(N+1)^2}{2}||\widetilde{\partial_x X} - \partial_x X||_2^2,
\end{aligned}
$$

which completes the proof. $\square$

The same guarantee holds for $\tilde{X}_h$ and the proof is almost exactly the same.

Unfortunately, this is a rather weak result. Because of the integration carried out, any error in the edge reconstruction gets smeared, resulting in an overall reconstruction error that is $O(N)$ times worse than the edge recovery error. This is where a post-processing step can be useful. By utilizing the Haar wavelet method, one can significantly improve upon this error.

Proposition III.3. *Let $\tilde{X}$ be the result of combining $\tilde{X}_v$ and $\tilde{X}_h$ by the Haar wavelet post-processing technique. Assuming that $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ are each no more than T-sparse, the error $||\tilde{X} - X||_2$ satisfies the following two inequalities:*

$$||\tilde{X} - X||_2 \ \leq \ \sqrt{T\left(\frac{5}{8}\log N + \frac{1}{2}\right)}\left(||\widetilde{\partial_x X} - \partial_x X||_2 + ||\widetilde{\partial_y X} - \partial_y X||_2\right)$$

$$||\tilde{X} - X||_2 \ \leq \ \sqrt{T\left(\frac{5}{2}\log N + 2\right)}\max\left(||\widetilde{\partial_x X} - \partial_x X||_2, ||\widetilde{\partial_y X} - \partial_y X||_2\right).$$

Proof. The first step is to introduce some specialized notation that will be needed to present a rigorous proof. First observe that every entry (except the one in the first row and first column) in the matrix representation of a full Haar Wavelet Transform corresponds to some vertical, horizontal, or diagonal differencing operation conducted on the original image. As an example, in the 8×8 case, one has the following representation:

$$\begin{pmatrix}
a & v & v & v & v & v & v & v \\
h & d & v & v & v & v & v & v \\
h & h & d & d & v & v & v & v \\
h & h & d & d & v & v & v & v \\
h & h & h & h & d & d & d & d \\
h & h & h & h & d & d & d & d \\
h & h & h & h & d & d & d & d \\
h & h & h & h & d & d & d & d
\end{pmatrix}$$

Figure 3.9: The directions of differences in an 8×8 Haar Wavelet expansion.

Here v, h, and d represent vertical, horizontal, and diagonal differences respectively and a is the overall averaging term. The next step is to define notation to represent the Haar wavelet transform restricted to certain types of differences. Superscripts can be used to accomplish this. For example, $\mathcal{H}^{d,v}$ will represent the Haar wavelet transform restricted to just diagonal and vertical differences. For some image X, the image $\mathcal{H}^{d,v}X$ will equal $\mathcal{H}X$ at all v and d entries and zero otherwise. Based on this notation, one can write $\tilde{X}$ as:

$$
(3.16) \qquad \tilde{X} = \mathcal{H}^{-1}\left[\mathcal{H}^v\tilde{X}_v + \mathcal{H}^h\tilde{X}_h + \frac{1}{2}\mathcal{H}^{a,d}\left(\tilde{X}_v + \tilde{X}_h\right)\right]
$$

$$
(3.17) \qquad = \mathcal{H}^{-1}\left[\left(\mathcal{H}^v + \frac{1}{2}\mathcal{H}^{a,d}\right)\tilde{X}_v + \left(\mathcal{H}^h + \frac{1}{2}\mathcal{H}^{a,d}\right)\tilde{X}_h\right].
$$

Next define the atomic $N \times N$ vertical 'streak' image $S^v_{a,b}$ pixel-wise as

$$
(3.18) \qquad S^v_{a,b}(n,m) = \begin{cases} 1 & \text{if } n \geq a \text{ and } m = b \\ 0 & \text{otherwise} \end{cases}
$$

In other words, this image is zero-valued except for a vertical line segment beginning at coordinate (a,b) and extending downwards to (N,b). An example of such an image is the left most image in Figure 3.10. Now examine the wavelet transform of $S^v_{a,b}$ restricted to vertical and diagonal differences. At any scale $j > 0$, at most two pixels of magnitude no greater than $1/2$ each will be introduced, one in the diagonal difference region and one in the vertical difference region. This implies that $||(\mathcal{H}^v + (1/2)\mathcal{H}^d)S^v_{a,b}||_2^2 \leq \frac{5}{16}\log N$. It is easy to see that the averaging term $|\mathcal{H}^a S^v_{a,b}| \leq 1$, and so

$$
(3.19) \qquad \left\|(\mathcal{H}^v + \frac{1}{2}\mathcal{H}^{a,d})S^v_{a,b}\right\|_2^2 \leq \frac{5}{16}\log N + \frac{1}{4}.
$$

The $\log N$ factor comes from the fact that there are $\log N$ scales in the full wavelet decomposition.

One can similarly define the horizontal streak image $S_{c,d}^h$. The same exact analysis as before also yields

$$(3.20) \qquad \left\|(\mathcal{H}^h + \frac{1}{2}\mathcal{H}^{a,d})S_{c,d}^h\right\|_2^2 \leq \frac{5}{16}\log N + \frac{1}{4}.$$

A graphical illustration of this analysis can be seen below in Figure 3.10. Notice how the image energy is reduced from $O(N)$ in the original atomic streak to $O(\sqrt{\log N})$ in the partial wavelet expansion shown on the right. The next major observation

$$S_{20,28}^v \qquad\qquad \mathcal{H}S_{20,28}^v \qquad\qquad (\mathcal{H}^v + 0.5\mathcal{H}^{a,d})S_{20,28}^v$$
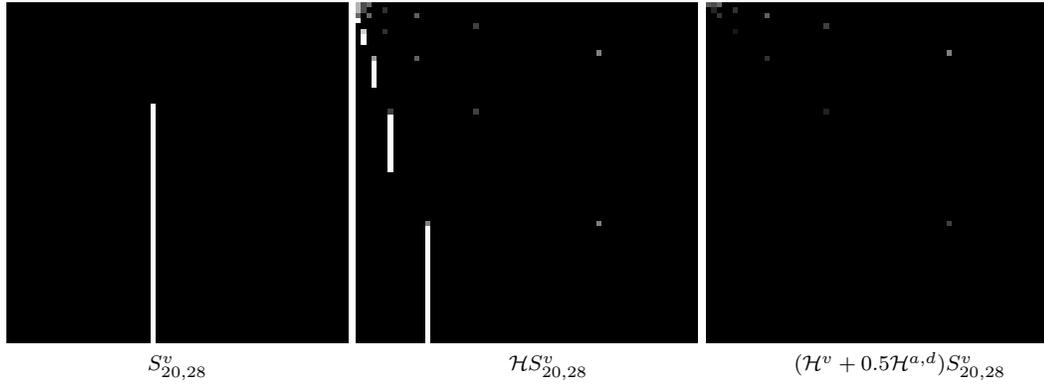
Figure 3.10: Single streak image along with its full Haar Wavelet Transform as well as its Haar Wavelet Transform restricted to vertical and diagonal differences only.

is that the errors $\tilde{X}_v - X$ and $\tilde{X}_h - X$ can be written as the following two linear combinations:

$$(3.21) \qquad\qquad \tilde{X}_v - X \; = \; \sum_{k=1}^{T_v} \alpha_k S_{a_k,b_k}^v$$

$$(3.22) \qquad\qquad \tilde{X}_h - X \; = \; \sum_{k=1}^{T_h} \beta_k S_{c_k,d_k}^h.$$

where $T_v \leq 2T$ and $T_h \leq 2T$ represent the respective sparsities of $\widetilde{\partial_x X} - \partial_x X$ and $\widetilde{\partial_y X} - \partial_y X$. The α_k's and β_k's represent the magnitudes of the nonzero positions of these errors, which occur at positions (a_k, b_k) and (c_k, d_k) respectively.

Now observe that

$$\mathcal{H}\left(\tilde{X} - X\right) = \left(\mathcal{H}^v + \frac{1}{2}\mathcal{H}^{a,d}\right)\left(\tilde{X}_v - X\right) + \left(\mathcal{H}^h + \frac{1}{2}\mathcal{H}^{a,d}\right)\left(\tilde{X}_h - X\right)$$

$$= \sum_{k=1}^{T_v} \alpha_k \left(\mathcal{H}^v + \frac{1}{2}\mathcal{H}^{a,d}\right) S^v_{a_k,b_k} + \sum_{k=1}^{T_h} \beta_k \left(\mathcal{H}^h + \frac{1}{2}\mathcal{H}^{a,d}\right) S^h_{c_k,d_k}$$

Now by combining this last equation with Inequalities 3.19 and 3.20, the triangle inequality, and the fact that $\mathcal{H}$ is unitary, one obtains that

$$\left\|\tilde{X} - X\right\|_2 \leq \sqrt{\left(\frac{5}{16}\log N + \frac{1}{4}\right)}\left[\sum_{k=1}^{T_v} |\alpha_k| + \sum_{k=1}^{T_h} |\beta_k|\right]$$

$$= \sqrt{\left(\frac{5}{16}\log N + \frac{1}{4}\right)}\left(||\widetilde{\partial_x X} - \partial_x X||_1 + ||\widetilde{\partial_y X} - \partial_y X||_1\right)$$

$$\leq \sqrt{2T\left(\frac{5}{16}\log N + \frac{1}{4}\right)}\left(||\widetilde{\partial_x X} - \partial_x X||_2 + ||\widetilde{\partial_y X} - \partial_y X||_2\right)$$

where the extra $2T$ factor comes from the fact that both edge error images are no more than $2T$-sparse. This completes the proof. $\square$

This is a significant improvement. In a nutshell, it states that the overall reconstruction error is only $O(\sqrt{T \log N})$ times worse than the edge recovery error introduced by the various orthogonal matching pursuit algorithms. This is asymptotically much better than $O(N)$. Additional noise can also be removed with soft-thresholding techniques such as wavelet shrinkage, which was introduced my Donoho et. al. in the mid 1990's (see [12, 14]). However, this topic will not be studied any further in this work.

The next objective is to prove a result with respect to the capabilities of the Least Squares integration method. With the choice of $\lambda = \infty$ and $\beta = 0$, one can show that the solution to (3.11) satisfies the following reconstruction performance guarantee.

Proposition III.4. *Given approximations $\widetilde{\partial_x X}$ and $\widetilde{\partial_y X}$ of $\partial_x X$ and $\partial_y Y$, then the*

solution $\tilde{X}$ of Equation 3.11 will satisfy:

$$\left\| \tilde{X} - X \right\|_2 \leq O\left(\frac{N}{\sqrt{k_1^2 + k_2^2}} \right) \max\left(\left\| \widetilde{\partial_x X} - \partial_x X \right\|_2, \left\| \widetilde{\partial_y X} - \partial_y X \right\|_2 \right).$$

where

$$(k_1, k_2) = \operatorname*{argmin}_{(\omega_1, \omega_2) \notin \Omega} \omega_1^2 + \omega_2^2.$$

Proof. Observe that for each $(\omega_1, \omega_2) \in \Omega$, $|\hat{\tilde{X}}_{\omega_1,\omega_2} - \hat{X}_{\omega_1,\omega_2}|^2 = 0$ by definition. Outside of Ω,

$$
\begin{aligned}
|\hat{\tilde{X}}_{\omega_1,\omega_2} - \hat{X}_{\omega_1,\omega_2}|^2 &= \left| \frac{\left(1 - e^{2\pi i \omega_1/N}\right) \widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} + \left(1 - e^{2\pi i \omega_2/N}\right) \widehat{\widetilde{\partial_y X}}_{\omega_1,\omega_2}}{|1 - e^{-2\pi i \omega_1/N}|^2 + |1 - e^{-2\pi i \omega_2/N}|^2} \right. \\
&\quad \left. - \frac{\left(1 - e^{2\pi i \omega_1/N}\right) \widehat{\partial_x X}_{\omega_1,\omega_2} + \left(1 - e^{2\pi i \omega_2/N}\right) \widehat{\partial_y X}_{\omega_1,\omega_2}}{|1 - e^{-2\pi i \omega_1/N}|^2 + |1 - e^{-2\pi i \omega_2/N}|^2} \right|^2 \\
&= \left| \frac{\left(1 - e^{2\pi i \omega_1/N}\right)}{|1 - e^{-2\pi i \omega_1/N}|^2 + |1 - e^{-2\pi i \omega_2/N}|^2} \left(\widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} - \widehat{\partial_x X}_{\omega_1,\omega_2} \right) \right. \\
&\quad \left. + \frac{\left(1 - e^{2\pi i \omega_2/N}\right)}{|1 - e^{-2\pi i \omega_1/N}|^2 + |1 - e^{-2\pi i \omega_2/N}|^2} \left(\widehat{\widetilde{\partial_y X}}_{\omega_1,\omega_2} - \widehat{\partial_y X}_{\omega_1,\omega_2} \right) \right|^2.
\end{aligned}
$$

Now utilize the fact that for any $a, b \in \mathbb{C}$, $|a + b|^2 = |a|^2 + |b|^2 + 2\mathrm{Re}(ab) \leq 4\max(|a|^2, |b|^2)$ and assume without loss of generality (so it isn't necessary to keep on writing out maximums) that

$$
\begin{aligned}
&\frac{|1 - e^{2\pi i k_1/N}|}{|1 - e^{-2\pi i k_1/N}|^2 + |1 - e^{-2\pi i k_2/N}|^2} \left| \widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} - \widehat{\partial_x X}_{\omega_1,\omega_2} \right| \\
&\geq \frac{|1 - e^{2\pi i k_2/N}|}{|1 - e^{-2\pi i k_1/N}|^2 + |1 - e^{-2\pi i k_2/N}|^2} \left| \widehat{\widetilde{\partial_y X}}_{\omega_1,\omega_2} - \widehat{\partial_y X}_{\omega_1,\omega_2} \right|
\end{aligned}
$$

to obtain that:

$$
\begin{aligned}
|\hat{\tilde{X}}_{\omega_1,\omega_2} - \hat{X}_{\omega_1,\omega_2}|^2 &\leq \frac{4 \left|1 - e^{-2\pi i k_1/N}\right|^2}{\left(|1 - e^{-2\pi i k_1/N}|^2 + |1 - e^{-2\pi i k_2/N}|^2\right)^2} \left| \widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} - \widehat{\partial_x X}_{\omega_1,\omega_2} \right|^2 \\
&\leq \frac{4}{|1 - e^{-2\pi i k_1/N}|^2 + |1 - e^{-2\pi i k_2/N}|^2} \left| \widehat{\widetilde{\partial_x X}}_{\omega_1,\omega_2} - \widehat{\partial_x X}_{\omega_1,\omega_2} \right|^2.
\end{aligned}
$$

Now simply use the Taylor expansion of the denominator and Parseval's Theorem to get the desired result. $\qquad\square$

As can be seen, the performance of this method depends on the selection of Ω. If Ω contains all the low frequencies within some radius $r = k_1^2 + k_2^2$, then the final reconstruction error will be $O(N/r)$ times worse than the maximum edge reconstruction error. This is much better than the $O(N/\sqrt{2})$ factor obtained with straight-forward integration. In general, if Ω contains mostly low frequencies, then this method will generate better results than if Ω contained mostly high frequencies. As a result, this "integration" is very appropriate in applications such as CT where Ω will consist of radial lines that congregate near the DC frequency. For the same reason, it may also be useful in MRI applications where the Fourier Space is sampled according to a spiral trajectory (see [32] and [57]).

The following corollary illustrates the stability that this method exhibits if all low frequencies below a certain threshold are included within the Fourier sampling frequencies Ω.

Corollary III.5. *Suppose that $\{(\omega_1, \omega_2)| - \frac{N}{6} \leq \omega_1 \leq \frac{N}{6} \text{ and } - \frac{N}{6} \leq \omega_2 \leq \frac{N}{6}\} \subseteq \Omega$ where all frequencies are given modulo N. Then the least squares method of integration will return a reconstruction such that*

$$\left\| \tilde{X} - X \right\|_2 \leq 2 \cdot \max \left(\left\| \widetilde{\partial_x X} - \partial_x X \right\|_2, \left\| \widetilde{\partial_y X} - \partial_y X \right\|_2 \right).$$

Now if $\{(\omega_1, \omega_2)| - \frac{N}{4} \leq \omega_1 \leq \frac{N}{4} \text{ and } - \frac{N}{4} \leq \omega_2 \leq \frac{N}{4}\} \subseteq \Omega$, then the Least Squares method will return a reconstruction such that

$$\left\| \tilde{X} - X \right\|_2 \leq \sqrt{2} \cdot \max \left(\left\| \widetilde{\partial_x X} - \partial_x X \right\|_2, \left\| \widetilde{\partial_y X} - \partial_y X \right\|_2 \right).$$

In general if $\{(\omega_1, \omega_2)| - \epsilon N \leq \omega_1 \leq \epsilon N \text{ and } - \epsilon N \leq \omega_2 \leq -\epsilon N\} \subseteq \Omega$, then the final reconstruction error will satisfy:

$$\left\| \tilde{X} - X \right\|_2 \leq O \left(\frac{1}{\pi \epsilon} \right) \max \left(\left\| \widetilde{\partial_x X} - \partial_x X \right\|_2, \left\| \widetilde{\partial_y X} - \partial_y X \right\|_2 \right).$$

.

Figure 3.11 illustrates the first two results of this corollary. If Ω contains the larger square which contains all frequencies whose magnitudes (modulo N) are less than $N/4$, then the final reconstruction error will only be at most $\sqrt{2}$ times worse than the maximum edge reconstruction error. Similarly, if Ω contains all frequencies with magnitudes less than $N/6$, then the final error will be multiplied by at most a factor of two. Interestingly enough, even though the Haar Wavelet method gives
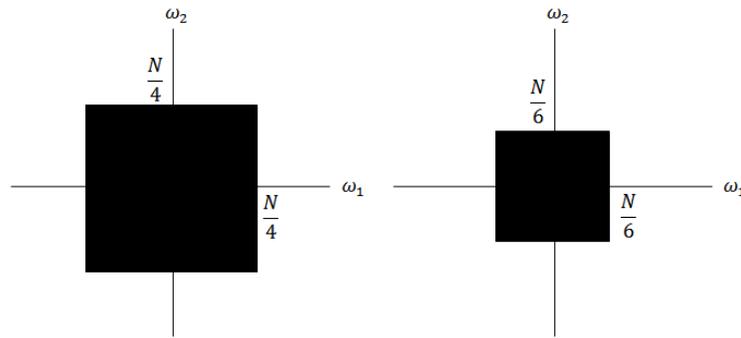
Figure 3.11: Sets of low frequencies that, when included in Ω, will yield stable edge integration performance guarantees.

the best asymptotical performance guarantee (e.g. $O(\sqrt{T \log N})$), in practice, the Least Squares approach typically yields ℓ_2 reconstruction errors of the same order. In fact, the Least Squares integration technique may be considered superior to the Haar wavelet method because it does an excellent job concealing errors produced during the edge reconstruction step. In the wavelet-based method, these errors will typically manifest themselves as blockiness, which is highly unappealing to the eye. Thus, with respect to the human "eye-ball norm", the Least Squares method performs better. The following section will consider the empirical performance of GRADIENTMP with both Haar Wavelet integration and least-squares integration schemes, as well as other techniques (e.g. TV minimization), in more detail.

3.5   Empirical Performance

This section compares the actual performance of GRADIENTMP, with the various methods of integration, against that of total-variation minimization. For the latter case, this work utilized both the L1-Magic suite of TV-optimization algorithms (with strict Fourier constraints) (see [48]) as well as a home-brewed implementation of the FTVD algorithm (see [64]).

The first step of this analysis is to show reconstructions of a 256×256 Shepp-Logan phantom from 10% of its Fourier coefficients. Figure 3.12 shows the original image, its full Fourier Transform, and it's partial Fourier Transform over Ω.

Original Phantom   Full Fourier Transform   Fourier Transform over Ω
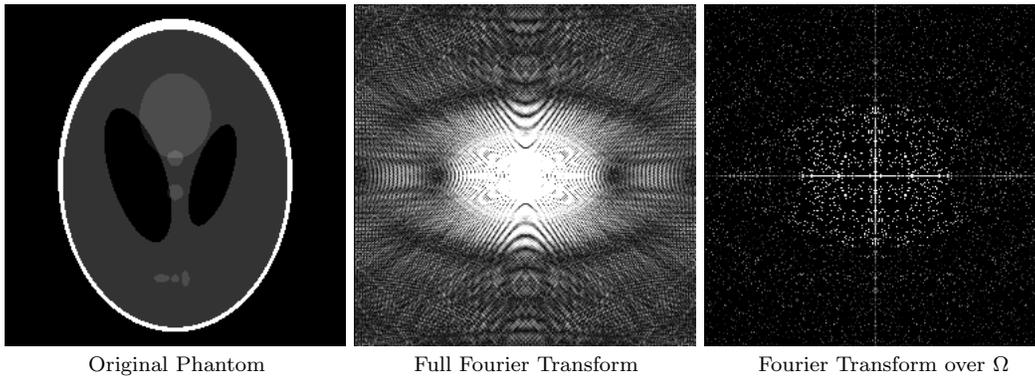
Figure 3.12: Original Shepp-Logan Phantom along with its full and partial Fourier Transforms.

The reconstructions of this phantom using GRADIENTMP with four different matching pursuit algorithms and the Haar Wavelet integration method, L1-Magic, and FTVD are shown below in Figure 3.13.

In this experiment, KOMP was performed using 22 iterations and $K = 100$.

All of the reconstructions were of good quality except for CoSAMP. The reason for CoSAMPs failure is that at every iteration, it must perform a least squares iteration on $3T \approx 6000$ atoms. This is a very poorly conditioned problem if only ten percent of the Fourier coefficient are given. As a result, CoSAMP was working well within its "break-down" zone.

GRADIENTMP (OMP)
MSE = .022
Runtime ≈ 4 min.

GRADIENTMP (KOMP)
MSE = .117
Runtime ≈ 24 sec.

GRADIENTMP (CoSAMP)
MSE = 36.1
Runtime ≈ 28 sec.

GRADIENTMP (CoSAMP)
MSE = .070
Runtime ≈ 9.5 hours

GRADIENTMP (CoSAMP)
MSE = 3.10
Runtime ≈ 4.5 minutes
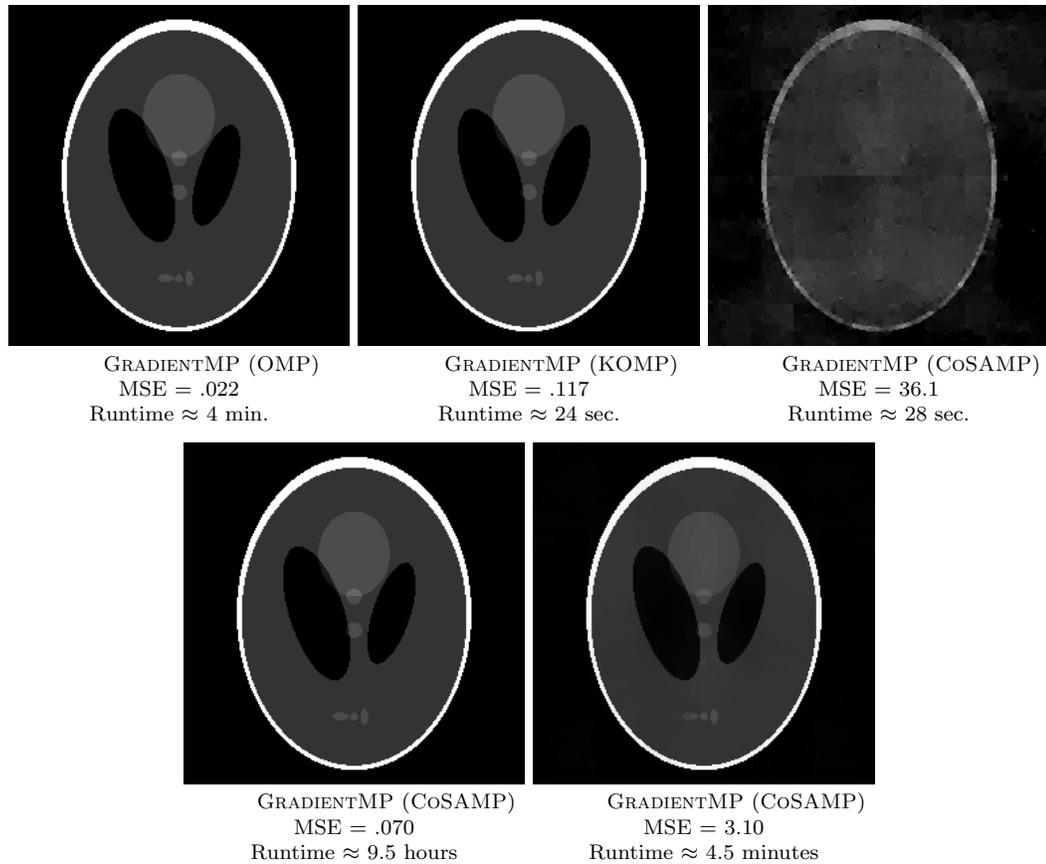
Figure 3.13: Various reconstruction of the Shepp-Logan phantom from 10% of its Fourier Coefficients.

The L1-Magic reconstruction was of very high quality, but took over 9 hours to run. The runtime of FTVd was not extremely fast, but is comparable to that of GRADIENTMP with plain OMP. The reconstruction with FTVd was not perfect, but definitely was of very high quality. The point to remember with TV minimization routines is that one can often tradeoff reconstruction accuracy with runtime. Thus, FTVd might be able to return a better reconstruction, but then its runtime would become too large.

In terms of the greedy algorithms, GRADIENTMP with OMP returned the best result. However, its runtime was a bit slow. GRADIENTMP with KOMP and CoSAMP were much faster; however, only GRADIENTMP with KOMP returned

an image with a mean square error comparable to that of GRADIENTMP with OMP.

This experiment utilized a set Ω of randomly selected frequencies. Such sampling grids can be approximated fairly well in MRI using cleverly designed k-space trajectories (see [34] for example). However, with the traditional setup of a CT machine, such a grid may not be possible. With CT, one can only obtain estimates of Fourier coefficients along radial lines such as those shown in Figure 1.9. This presents a challenge for Gradient Matching Pursuit because of its inherent differencing schemes: Generating Fourier observations of edges is a high pass operation that attenuates low frequencies that are emphasized by radial frequency grids. However, despite this issue, it will be seen that utilizing the Least Squares method of edge integration produces satisfactory results mainly because of the low frequency region of concentration in Ω.

In the following experiment, the usual 256 by 256 Shepp Logan phantom is recovered from Fourier measurements given along 22 equally spaced radial lines as shown in Figure 1.9. The following algorithms were used: GRADIENTMP with the Haar Wavelet Method, GRADIENTMP with the Least Squares Method, TV-Minimization (L1-Magic), FTVD, and naive Fourier Back Projection. Both instances of GRADIENTMP utilized KOMP with $K = 100$. The resulting reconstructions are shown in Figure 3.14.

Of course, the back projected reconstruction is terrible and inappropriate for any sort of medical imaging application. The best reconstruction was returned by L1-Magic. Even though the original image was recovered exactly, the use of this algorithm is also inappropriate in a medical setting because the image processing time was several hours. FTVD significantly cut down on this time; however, the final result is blurrier than the reconstructions produced by both instances of GRA-

| Original | GRADIENTMP (Haar Wavelets)
MSE = 3.11 | GRADIENTMP (Least Squares)
MSE = 2.44 |
| TV-Min. (L1-Magic)
MSE = 0 | TV-Min. (FTVD)
MSE = 5.58 | Back Projection
MSE = 33.8 |
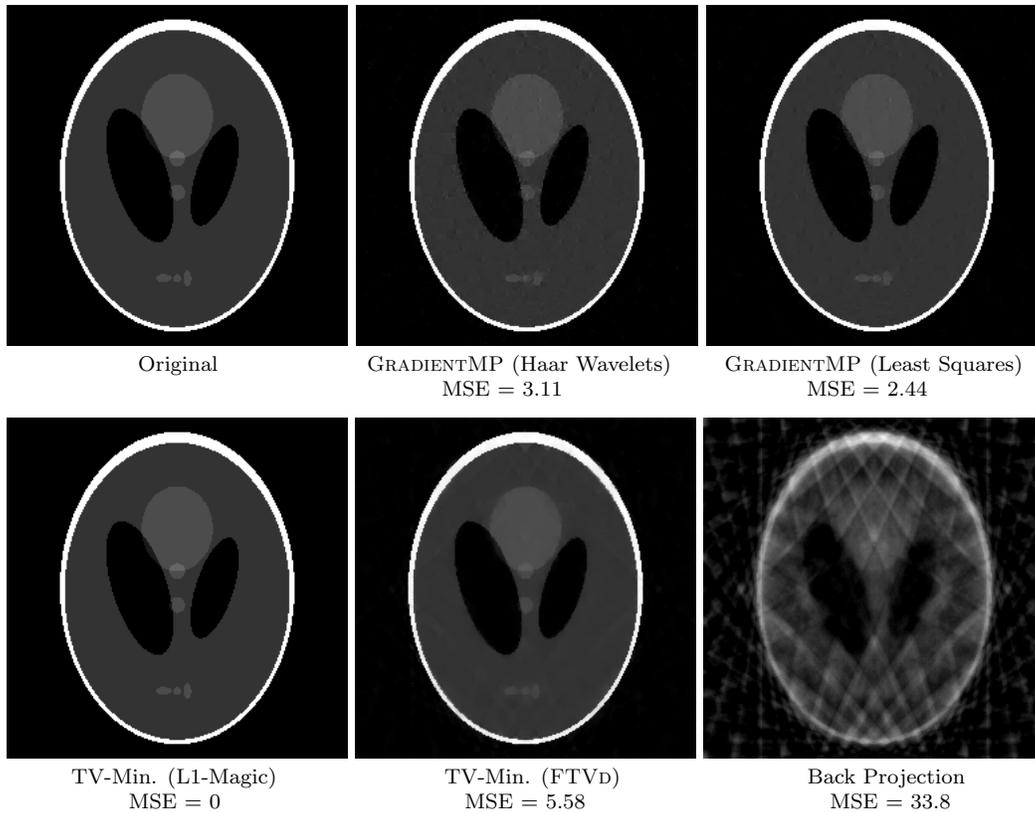
Figure 3.14: Various reconstruction of the Shepp-Logan phantom from Fourier Coefficients along radial lines.

DIENTMP. With respect to only the Gradient Matching Pursuit generated images, the one utilizing the Least Squares integration method is slightly better both visually and in terms of mean square error. The Haar Wavelet based reconstruction suffers from minor blockiness.

In the next experiment, a 64×64 version of the Shepp-Logan phantom was recovered using GRADIENTMP (KOMP with $K = 25$ and 21 iterations) with both the Haar Wavelet and Least Squares methods, L1-Magic, and FTVD utilizing the following setup: For every value of p between .04 and .4 in increments of .04, 10 different random frequency masks Ω were produced consisting of fraction p of all possible 2D frequencies. Then the four algorithms were used to recover the phantom and for each p, the average normalized mean square error (NRMSE) and runtime

was computed. The results of this experiment are shown below in Figure 3.15
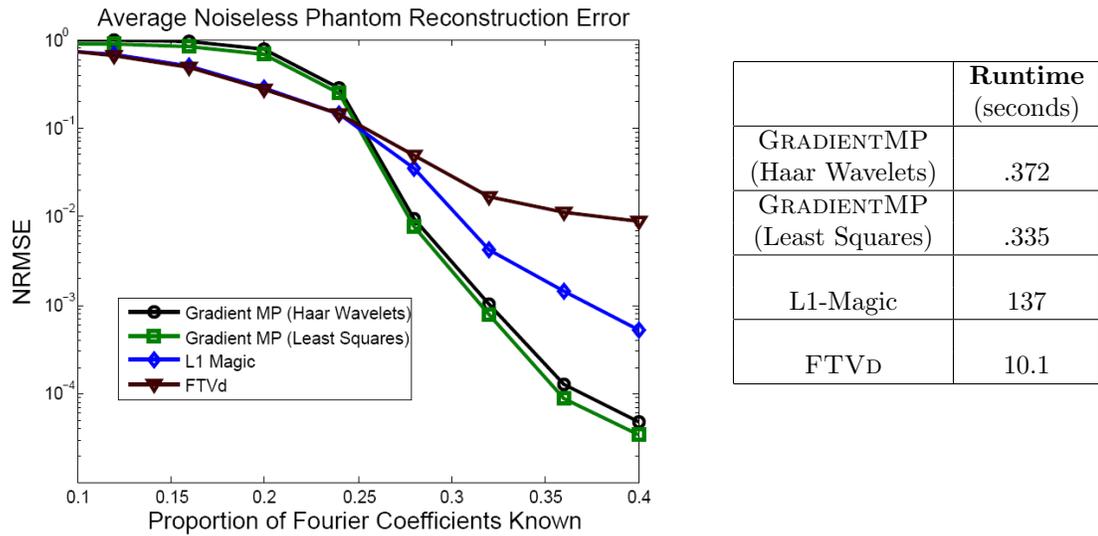
Figure 3.15: Reconstruction NRMSEs and Runtimes of GRADIENTMP and TV-Minimization with respect to recovering a 64×64 Shepp-Logan phantom.

In terms of the NRMSE, TV-minimization takes an early lead. But as one passes the point where at least 25% of all possible Fourier coefficients are provided, both variations of GRADIENTMP become the better performers. Of course, this latter region of p values is the more interesting case because all the algorithms do rather poorly when $p < .25$. Thus, in terms of accuracy, GRADIENTMP performs better than both L1-Magic and FTVD. In terms of GRADIENTMP alone, the Least Squares method returns results that are slightly more accurate than those produced by the Haar Wavelet method. Now in terms of runtime, both variations of GRADIENTMP were extremely fast and ran in approximately one-third of a second. FTVD was roughly 30 times slower than this and L1-Magic was about 400 times slower. Thus, in this ideal setting, GRADIENTMP is the all-around best choice for reconstructing the phantom.

The next experiment is very similar in nature except that this time, a signifi-

cant amount of Gaussian white noise will be added to each 64 ×64 phantom. This time, the proportion of frequencies was fixed with $p = .65$ over a symmetric Fourier sampling grid. The noise standard deviation was varied from $\sigma = .01$ to $\sigma = .1$ in increments of .01. GRADIENTMP was used before except with an increased number of iterations (30) for the purposes of noise-induced error reduction. L1-Magic was used with an inequality constraint on Fourier deviations from the given data. Finally, FTVD was used without modification. A plot of the resulting reconstruction normalized errors and runtimes is shown below in Figure 3.16.

	Runtime (seconds)
GRADIENTMP (Haar Wavelets)	.510
GRADIENTMP (Least Squares)	.481
L1-Magic	49.2
FTVD	6.86
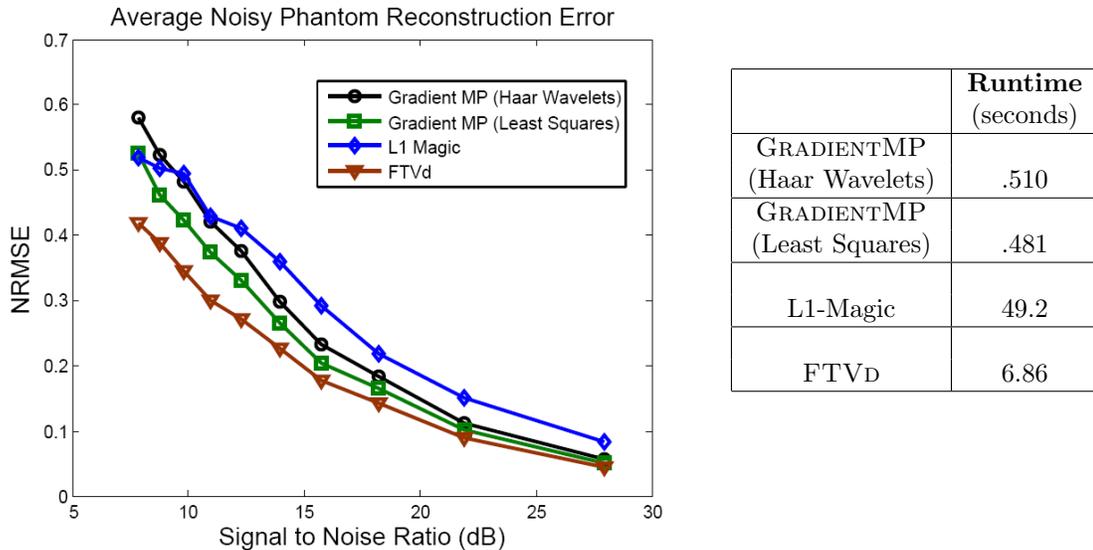
Figure 3.16: Reconstruction NRMSEs and Runtimes of GRADIENTMP and TV-Minimization with respect to recovering 64×64 Shepp-Logan phantoms corrupted with a significant amount of white Gaussian noise.

This is where FTVD had the chance to shine. It outperformed all the other algorithms in terms of reconstruction error. Both variations of GRADIENTMP followed closely behind. L1 Magic was behind the others because of sensitivity issues with the parameters associated with its underlying algorithm. These parameters can be extremely difficult to select appropriately given a noisy signal. As shall be seen later in this section, just because FTVD had slightly NRMSEs, this does not necessarily

mean that the FTVD reconstructions look better. FTVD tends to perform the best in this situation because of its ability to smooth noisy data. This may result in the smoothing of important texture as well. In terms of runtime, GRADIENTMP was again by far the fastest algorithm.

Of course, one of the most important performance criteria that has not been discussed thus far is GRADIENTMP's ability to handle natural images. To this end, the following is an attempt to recover the famous image 'Peppers' from 40% of all possible Fourier Coefficients (over a symmetric grid) using two techniques: GRADIENTMP (Least Squares) with KOMP($K = 200$, 30 iterations) and total-variation minimimzation (FTVD). The original image, its Fourier back-projection, and the two reconstructions are shown below in Figure 3.5.

Original

Back Projection (Raw Data)
MSE = 35.1

GRADIENTMP
MSE = 31.7

TV-Minimization (FTVD)
MSE = 27.8
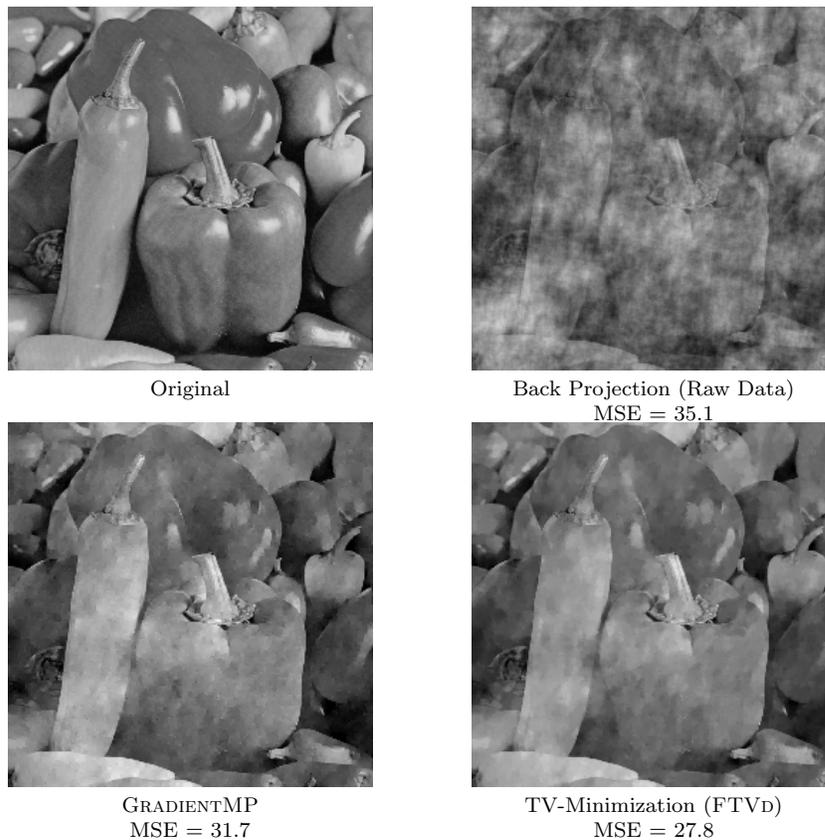
Figure 3.17: Original peppers image and reconstructions using Fourier Back Projection, GRADIENTMP, and FTVD.

In this particular example, the two reconstructions are very similar. In terms of mean square error, the TV-minimization reconstruction was slightly better. However, it took FTVD approximately one minute to run whereas GRADIENTMP only took 20 seconds.

The next experiment tests the same two algorithms on a real-life medical image. This time, a more conservative frequency selection consisting of 50% of all available frequencies randomly chosen over a symmetric grid was employed. The original image as well as its reconstructions are shown in Figure 3.5.
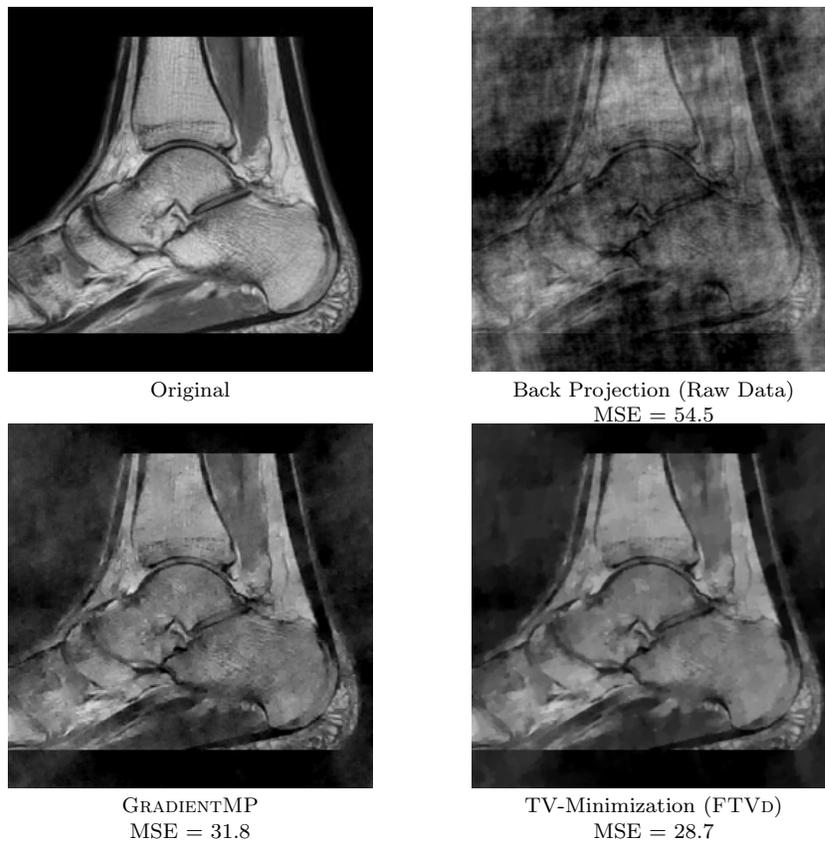
Figure 3.18: Original medical image and reconstructions using Fourier Back Projection, GRADI-ENTMP, and FTVD.

Again, both reconstructions are very similar in quality and all major anatomical objects can be seen in each. In terms of ℓ_2 error, TV-minimization performed slightly better, but the difference is most certainly not significant. GRADIENTMPdid a better

job in terms of retaining the texture in the original image. The TV-minimization reconstruction is somewhat blurry. For that reason, one may argue that with respect to the "eyeball" norm, GRADIENTMPactually performs better. The runtime of GRADIENTMPwas 20 seconds whereas the runtime of FTVᴅ was a bit over a minute.

To summarize, this chapter introduced Gradient Matching Pursuit, a novel algorithm to recover sparse gradient images from small sets of Fourier Data. It works by using greedy orthogonalized matching pursuit algorithms to recover an image in the edge (or gradient) domain and then utilizes one of several inverse filtering methods to form a final estimate of the original image. It was shown that given a set of Fourier samples, Gradient Matching Pursuit will produce an image that is of similar quality to those produced by existing total variation minimization routines, but in much less time. Unfortunately, these methods are somewhat controversial in that many discriminating radiologists would be unsatisfied with image reconstructions generated from both GRADIENTMP and total variation minimization, even for modest size frequency sampling sets. Both exhibit some degree of overall cloudiness and blurring of very fine details in their output. However, this does not mean that the application of compressive sensing to medical imaging is a useless endeavor. Modern motion-based MRI, an application in which a lot of data must be collected over a short time period, may benefit greatly from such techniques. In such situations, high resolution scans may be unrealistic with current technology and the reconstructions shown in the above figures are more likely to resemble the quality that one would obtain in practice. Of course, this is an area of future work that will not be discussed any further here. Instead, the focus will be shifted in the next chapter to an application of sparse approximation theory in medical imaging, which is the correction of non-homogeneous sensitivity patterns of MRI RF transmission coils.

CHAPTER IV

Parallel Approximation Theory and MRI Excitation

4.1 MRI Parallel Excitation

In Chapter I Section 1.3, a mathematical description of MRI in an ideal setting was presented. This section will relax the assumption of ideality and consider situations in which one is presented with an imperfect RF transmitter coil with a sensitivity pattern given by $S_{\text{space}}(x, y, z)$. This function is a measure of the gain of the RF pulse in various regions of the scanning area. Over a thin slice, it may be assumed that this sensitivity pattern is constant valued in the longitudinal direction (z-axis) and therefore the sensitivity pattern can be written as

$$S_{\text{space}}(x, y, z) = S_{\text{in-plane}}(x, y)S_{\text{long}}(z).$$

Without loss of generality, S_{long} can be considered to be identically equal to unity. For convenience, the subscript on $S_{\text{in-plane}}$ will be dropped, i.e. let $S = S_{\text{in-plane}}$. By utilizing a single RF pulse, then the MRI machine will recover Fourier Transform coefficients of an image (over a thin slice) proportional to $\rho(x, y)S(x, y)$. Figure 4.1 shows a typical sensitivity pattern for an MRI machine with a single transmission coil.

If the sensitivity pattern is highly non-uniform, then unwanted image artifacts are guaranteed to be present. To correct this issue, some more advanced MRI systems
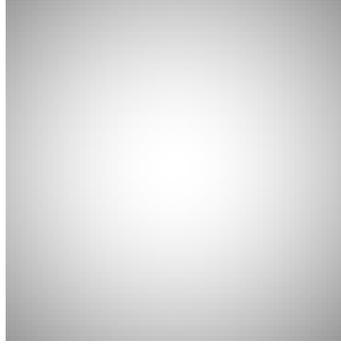
Figure 4.1: Typical single coil sensitivity pattern
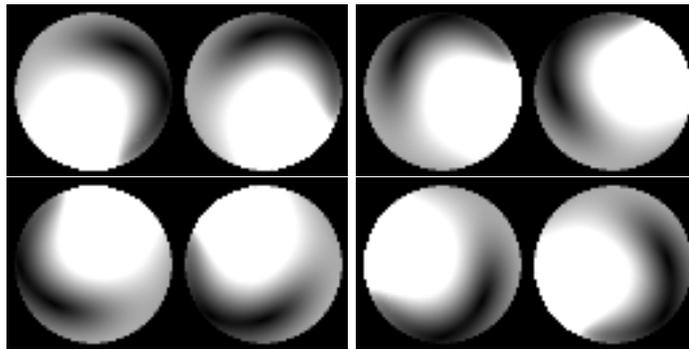
Figure 4.2: Typical sensitivity patterns for an 8-coil MRI machine.

possess multiple transmission coils. Assume that there are K such coils and that for each k, coil k has a thin-slice sensitivity pattern given by $S_k(x, y)$. Typical sensitivity patterns for an 8-coil MRI setup are shown below in Figure 4.2.

A sensitivity pattern is not a fixed property of the MRI scanner; rather, it depends on the subject that is sitting inside it. The double angle method (see [55] and [29]) can be used to efficiently calculate the sensitivity pattern of a given transmission coil. The first step is to generate two quick images of the desired slice. The first one is produced by transmitting an RF pulse inducing a tip angle of θ and the second image is produced using a tip angle of 2θ, hence the name "double-angle." Observe that based on Equation 1.13, the following proportionality is approximately satisfied:

$$\frac{I_1(x, y)}{I_2(x, y)} = \frac{\sin(\theta(x, y))}{\sin(2\theta(x, y))}$$

where I_1 and I_2 are the intensities of the images corresponding to tip angles θ and 2θ respectively. It is the value of $\theta(x,y)$ that determines the sensitivity pattern because in areas of poor sensitivity, the net proton magnetization vector will not tip as much as desired. By utilizing the trigonometric identity $\sin(2\theta) = 2\sin\theta\cos\theta$ and performing some basic algebra, one can see that

$$S(x,y) \propto \theta(x,y) = \cos^{-1}\left(\frac{I_2(x,y)}{2I_1(x,y)}\right).$$

Assuming again that only one RF pulse is transmitted by each coil, then the final reconstructed image will be proportional to $\rho(x,y)[S_1(x,y) + \cdots + S_K(x,y)]$. While this may improve the overall reconstruction quality, designing transmission coils that closely approximate a partition of unity over the scanning region is a highly non-trivial hardware problem. A clever way to address this issue is to allow each coil to transmit a short sequence of T RF pulses with each pulse being modulated by a two-dimensional waveform induced by the gradient magnet coils. The 2D frequencies associated with these waveforms are often referred to in the literature as phase encoding locations. The recovered image will now take the form $\rho(x,y)d(x,y)$ where

$$(4.1) \qquad d(x,y) = \sum_{k=1}^{K} S_k(x,y) \sum_{j=1}^{T} b_{k,j} e^{it\gamma(k_{x,j}x + k_{y,j}y)}.$$

Here, $d(x,y)$ is called the in-plane excitation profile and the $b_{k,j}$s are complex-valued weights associated with the phase encoding locations that will be selected in order to make d as close to being constant-valued as possible over the imaging region of interest. Over other regions, the value of d is unimportant. An example of a typical idealized excitation pattern is shown below in Figure 4.3.

In [69], Equation 4.1 is discretized as follows:

$$(4.2) \qquad d = S_1 F b_1 + S_2 F b_2 + \cdots + S_K F b_K$$
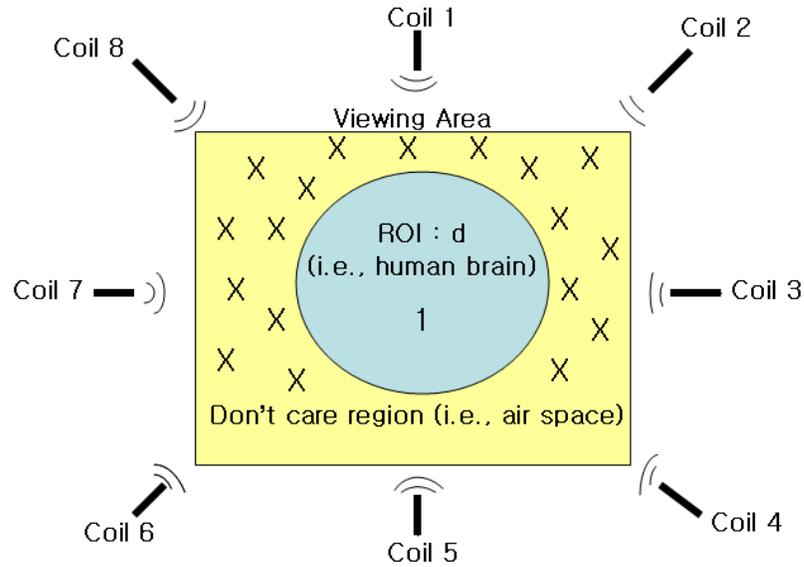
Figure 4.3: An idealized in-plane excitation profile that is identically one over the region of interest.

where S_k is a diagonal matrix representing the discrete sensitivity pattern of the kth coil, F is a two-dimensional Fourier encoding matrix, and b_k is a vector of the complex-valued weights for the RF pulses transmitted by the coil. The objective is to find a jointly T-sparse set of coefficient vectors $b = [b_1, \cdots, b_K]$ that will induce a uniform excitation pattern (e.g. $d = 1$) over the region of interest. The T non-zero entries of each b_k will correspond to the amplification of the T RF pulses that will be transmitted by the kth coil. Figure 4.4 graphically illustrates this pulse sequence.

	Time $\longrightarrow$			
Gradient Coils	f_1	f_2	f_3	$\cdots$
RF Coil 1	b_{11}*subpulse	b_{12}*subpulse	b_{13}*subpulse	$\cdots$
RF Coil 2	b_{21}*subpulse	b_{22}*subpulse	b_{23}*subpulse	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Figure 4.4: An illustration of a pulse sequence in MRI parallel excitation.

The time period between the transmission of subsequent pulses is very short. As a result, for simple single slice scans requiring only a small number of pulse segments, the order in which the pulses are transmitted is not important. In more complicated situations, such as 3D excitation or dynamic imaging applications, which will not be addressed here, this assumption is not valid.

The discrete version of the MRI Parallel Excitation problem motivates the more general Parallel Sparse Approximation Problem, which is described in the next section.

4.2 Parallel Sparse Approximation Theory and Parallel Orthogonal Matching Pursuit

There are two predominant paradigms for sparse approximation, both of which were discussed in Chapter I. In the first setting, $y \in \mathbb{C}^M$ is a signal vector that is to be synthesized as $y = \Phi x$, Φ is a dictionary matrix whose columns correspond to elementary signal vectors, and $x \in \mathbb{C}^N$ is a coefficient vector. The goal is to create an approximation of y using only a few columns of Φ, enforcing sparsity on x. Frequently in applications, such as hyperspectral imaging [65], sparse-gradient image recovery [37], etc., one is required to solve a related sequence of K sparse approximation problems of the form $y_k = \Phi x_k$ with $k = 1, \cdots, K$. If the x_ks have a common support set, then these K problems can be coupled leading to the second setting known as simultaneous sparse approximation theory. As before, one can refer to these two settings as compressive sensing and simultaneous compressive sensing if roles of the x's and y's are interchanged.

This work presents a variant of the sparse approximation problem that interpolates between these two paradigms, which shall be referred to as Parallel Approximation. In this problem, assume that one has a set of sparse coefficient vectors $\{x_1, \cdots, x_K\} \subset$

$\mathbb{C}^N$ with a common support set which additively synthesize a single signal using K highly incoherent dictionaries. In other words, one has K dictionary matrices $\{\Phi_1, \cdots, \Phi_K\} \subset \mathbb{C}^{M \times N}$ and a vector to approximate $y \in \mathbb{C}^M$ given by

$$(4.3) \qquad y = \Phi_1 x_1 + \Phi_2 x_2 + \cdots \Phi_K x_K.$$

The objective is to efficiently compute the x_ks from y and the Φ_ks while enforcing simultaneous sparsity on the x_ks. A graphical depiction of the basic parallel approximation problem is shown in Figure 4.5.
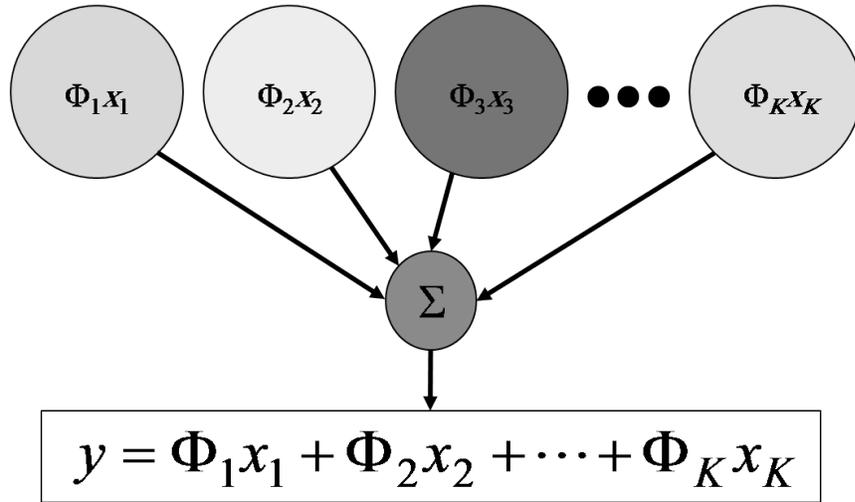
Figure 4.5: A graphical depiction of the multiple input single output structure of the basic parallel compressed sensing / sparse approximation problem.

It is easy to see that the discretized form of the MRI Parallel Excitation problem is a special case of the Parallel Sparse Approximation Problem. While the general Parallel Sparse Approximation setup is fairly new, it does appear in the literature in its mathematically equivalent forms such as the multiple input single output (MISO) forumulation in [67] and the block-sparse approximation problem in [54]. Predictably, the benchmark algorithmic approach for solving such problems is convex optimization. Letting $x \in \mathbb{C}^{NK}$ be the concatenation of $x_1, \cdots, x_K$ and

$\Phi = [\Phi_1, \cdots, \Phi_K] \in \mathbb{C}^{M \times NK}$, these optimization problems typically take the form:

$$\tilde{x} = \operatorname*{argmin}_{x} \|x\|_{1,p} \text{ s.t. } \Phi x = y$$

in the case of a constrained problem or

$$\tilde{x} = \operatorname*{argmin}_{x} \|x\|_{1,p} + \lambda \|y - \Phi x\|_2$$

in the case of an unconstrained problem. The norm $||\cdot||_{p,q}$ is calculated as follows: If x is an $N \times K$ matrix whose columns are the x_ks, then compute the ℓ_q form of each row and the find the ℓ_p norm of the resulting N-vector. For formally, the norm is defined as follows. Let $x'_n = [x_1(n), \cdots, x_K(n)] \in \mathbb{C}^K$. Then

$$\|x\|_{p,q} := \left\| \|x'_1\|_q, \|x'_2\|_q, \cdots, \|x'_N\|_q \right\|_p.$$

While convex optimization will produce high quality signal reconstructions, it tends to be fairly slow. As a result, we introduce a faster greedy algorithm known as Parallel Orthogonal Matching Pursuit (or POMP for short) to solve the same problem. The POMP algorithm is a very natural generalization of standard OMP and its pseudo-code representation is shown below in Figure 4.6.

In order to prove the correctness of POMP, it is necessary to introduce two definitions. First, one defines the cumulative coherence function $\mu_k(T)$ of dictionary Φ_k the same way as presented in Chapter I Section 1.2. Next, one needs to define a notion of coherence between atoms of different dictionaries. To that end, the following is a definition for the cross-cumulative coherence between K dictionaries.

```
            Algorithm:  Parallel Orthogonal Matching Pursuit

 Inputs:    A target signal to approximate :   y
            Dictionary matrices :   Φ_k,  k = 1, ··· , K
            Number of iterations T.
 Outputs:   T term approximations x̃_k s of the x_k s
            Residual r_T.
```

Initialize residual $r_0 = y$, index set $\Lambda = \emptyset$.
For t from 1 to T {
    Let

(4.4)
$$\lambda_t = \operatorname*{argmax}_{i} \sum_{k=1}^{K} \left| (\phi_k^i)^* r_t \right|.$$

    Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$.
    Let p_k be the projection of the residual onto the selected
      vectors $\{\phi_k^i | 1 \le k \le K, i \in \Lambda_t\}$.
    Set $r_t = r_0 - p_t$.   }
Solve for the $\widetilde{x_k}$'s by using the coefficients of the ϕ_k^i ss
  determined when solving for p_T.

Figure 4.6:   Pseudocode for Parallel Orthogonal Matching Pursuit.

Definition IV.1. The cross-cumulative coherence function $\nu_k(T)$ of a dictionary Φ_k from a collection of K dictionaries $\Phi_1, \cdots, \Phi_K$ is defined to be:

$$\nu_k(T) = \max_{i} \max_{\substack{|\Lambda|=T \\ \ell \neq k}} \sum_{j \in \Lambda} \left| (\phi_k^i)^* \phi_\ell^j \right|.$$

With the definitions of cumulative coherence and cross-cumulative coherence in mind, one is prepared to state a sufficient condition that will ensure that POMP will recover $x_1, \cdots, x_K$ from its measurements y.

Proposition IV.2. *Suppose one has an ensemble of K dictionaries $\Phi = [\Phi_1, \cdots, \Phi_K]$ with cumulative coherences μ_k and cross cumulative coherences ν_k, $k = 1, \cdots, K$ that satisfy*

(4.5)
$$\sum_{k=1}^{K} \mu_k(T) + \mu_k(T-1) + 2(K-1)\nu_k(T) < 1.$$

Then POMP *will select a correct common atom at every iteration and therefore recover the x_k s exactly.*

Proof. Suppose that after t iterations, POMP has selected only correct atoms. Then it follows that the residual r_t in an element of the space $\sum_{k=1}^{K} \text{colspan}(\Phi_{k,\text{opt}})$ where $\Phi_{k,\text{opt}}$ is the $M \times T$ submatrix of Φ_k which consist of the T columns corresponding to the correct atoms. Now write r as

$$r = \sum_{k=1}^{K} \sum_{i=1}^{T} c_{k,t} \phi_{k,\text{opt}}^{i}$$

where $\phi_{k,\text{opt}}^{i}$ is the i-th column of $\Phi_{k,\text{opt}}$. Now without loss of generality, assume that $\sum_{k=1}^{K} |c_{k,1}| \geq \sum_{k=1}^{K} |c_{k,i}|$ for each i. Otherwise, just reorder the columns of Φ_{opt}. Then it is possible to derive the following useful inequalities based on cumulative coherence and cross cumulative coherence estimates: For any $\phi_{k,\text{opt}}^{1}$ from the first column of some $\Phi_{k,\text{opt}}$,

$$|(\phi_{k,\text{opt}}^{1})^{*}r| \geq |c_{k,1}| - \left(\sum_{k=1}^{K} |c_{k,1}| \right) (\mu_k(T-1) + (K-1)\nu_k(T)).$$

Summing over k yields:

$$(4.6) \qquad \sum_{k=1}^{K} |(\phi_{k,\text{opt}}^{1})^{*}r| \geq \left(\sum_{k=1}^{K} |c_{k,1}| \right) \left(1 - \sum_{k=1}^{K} [\mu_k(T-1) + (K-1)\nu_k(T)] \right).$$

Similarly, for each k, define Ψ_k to be the $M \times (N-T)$ submatrix of Φ_k consisting of the incorrect atoms. Let ψ_k^{i} denote the ith column of Ψ_k. Then for a fixed k and i, one can obtain the estimate

$$|(\psi_k^{i})^{*}r| \leq \left(\sum_{k=1}^{K} |c_{k,1}| \right) (\mu_1(T) + (K-1)\nu_1(T)).$$

Again, keep i fixed and sum over k to obtain:

$$(4.7) \qquad \sum_{k=1}^{K} |(\psi_k^{i})^{*}r| \leq \left(\sum_{k=1}^{K} |c_{k,1}| \right) \left(\sum_{k=1}^{K} [\mu_1(T) + (K-1)\nu_1(T)] \right).$$

Now observe that P-OMP will definitely pick a correct atom if

$$\sum_{k=1}^{K} |(\phi_{k,\text{opt}}^{1})^{*}r| \geq \sum_{k=1}^{K} |(\psi_k^{i})^{*}r|.$$

Combining Inequalities 4.6 and 4.7 gives a sufficient condition for this, which is:

$$\left(\sum_{k=1}^{K} |c_{k,1}|\right)\left(\sum_{k=1}^{K} [\mu_1(T) + (K-1)\nu_1(T)]\right)$$
$$\leq \left(\sum_{k=1}^{K} |c_{k,1}|\right)\left(1 - \sum_{k=1}^{K} [\mu_k(T-1) + (K-1)\nu_k(T)]\right).$$

Rearranging terms now yields:

$$\sum_{k=1}^{K} \mu_k(T) + \mu_k(T-1) + 2(K-1)\nu_k(T) < 1,$$

which completes the proof. $\square$

An interesting observation is that by setting $K = 1$ in (4.5), then one gets Tropp's sufficient condition for OMP shown in (1.8). Indeed, by carefully examining the Parallel Orthogonal Matching Pursuit procedure, it is apparent that it is equivalent to ordinary Orthogonal Matching Pursuit when $K = 1$.

In terms of sparse approximation, one can obtain a result very similar to those shown in Equations 1.9 and 1.10. For convenience, define the function $\xi(T)$ to be

$$\xi(T) = \sum_{k=1}^{K} \mu_k(T) + \mu_k(T-1) + 2(K-1)\nu_k(T)$$

Let Λ_{opt} represent the column indices of the Φ_i that form the optimal T term representation y_T for some signal y. Now one can prove the following lemma:

Lemma IV.3. *Let $y \in \mathbb{C}^M$ be an arbitrary signal and let y_T be it's optimal T-term representation of the form*

$$y_T = \sum_{k=1}^{K} \Phi_k x_k$$

where the x_ks are jointly T-sparse coefficient vectors. Suppose that $\xi(T) < 0.5$ and that at iterations $1, \cdots, t-1$, POMP has selected only atoms indexed by Λ_{opt}. Then at iteration t, POMP will select another atom from Λ_{opt} if

$$(4.8) \qquad \|y - y_t\|_2 > \sqrt{1 + \frac{K^2 T\,(1 - \xi(T))}{(1 - 2\xi(T))^2}}\,\|y - y_T\|_2\,.$$

Proof. This proof is another adaptation of Tropp's work in [58, 60]. Observe that POMP will select a correct atom from Λ_{opt} at iteration t if the "greedy-selection ratio"

$$\rho(r_t) \equiv \frac{\|\Phi^* r_t\|_{1,\infty}}{\|\Psi^*_{\text{opt}} r_t\|_{1,\infty}}$$

satisfies $\rho(r_t) < 1$ where $\Psi = [\Psi_1, \cdots, \Psi_K]$ are the non-optimal atoms as in the proof of the previous proposition. Now one has that

$$\begin{aligned}
\rho(r_t) &= \frac{\|\Psi^* (y - y_t)\|_{1,\infty}}{\|\Phi_{\text{opt}} (y - y_t)\|_{1,\infty}} \\
&\leq \frac{\|\Psi^* (y - y_T)\|_{1,\infty} + \|\Psi^* (y_T - y_t)\|_{1,\infty}}{\|\Phi^*_{\text{opt}} (y - y_T) + \Phi^*_{\text{opt}} (y_T - y_t)\|_{1,\infty}} \\
&= \frac{\|\Psi^* (y - y_T)\|_{1,\infty}}{\|\Phi^*_{\text{opt}} (y_T - y_t)\|_{1,\infty}} + \frac{\|\Psi^* (y_T - y_t)\|_{1,\infty}}{\|\Phi_{\text{opt}} (y_T - y_t)\|_{1,\infty}}.
\end{aligned}$$

The above expressions utilize the fact that y_T is the orthogonal projection of y onto the columns of Φ_{opt}, i.e. $\Phi^*_{\text{opt}}(y - y_T) = 0$. Now it's easy to see that the second term satisfies

$$\frac{\|\Psi^* (y_T - y_t)\|_{1,\infty}}{\|\Phi_{\text{opt}} (y_T - y_t)\|_{1,\infty}} \leq \frac{\xi(T)}{1 - \xi(T)}$$

based on arguments presented in the proof of the previous proposition. As for the first term, one has that

$$\begin{aligned}
\frac{\|\Psi^* (y - y_T)\|_{1,\infty}}{\|\Phi^*_{\text{opt}} (y_T - y_t)\|_{1,\infty}} &= \frac{\max_i \sum_{k=1}^{K} |(\psi_k^i)^* (y - y_T)|}{\|\Phi^*_{\text{opt}} (y_T - y_t)\|_{1,\infty}} \\
&\leq \frac{\max_i \sum_{k=1}^{K} \|\psi_k^i\|_2 \|y - y_T\|_2}{\|\Phi^*_{\text{opt}} (y_T - y_t)\|_{1,\infty}} \\
&\leq \frac{K\sqrt{T} \|y - y_T\|_2}{\|\Phi^*_{\text{opt}} (y_T - y_t)\|_{2,2}}.
\end{aligned}$$

These follow from simple Cauchy-Schwartz inequality arguments. Next utilize the properties of the ℓ_2 norm and the Gershgorin Disc Theorem (as shown in [58]) to

show that

$$\frac{K\sqrt{T}\,\|y-y_T\|_2}{\left\|\Phi^*_{\text{opt}}\,(y_T-y_t)\right\|_{2,2}} \leq \frac{K\sqrt{T}\,\|y-y_T\|_2}{\sigma_{\min}\left(\Phi^*_{\text{opt}}\right)\|y_T-y_t\|_2}$$

$$\leq \frac{K\sqrt{T}\,\|y-y_T\|_2}{\sqrt{1-\xi(T)}\,\|y_T-y_t\|_2}.$$

Putting the pieces together, one sees that a sufficient condition to ensure that a correct atom is chosen is that

$$\frac{K\sqrt{T}\,\|y-y_T\|_2}{\sqrt{1-\xi(T)}\,\|y_T-y_t\|_2} + \frac{\xi(T)}{1-\xi(T)} < 1.$$

Now rearrange the terms to obtain

$$\|y_T-y_t\|_2 > \frac{K\sqrt{T\,(1-\xi(T))}}{1-2\xi(T)}\,\|y-y_T\|_2.$$

Since $y-y_T$ is orthogonal to y_T-y_t (i.e. $y_T-y_t \in \text{span}(\Phi_{\text{opt}})$ and $y-y_T \in \text{span}(\Phi^\perp_{\text{opt}})$), it follows from the Pythagorean Theorem that

$$\|y-y_t\|_2 > \sqrt{1 + \frac{K^2 T\,(1-\xi(T))}{(1-2\xi(T))^2}}\,\|y-y_T\|,$$

which completes the proof. $\qquad\square$

From this lemma, the following useful result immediately follows:

Proposition IV.4. *Let $y \in \mathbb{C}^M$ be any signal and let y_T be its best T-term representation in $\Phi_1, \cdots \Phi_K$, then assuming $\xi(T) < 0.5$, POMP will return a reconstruction $\tilde{y}$ of y that satisfies*

(4.9) $$\|y-\tilde{y}\|_2 \leq \sqrt{1 + \frac{K^2 T\,(1-\xi(T))}{(1-2\xi(T))^2}}\,\|y-y_T\|_2.$$

Thus, one sees that the reconstruction error generated by the output of POMP is not too much worse than the optimal jointly T-sparse representation: the sub-optimality constant does not depend on the signal size but rather only the sparsity

of the signal, the number of dictionaries, and the respective coherences of these dictionaries.

A common generalization of POMP is to replace the column index selection criterion (4.4) with a p-norm to obtain the alternative expression

$$\lambda_t = \underset{i}{\operatorname{argmax}} \sum_{k=1}^{K} \left| (\phi_k^i)^* r_t \right|^p.$$

This generalization will be denoted as POMP_p. For the special case $p = 2$, it is possible to modify the last proposition and obtain the following sparse approximation guarantee for POMP_2.

Proposition IV.5. *Let $y \in \mathbb{C}^M$ be any signal and let y_T be its best T-term representation in $\Phi_1, \cdots \Phi_K$, then assuming $\xi(T) < 0.5$, POMP_2 will return a reconstruction $\tilde{y}$ of y that satisfies*

(4.10)
$$\|y - \tilde{y}\|_2 \leq \sqrt{1 + \frac{K^3 T \left(1 - \xi(T)\right)}{\left(1 - \xi(T)(1 + \sqrt{K})\right)^2}} \, \|y - y_T\|_2.$$

Proof. In this situation, the greedy selection ratio takes the form:

$$\rho_t = \frac{\|\Psi r_t\|_{2,\infty}}{\|\Phi_{\text{opt}} r_t\|_{2,\infty}} \leq \frac{\sqrt{K} \, \|\Psi r_t\|_{1,\infty}}{\|\Phi_{\text{opt}} r_t\|_{1,\infty}}.$$

The remainder of the proof is exactly the same as it was for the previous lemma and proposition. $\square$

Observe again that for $K = 1$, this result is exactly equivalent to those obtainable for OMP and POMP_1. This is because all three algorithms are identical when $K = 1$.

The last several results deal with the application of cumulative coherence and cross-coherence functions to derive sparse approximation results pertaining to Parallel Orthogonal Matching Pursuit. In order to generate a more complete theory, we

now shift our attention to POMP as a compressive sensing algorithm and prove additional results. Proceeding in this direction, we describe the Parallel Approximation analogue of restricted isometry numbers.

Suppose one has a set of signals $x = [x_1; \cdots ; x_K]$ that are jointly T sparse and a set of measurements of the form $y = \Phi x = \Phi_1 x_1 + \cdots + \Phi_K x_K$. It is possible to derive sufficient conditions based on the restricted isometry numbers $\delta_{1,T}, \cdots, \delta_{K,T}$ of the measurement matrices $\Phi_1, \cdots, \Phi_K$ and a new, but similar concept, which will be defined as the restricted cross-isometry numbers:

Definition IV.6. For dictionaries Φ_k and Φ_ℓ where $k \neq \ell$, the restricted cross-isometry numbers $\epsilon_{k,\ell,T}$ are defined, as a function of T, to be the smallest numbers such that

$$\left\| \Phi_{k,A}^* \Phi_\ell x \right\|_2 \leq \epsilon_{k,\ell,T} \left\| x \right\|_2$$

for any signal $x \in \mathbb{C}^N$ and any set A such that the sum of the cardinalities of $\text{supp}(x)$ and A do not exceed T.

For convenience, let $\epsilon_{s,T} := \max_{k,\ell} \epsilon_{k,\ell,T}$ and let $\delta'_{s,T} := \max_k \delta_{k,T}$. Now bound everything by $\delta_{s,T} := \max(\delta'_{s,T}, \epsilon_{s,T})$. With all this in mind, one is ready to prove the following lemma pertaining to the restricted isometry properties of the overall measurement ensemble $\Phi = [\Phi_1, \cdots, \Phi_K]$, which will allow one to generate a result pertaining the performance of POMP_2. Then this will be extended to encompass the case $p = 1$.

Proposition IV.7. *Let $\Phi_1, \cdots, \Phi_K$ be K dictionaries with restricted isometry numbers $\delta_{k,T}$ and restricted cross-isometry numbers $\epsilon_{k,T}$. Let $x = [x_1, \cdots, x_K]$ be any sequence of jointly T-sparse signals with common support Λ. Then the following*

property regarding Φ holds.

$$\|\Phi_\Lambda^*\Phi x\|_{2,2} = \left\| \begin{pmatrix} \Phi_{1,\Lambda}^*\Phi_1 & \Phi_{1,\Lambda}^*\Phi_2 & \cdots & \Phi_{1,\Lambda}^*\Phi_K \\ \Phi_{2,\Lambda}^*\Phi_1 & \Phi_{2,\Lambda}^*\Phi_2 & \cdots & \Phi_{2,\Lambda}^*\Phi_K \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{K,\Lambda}^*\Phi_1 & \Phi_{K,\Lambda}^*\Phi_2 & \cdots & \Phi_{K,\Lambda}^*\Phi_K \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} \right\|_{2,2}$$

$$\geq \left(1 - \delta'_{s,T+1} - K\epsilon_{s,T+1}\right) \|x\|_{2,2}$$

$$\geq \left(1 - (K+1)\delta_{s,T+1}\right) \|x\|_{2,2}$$

In addition, for any singleton set $A := \{a\}$ where $a \notin \Lambda$,

$$\|\Phi_A^*\Phi x\|_{2,2} \leq \left(\delta'_{s,T+1} + K\epsilon_{s,T+1}\right) \|x\|_{2,2}$$

$$\leq (K+1)\delta_{s,T+1} \|x\|_{2,2}.$$

Proof. The proof is simply a matter of bookkeeping to keep track of the various restricted isometry and cross-isometry numbers involved in the overall construction of Φ. $\square$

Based on this overall restricted isometry condition of Φ, one can present a sufficient condition that will guarantee that POMP_2 will recover any jointly T-sparse ensemble of signals from its single measurement vector.

Proposition IV.8. *Let $\Phi_1, \cdots, \Phi_K$ be K dictionaries whose restricted isometry and cross-isometry numbers yield the following expression:*

$$\delta_{s,T+1} < \frac{1}{(K+1)\left(1 + \sqrt{T}\right)}.$$

Then POMP_2 will recover any jointly T-sparse ensemble of signals $x_1, \cdots, x_K$ from $y = \Phi_1 x_1 + \cdots \Phi_K x_K$.

Proof. Proceed by induction. Suppose that after t iterations, POMP$_2$ has only selected correct atoms from Λ. Let $r_t = \Phi_1 c_{t,1} + \cdots \Phi_K c_{t,K}$ where the $c_{t,k}$s are coefficient vectors that realize the residual. Let c_t be the concatenation of the $c_{t,k}$s. Consider the following two observations. First,

$$\|\Phi_\Lambda^* \Phi c_t\|_{\infty,2} \geq \frac{\|\Phi_\Lambda^* \Phi c_t\|_{2,2}}{\sqrt{T}} \geq \frac{(1 - (K+1)\delta_{s,T+1})\, \|c_t\|_{2,2}}{\sqrt{T}}.$$

Secondly, for any singleton set $A = \{a\}$ where $a \notin \Lambda$,

$$\|\Phi_A^* \Phi c_t\|_{\infty,2} = \|\Phi_A^* \Phi c_t\|_{2,2} \leq (K+1)\delta_{s,T+1}\, \|c_t\|_{2,2}\,.$$

Just like in the proof of Lemma II.3, POMP$_2$ will correctly identify the next atom if $\|\Phi_A \Phi c_t\|_{\infty,2} \leq \|\Phi_\Lambda \Phi c_t\|_{\infty,2}$. This will be guaranteed if

$$(K+1)\delta_{s,T+1}\, \|c_t\|_{2,2} \leq \frac{(1 - (K+1)\delta_{s,T+1})\, \|c_t\|_{2,2}}{\sqrt{T}}.$$

Rearranging terms gives (IV.8). □

This result can be extended for other values of p by slightly modifying the restricted isometry condition above. For the case $p = 1$, one has:

Proposition IV.9. *Let $\Phi_1, \cdots, \Phi_K$ be K dictionaries whose restricted isometry and cross-isometry numbers yield the following expression:*

$$\delta_{s,T+1} < \frac{1}{(K+1)\left(1 + \sqrt{KT}\right)}.$$

Then POMP$_1$ will recover any jointly T-sparse ensemble of signals $x_1, \cdots, x_K$ from its measurement vector.

Proof. The proof is exactly as before except for the utilization of the following inequalities:

$$\begin{aligned}
\|\Phi_\Lambda^* \Phi x\|_{\infty,1} &\geq \|\Phi_\Lambda^* \Phi x\|_{\infty,2} \\
\|\Phi_A^* \Phi x\|_{\infty,1} &\leq \sqrt{K}\, \|\Phi_A^* \Phi x\|_{\infty,2}\,.
\end{aligned}$$

All the subsequent steps remain unchanged. □

For general signal recovery, the same methods used over and over again throughout this work can be used to prove the following result:

Proposition IV.10. *Let $x = [x_1, \cdots, x_K]$ be a sequence of K arbitrary signals with optimal jointly T-sparse approximation $x_T = [x_{T,1}, \cdots, x_{T,K}]$. If a sequence of dictionaries $\Phi_1, \cdots, \Phi_K$ yield restricted isometry and cross-isometry numbers satisfying (IV.8) and one has access to noisy measurements of the form*

$$(4.11) \qquad y = \sum_{k=1}^{K} \Phi_k x_k + w = \sum_{k=1}^{K} \Phi_k x_{T,k} + e$$

where $e = \sum_{k=1}^{K} x_{T^C,k} + w$ and $x_{T^C} = x - x_T$, then POMP_2 will obtain an estimate $\tilde{x} = [\tilde{x}_1, \cdots, \tilde{x}_K]$ of x satisfying

$$\|x - \tilde{x}\|_{2,2} \leq \left(1 + \sqrt{K} D_{1,2}\right) \|x - x_T\|_{2,2} + \frac{D_{1,2}}{\sqrt{T}} \|x - x_T\|_1 + D_{1,2} \|w\|_2.$$

On the other hand, if Condition IV.9 is satisfied, then POMP_1 will obtain an estimate $\tilde{x}$ of x that satisfies:

$$\|x - \tilde{x}\|_{2,2} \leq \left(1 + \sqrt{K} D_{1,1}\right) \|x - x_T\|_{2,2} + \frac{D_{1,1}}{\sqrt{T}} \|x - x_T\|_1 + D_{1,1} \|w\|_2.$$

Here, $D_{1,2}$ grows roughly like $\sqrt{T}$ and $D_{1,1}$ grows like $\sqrt{KT}$.

Proof. First begin with the POMP_2 guarantee. Suppose that up until iteration t, only correct column indices have been selected. Observe that

$$\|\Phi_{\Lambda^C}^* r_t\|_{\infty,2} = \|\Phi_{\Lambda^C}^* \Phi_\Lambda (x_T - a_t) + \Phi_{\Lambda^C}^* e\|_{\infty,2}$$

$$\leq \|\Phi_{\Lambda^C}^* \Phi_\Lambda (x_T - a_t)\|_{\infty,2} + \|\Phi_{\Lambda^C}^* e\|_{\infty,2}$$

$$\leq (K+1)\delta_{s,T+1} \|x_T - a_t\|_{2,2} + \sqrt{K} \|e\|_2.$$

Here $a_t = [a_{t,1}, \cdots, a_{t,K}]$ are the coefficient vectors corresponding to the projection of y onto all Kt currently chosen column vectors of the Φ_ks. On the other hand, one also has:

$$\|\Phi_\Lambda^* r_t\|_{\infty,2} = \|\Phi_\Lambda^* \Phi_\Lambda (x_T - a_t) + \Phi_\Lambda^* e\|_{\infty,2}$$

$$\geq \frac{1}{\sqrt{T}} \left[\|\Phi_\Lambda^* \Phi_\Lambda (x_T - a_t)\|_{2,2} - \|\Phi_\Lambda^* e\|_{2,2} \right]$$

$$\geq \frac{1 - (K+1)\delta_{s,T+1}}{\sqrt{T}} \|x_T - a_t\|_{2,2} - \frac{\sqrt{K}\sqrt{1+\delta_{s,T}}}{\sqrt{T}} \|e\|_2 .$$

Now putting the two pieces together, a sufficient condition for a correct index to be chosen during iteration $t+1$ is that:

$$\|x_T - a_t\|_{2,2} > \frac{\sqrt{K}\left(\sqrt{T} + \sqrt{1+\delta_{s,t}}\right)}{1 - (K+1)\delta_{s,T+1}(1+\sqrt{T})} \|e\|_2 .$$

As before, define the stopping time t^* to be the first iteration where

$$\|x_T - a_{t^*}\|_{2,2} \leq \frac{\sqrt{K}\left(\sqrt{T} + \sqrt{1+\delta_{s,T}}\right)}{1 - (K+1)\delta_{s,T+1}(1+\sqrt{T})} \|e\|_2 .$$

The next step is to derive a bound on $\|x - \tilde{x}\|_{2,2}$. Observe that:

$$\|x - \tilde{x}\|_{2,2} \leq \|x_T - \tilde{x}\|_{2,2} + \|x_{T^C}\|_{2,2}$$

$$\leq \frac{1}{\sqrt{1 - (K+1)\delta_{s,2T}}} \|\Phi_{\Lambda'} (x_T - \tilde{x})\|_2 + \|x_{T^C}\|_{2,2} .$$

Here $\Lambda' = \text{supp}(\tilde{x}) \cup \Lambda$ which has cardinality at most $2T$. Now it is possible to further bound $\|x - \tilde{x}\|_{2,2}$ by:

$$\|x - \tilde{x}\|_{2,2} \leq \frac{1}{\sqrt{1 - (K+1)\delta_{s,2T}}} \left[\|\Phi_{\Lambda'} (x_T - \tilde{x}) + e\|_2 + \|e\|_2 \right] + \|x_{T^C}\|_{2,2}$$

$$\leq \frac{1}{\sqrt{1 - (K+1)\delta_{s,2T}}} \left[\|\Phi_{\Lambda'} (x_T - a_{t^*}) + e\|_2 + \|e\|_2 \right] + \|x_{T^C}\|_{2,2}$$

$$\leq \frac{1}{\sqrt{1 - (K+1)\delta_{s,2T}}} \left[\|\Phi_{\Lambda'} (x_T - a_{t^*})\|_2 + 2\|e\|_2 \right] + \|x_{T^C}\|_{2,2}$$

$$\leq \frac{\sqrt{1 + (K+1)\delta_{s,T}}}{\sqrt{1 - (K+1)\delta_{s,2T}}} \|x_T - a_{t^*}\|_{2,2} + \frac{2}{\sqrt{1 - (K+1)\delta_{s,2T}}} \|e\|_2 + \|x_{T^C}\|_{2,2} .$$

Now using the previously derived bound on $||x_T - a_{t^*}||_{2,2}$ gives:

$$||x_T - a_{t^*}||_{2,2}$$

$$\leq \left[\left(\frac{\sqrt{1+(K+1)\delta_{s,T}}}{\sqrt{1-(K+1)\delta_{s,2T}}} \right) \left(\frac{\sqrt{K}\left(\sqrt{T}+\sqrt{1+\delta_{s,T}}\right)}{1-(K+1)\delta_{s,T+1}(1+\sqrt{T})} \right) + \frac{2}{\sqrt{1-(K+1)\delta_{s,2T}}} \right]$$

$$\cdot ||e||_2 + ||x_{T^C}||_{2,2} .$$

The bracketed expression will be referred to as $D'_{1,2}$. The final step is to obtain a bound on the vector e:

$$||e||_2 \leq ||\Phi x_{T^C}||_2 + ||w||_2$$

$$\leq \sum_{k=1}^{K} ||\Phi_k x_{T_C,k}||_2 + ||w||_2$$

$$\leq \sum_{k=1}^{K} \sqrt{1+\delta_{s,T}} \left(||x_{T_C,k}||_2 + \frac{1}{\sqrt{T}} ||x_{T^C,k}||_1 \right) + ||w||_2$$

$$\leq \sqrt{1+\delta_{s,T}} \left[\sqrt{K} ||x_{T^C}||_{2,2} + \frac{1}{\sqrt{T}} ||x_{T^C}||_{1,1} \right] + ||w||_2 .$$

Set $D_{1,2} = \sqrt{1+\delta_{s,T}} D'_{1,2}$ and then put everything together to get:

$$||x - \tilde{x}||_{2,2} \leq \left(1 + \sqrt{K}D_{1,2}\right) ||x - x_T||_{2,2} + \frac{D_{1,2}}{\sqrt{T}} ||x - x_T||_1 + D_{1,2} ||w||_2 .$$

which completes the proof with respect to POMP_2. The analysis for POMP_1 is nearly identical. It isn't difficult to show that $D_{1,1} = \sqrt{1+\delta_{s,T}} D'_{1,1}$ where

$$D_{1,1} = \left(\frac{\sqrt{1+(K+1)\delta_{s,T}}}{\sqrt{1-(K+1)\delta_{s,2T}}} \right) \left(\frac{\sqrt{K}\left(\sqrt{KT}+\sqrt{1+\delta_{s,T}}\right)}{1-(K+1)\delta_{s,T+1}(1+\sqrt{KT})} \right)$$

$$+ \frac{2}{\sqrt{1-(K+1)\delta_{s,2T}}}.$$

This completes the proof. □

Of course, as suggested in Chapter II, these performance guarantees can be improved if one imposes a multiple column index selection scheme (e.g. K-POMP or

POMP with thresholding) and relax the restriction that only correct atoms must be chosen. As an example, consider the case of 2-POMP. For exactly jointly T-sparse signals, one can use the same sort of argument as in Propositions II.6 and II.8 in order to show the following:

Proposition IV.11. *Let* $\Phi_1, \cdots, \Phi_K$ *be a set of dictionaries whose restricted isometry and cross-isometry numbers satisfy:*

$$(4.12) \qquad \delta_{s,T+t+1} < \frac{1}{(K+1)\left(1 + \sqrt{\frac{T-t+1}{2}}\right)}$$

for each $t = 1, 2, \cdots, T$. *Then* 2-POMP$_2$ *will recover any jointly T-sparse signal from its measurements. Furthermore, if*

$$(4.13) \qquad \delta_{s,T+t+1} < \frac{1}{(K+1)\left(1 + \sqrt{\frac{K(T-t+1)}{2}}\right)}$$

for each iteration t, *then* 2-POMP$_1$ *will recover any jointly T sparse signal from its measurements.*

For a general recovery guarantee regarding 2-POMP$_2$, one can appeal again to the methods of Chapter II in order to obtain:

Proposition IV.12. *Let* $\Phi_1, \cdots, \Phi_K$ *be a set of dictionaries satisfying (4.12). Let* $x = [x_1, \cdots, x_K]$ *be a set of K general signals and x_{opt} be its optimal jointly T-sparse representation. Suppose one is given noisy measurements as shown in (4.11). Then* 2-POMP$_2$ *will recover a jointly $2T$ sparse approximation $\tilde{x}$ of x satisfying:*

$$\|x - \tilde{x}\|_{2,2} \leq (1 + \sqrt{K}D_{2,2})\|x - x_T\|_{2,2} + \frac{D_{2,2}}{\sqrt{T}}\|x - x_T\|_{1,1} + D_{2,2}\|w\|_2$$

where $D_{2,2} = \sqrt{1 + \delta_{s,T}} D'_{2,2}$ and

$$D'_{2,2} = \left(\frac{\sqrt{1 + (K+1)\delta_{s,2T}}}{\sqrt{1 - (K+1)\delta_{s,3T}}} \right) \left(\frac{\sqrt{K} \left[\sqrt{1 + \delta_{s,2}} \sqrt{\frac{T-t^*+1}{2}} + \sqrt{1 + \delta_{s,T}} \right]}{1 - (K+1)\delta_{s,T+t^*+1} \left(1 + \sqrt{\frac{T-t^*+1}{2}} \right)} \right)$$
$$+ \frac{2}{\sqrt{1 - (K+1)\delta_{s,3T}}}$$

Now if $\Phi_1, \cdots, \Phi_K$ satisfy (4.13), then POMP_1 will recover a jointly $2T$ sparse approximation $\tilde{x}$ of x satisfying:

$$\| x - \tilde{x} \|_{2,2} \leq (1 + \sqrt{K} D_{2,1}) \| x - x_T \|_{2,2} + \frac{D_{2,1}}{\sqrt{T}} \| x - x_T \|_{1,1} + D_{2,1} \| w \|_2$$

where $D_{2,1} = \sqrt{1 + \delta_{s,T}} D'_{2,1}$ and

$$D'_{2,1} = \left(\frac{\sqrt{1 + (K+1)\delta_{s,2T}}}{\sqrt{1 - (K+1)\delta_{s,3T}}} \right) \left(\frac{\sqrt{K} \left[\sqrt{1 + \delta_{s,2}} \sqrt{\frac{K(T-t^*+1)}{2}} + \sqrt{1 + \delta_{s,T}} \right]}{1 - (K+1)\delta_{s,T+t^*+1} \left(1 + \sqrt{\frac{K(T-t^*+1)}{2}} \right)} \right)$$
$$+ \frac{2}{\sqrt{1 - (K+1)\delta_{s,3T}}}.$$

As mentioned previously, performance guarantees based on a stopping times are somewhat undesirable. As a result, it is once again important to appeal to the rate at which the restricted isometry and cross-isometry numbers grow as a rate of the sparsity parameter T. More specifically, suppose $\delta_{s,T} \leq \delta_{s,2} T^\beta$. Consider the following proposition, which again is simply an extension of the analysis in Chapter II.

Proposition IV.13. *Suppose that $\Phi_1, \cdots, \Phi_K$ are a set of incoherent dictionaries whose restricted isometry and cross-isometry numbers obey the growth restriction $\delta_{s,\ell} < \delta_{s,2} \ell^\beta$ for $\beta < 0.5$. Then the following statements are true: First, assuming $T \geq 2$ and that*

$$\delta_{s,2} < \frac{1}{(K+1)(T+2) \left(1 + \sqrt{\frac{T}{2}} \right)},$$

then for any set of K signals $x = [x_1, \cdots, x_K]$, 2-POMP$_2$ can generate a jointly
$2T$-sparse approximation $\tilde{x}$ from the noisy measurements (4.11) satisfying:

$$\|x - \tilde{x}\|_{2,2} \leq (1 + \sqrt{K}\tilde{D}_{2,2}) \|x - x_T\|_{2,2} + \frac{\tilde{D}_{2,2}}{\sqrt{T}} \|x - x_T\|_{1,1} + \tilde{D}_{2,2} \|w\|_2$$

where $\tilde{D}_{2,2} = \sqrt{1 + \delta_{s,T}}\tilde{D}'_{2,2}$ and

$$\tilde{D}'_{2,2} = \left(\frac{\sqrt{1 + (K+1)\delta_{s,2}(2T)^\beta}}{\sqrt{1 - (K+1)\delta_{s,2}(3T)^\beta}} \right) \left(\frac{\sqrt{K}\left[\sqrt{1 + \delta_{s,2}}\sqrt{\frac{T}{2}} + \sqrt{1 + \delta_{s,2}T^\beta}\right]}{1 - (K+1)\delta_{s,2}T^\beta \left(1 + \sqrt{\frac{T}{2}}\right)} \right)$$
$$+ \frac{2}{\sqrt{1 - (K+1)\delta_{s,2}(3T)^\beta}}.$$

Second, assuming $T \geq 2$ and that

$$\delta_{s,2} < \frac{1}{(K+1)(T+2)\left(1 + \sqrt{\frac{KT}{2}}\right)},$$

then 2-POMP$_1$ can generate a jointly $2T$-sparse approximation $\tilde{x}$ satisfying:

$$\|x - \tilde{x}\|_{2,2} \leq (1 + \sqrt{K}\tilde{D}_{2,1}) \|x - x_T\|_{2,2} + \frac{\tilde{D}_{2,1}}{\sqrt{T}} \|x - x_T\|_{1,1} + \tilde{D}_{2,1} \|w\|_2$$

where $\tilde{D}_{2,1} = \sqrt{1 + \delta_{s,T}}\tilde{D}'_{2,1}$ and

$$\tilde{D}'_{2,1} = \left(\frac{\sqrt{1 + (K+1)\delta_{s,2}(2T)^\beta}}{\sqrt{1 - (K+1)\delta_{s,2}(3T)^\beta}} \right) \left(\frac{\sqrt{K}\left[\sqrt{1 + \delta_{s,2}}\sqrt{\frac{KT}{2}} + \sqrt{1 + \delta_{s,2}T^\beta}\right]}{1 - (K+1)\delta_{s,2}T^\beta \left(1 + \sqrt{\frac{KT}{2}}\right)} \right)$$
$$+ \frac{2}{\sqrt{1 - (K+1)\delta_{s,2}(3T)^\beta}}$$

For the sake of completeness, this work will present one more interesting varia-
tion of POMP whose name is Parallel Orthogonal Matching Pursuit with Projec-
tion (POMP$_{\text{proj}}$). In this particular version, the column index selection criterion is
changed from

$$\lambda_t = \underset{i}{\arg\max} \left\| \left(\Phi^i\right)^* r_t \right\|_p,$$

as is the case in POMP_p, into

$$\lambda_t = \underset{i}{\operatorname{argmax}} \left\| \left(\Phi^i \right) \left(\Phi^i \right)^\dagger r_t \right\|_2.$$

Here, the dagger $\dagger$ refers to the pseudo-inverse operator, i.e. $A^\dagger = (A^*A)^{-1}A^*$. In words, for each column index i, calculate the projection of the residual r_t onto the K vectors in the ith position of all K dictionaries. Then choose λ_t to be the value of i that yields the projection of greatest magnitude in the ℓ_2 sense. One might suspect that the added computational complexity of these projection would yield better results; however theoretical and empirical results thus far suggest otherwise. For purely jointly T-sparse signal recovery, one can prove the following result pertaining to $\text{POMP}_{\text{proj}}$.

Proposition IV.14. *Let $\Phi_1, \cdots, \Phi_K$ be dictionaries whose restricted isometry and cross-isometry numbers satisfy*

(4.14)
$$\delta_{s,T+1} < \frac{\gamma_{min}^3}{\gamma_{max}^3} \frac{1}{(K+1)(1+\sqrt{T})}$$

where γ_{max} and γ_{min} are the largest and smallest singular values respectively that occur in the set of all singular values of the matrices $\Phi^1, \cdots \Phi^N$. Then $\text{POMP}_{\text{proj}}$ will recover any jointly T-sparse set of signals $x_1, \cdots, x_K$ from their measurements.

Proof. Proceed by induction on the iteration number t as usual. Observe that for any correct column index $i \in \Lambda = \{\lambda_1, \cdots, \lambda_T\}$,

$$\left\| \left(\Phi^i \right) \left(\Phi^{i\dagger} \right) \Phi c_t \right\|_2$$
$$= \left\| \Phi^i \left(\Phi^{i*}\Phi^i \right)^{-1} \Phi^{i*} \left(\Phi^{\lambda_1} c'_{t,1} + \cdots + \Phi^{\lambda_T} c'_{t,T} \right) \right\|_2.$$
$$\geq \frac{\gamma_{\min}}{\gamma_{\max}^2} \left\| \Phi^{i*} \left(\Phi^{\lambda_1} c'_{t,1} + \cdots + \Phi^{\lambda_T} c'_{t,T} \right) \right\|_2.$$

Here each $c'_{t,\ell} \in \mathbb{C}^K$ contains entries of c_t corresponding to the columns vectors in Φ^{λ_ℓ}. Now observe that

$$\left\|\left\|\left(\Phi^{\lambda_1}\right)\left(\Phi^{\lambda_1\dagger}\right)\Phi c_t\right\|_2, \cdots, \left\|\left(\Phi^{\lambda_T}\right)\left(\Phi^{\lambda_T\dagger}\right)\Phi c_t\right\|_2\right\|_\infty$$
$$\geq \frac{\gamma_{\min}}{\gamma_{\max}^2}\frac{1}{\sqrt{T}}\left\|\left\|\Phi^{\lambda_1*}\left(\Phi^{\lambda_1}c'_{t,1} + \cdots \Phi^{\lambda_T}c'_{t,T}\right)\right\|_2, \cdots, \left\|\Phi^{\lambda_T*}\left(\Phi^{\lambda_1}c'_{t,1} + \cdots \Phi^{\lambda_T}c'_{t,T}\right)\right\|_2\right\|_2$$

Because of the $\|\cdot\|_{2,2}$ norm being used, one can rearrange terms to rewrite the last statement as

$$\frac{\gamma_{\min}}{\gamma_{\max}^2}\frac{1}{\sqrt{T}}\left\|\left\|\Phi_1^*\left(\Phi_1 x_1 + \cdots + \Phi_K x_K\right)\right\|_2, \cdots, \left\|\Phi_K^*\left(\Phi_1 + \cdots + \Phi_K x_K\right)\right\|_2\right\|_2.$$

Proposition IV.7 gives a lower bound for this quantity, which yields the result:

$$\left\|\left\|\left(\Phi^{\lambda_1}\right)\left(\Phi^{\lambda_1\dagger}\right)\Phi c_t\right\|_2, \cdots, \left\|\left(\Phi^{\lambda_T}\right)\left(\Phi^{\lambda_T\dagger}\right)\Phi c_t\right\|_2\right\|_\infty$$
$$\geq \frac{\gamma_{\min}}{\gamma_{\max}^2}\frac{1}{\sqrt{T}}\left(1 - (K+1)\delta_{s,T+1}\right)\|c_t\|_{2,2}.$$

Using a similar argument, it is easy to show that for any $j \notin \Lambda$,

$$\left\|\Phi^j\Phi^{j\dagger}\Phi c_t\right\|_{\infty,2} \leq \frac{\gamma_{\max}}{\gamma_{\min}^2}(K+1)\delta_{s,T+1}\|c_t\|_{2,2}.$$

Combining the last two inequalities in the usual manner gives the final result. $\square$

This result is very similar to that of POMP$_2$ except for the $\gamma_{\max}^3/\gamma_{\min}^3$ factor, which is somewhat related to the condition number of the matrices Φ^i $(1 \leq i \leq N)$. If these matrices are well-conditioned, then the RIP for POMP$_{\text{proj}}$ is only slightly weaker than that for POMP$_2$. In terms of general signal recovery, the same usual methods as before can be used to prove:

Proposition IV.15. *Let $\Phi_1, \cdots \Phi_K$ be K highly incoherent dictionaries whose restricted isometry and cross-isometry numbers satisfy (4.14). Let $x = x_1, \cdots, x_K$ be a set of any K signals with an optimal jointly T-sparse representation thereof denoted*

as x_T. Then, given noisy measurements of the form (4.11), POMP_{proj} will return a jointly T-sparse approximation $\tilde{x}$ of x that satisfies the following error bound:

$$\|x - \tilde{x}\|_{2,2} \le (1 + \sqrt{K}D_{\text{proj}})\|x - x_T\|_{2,2} + \frac{D_{\text{proj}}}{\sqrt{T}}\|x - x_T\|_{1,1} + D_{\text{proj}}\|w\|_2$$

where $D_{\text{proj}} = \sqrt{1 + \delta_{s,T}}D'_{\text{proj}}$ and

$$D'_{\text{proj}} = \left(\frac{\sqrt{1 + (K+1)\delta_{s,T}}}{\sqrt{1 - (K+1)\delta_{s,2T}}}\right)\left(\frac{\sqrt{K}\left[\frac{\gamma_{\max}}{\gamma_{\min}}\sqrt{T} + \sqrt{1 + \delta_{s,T}}\right]}{1 - (K+1)\delta_{s,T+t^*+1}\left(1 + \frac{\gamma_{\max}}{\gamma_{\min}}\sqrt{T}\right)}\right)$$
$$+ \frac{2}{\sqrt{1 - (K+1)\delta_{s,2T}}}.$$

Again, this result is slightly weaker than that of POMP_2; however, $\text{POMP}_{\text{proj}}$ will nevertheless be utilized in the following section on MRI Parallel Excitation because is was part of the machinery discussed (without proof) in [40].

4.3 Parallel Excitation Experiment

By setting $\Phi_k = S_k F$ and $x_k = b_k$, it becomes clear that the discrete MRI Excitation problem shown in Equation 4.2 is really nothing more than the basic parallel approximation problem of the previous section. In [68], this problem is solved by the following convex program:

$$(4.15) \qquad \tilde{b} = \underset{b=[b_1,\cdots b_k]}{\text{argmin}} \left\|d - \sum_{k=1}^{K} S_k F b_k\right\|_2 + \lambda\|b\|_{1,2}.$$

While this method produces high quality results, it will be shown that a special version of Parallel Orthogonal Matching Pursuit can achieve similar results in far less time. This, of course, is of high importance because the patient must be sitting still throughout the pulse design processing because even the slightest movement can throw off the calculation of sensitivity patterns.

In the spirit of Chapter II as well as that of the last section, one can adapt the basic $\text{POMP}_{p \text{ or proj}}$ as follows: First, select a threshold τ. For POMP_p, modify the

step involving the augmentation of the indexing set into:

$$\Lambda_t = \Lambda_{t-1} \cup \left\{ i : \left\| \left(\Phi^i \right)^* r_t \right\|_p \geq \tau \left\| \Phi^* r_t \right\|_{\infty,2} \right\}.$$

For $\mathrm{POMP}_{\mathrm{proj}}$, this modification becomes

$$\Lambda_t = \Lambda_{t-1} \cup \left\{ i : \left\| \left(\Phi^i \right) \left(\Phi^i \right)^\dagger r_t \right\|_2 \geq \tau \max_j \left\| \left(\Phi^j \right) \left(\Phi^j \right)^\dagger r_t \right\|_2 \right\}.$$

Secondly, if at iteration t, only one element is selected for addition into the index-ing set, then force the algorithm to select the "second place" candidate element as well. Thus, at every iteration, at least two indices are chosen. Because appropriate thresholds are functions of restricted isometry numbers that are hard to compute, this modification effectively counteracts the negative consequence of selecting a threshold that is too large.

To compare the performance of Parallel Orthogonal Matching Pursuit against that of the convex optimization algorithms that solve (4.15), consider the following experiment in which the goal is to excite a uniform circular pattern with radius 10.125 cm in a viewing area that is 24 cm × 24 cm. The entire problem setup is discretized over a 64 × 64 uniformly spaced grid. This desired in-plane excitation pattern is depicted in Figure 4.7.
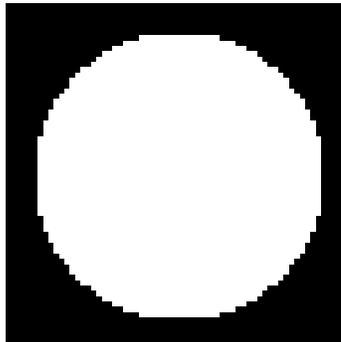
Figure 4.7: A circular excitation pattern of radius 10.125 cm over a 24 cm by 24 cm viewing area

The convex optimization approach was implemented using the SeDuMi convex

optimization package (see [56]). The POMP algorithms were implemented in MAT-LAB by [66]. $POMP_1$, $POMP_2$, and $POMP_{proj}$ were all utilized for the purpose of additional comparsion between these three variations of the original POMP concept.

The metrics used to compare the algorithms were runtime and the normalized root mean square error (NRMSE) of the induced excitation pattern compared to the ideal pattern shown above in Figure 4.7.

The above setup was first run utilizing an emulation of a single coil MRI setup with the Gaussian sensitivity pattern shown in Figure 4.1. Figure 4.8 shows a plot of the NRMSE of all four algorithms as a function of the number of RF pulses selected for transmission.
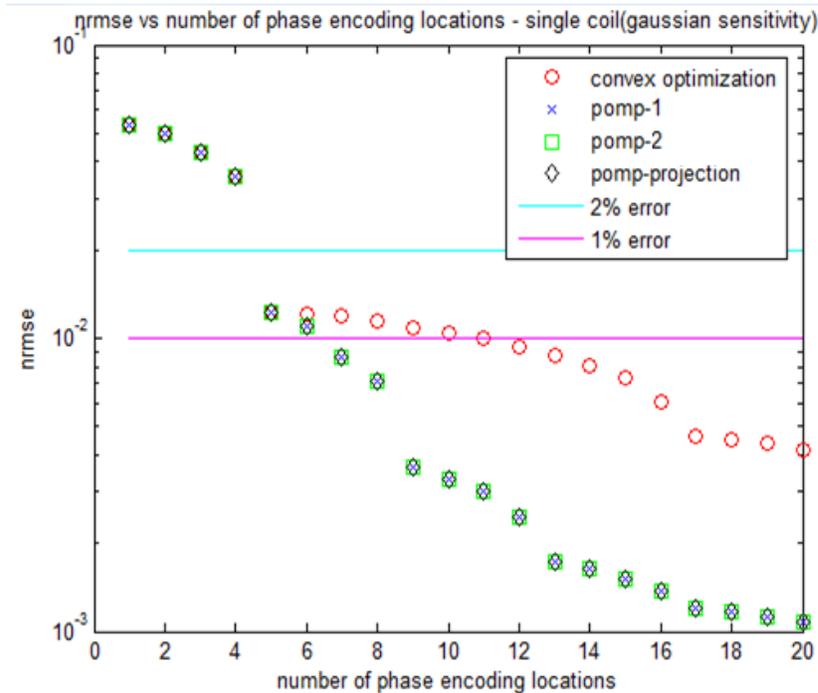
Figure 4.8: The NRMSEs of the excitation patterns induced by convex optimization and the three variations of POMP in the case of a single coil MRI setup

The results for the three variations of POMP are identical as expected. This is because for the single coil case where $K = 1$, $POMP_1$, $POMP_2$, and $POMP_{proj}$ are

all mathematically equivalent algorithms. In terms of convex optimization versus POMP, POMP did a better job in inducing an excitation pattern that is approximately uniform in the region of interest. In terms of runtime, each version of POMP ran in under five seconds whereas the convex optimization implementation took approximately 16 minutes to run.

Next, the same experiment was run utilizing an emulation of an eight-coil MRI setup with the sensitivity patterns shown in Figure 4.2. This time, convex optimization fared much better as shown in Figure 4.9.
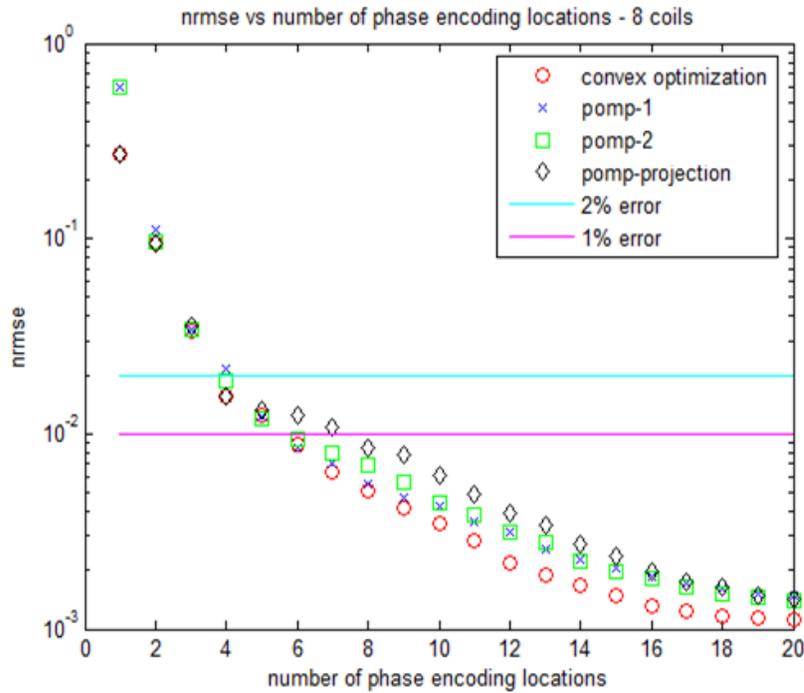
Figure 4.9: The NRMSEs of the excitation patterns induced by convex optimization and the three variations of POMP in the case of an 8-coil MRI setup

While convex optimization did technically win the NRMSE contest, POMP's performance was only a hair behind. In fact, the difference in performance can be considered almost negligible. In terms of the comparison of the three flavors of POMP, $POMP_1$ did slightly better than $POMP_2$ in the region of high quality

excitation pattern generations, i.e. when the number of selected phase encoding locations was greater than six. POMP$_{\text{proj}}$ lagged behind all the other algorithms; however, it too was not that far in the rear. In all three cases, POMP ran in under 25 seconds whereas convex optimization had a runtime of approximately 48 minutes.

The results of the above experiment are significant. They show that without any significant loss of quality, it is possible to use POMP to quickly determine a nearly optimal set of phase-encoding locations and weights for RF excitation in far less time than traditional convex optimization approaches.

To summarize, this chapter introduces Parallel Orthogonal Matching Pursuit as a fast and powerful tool for solving parallel sparse approximation and compressive sensing problems. One such important problem is that of Parallel Excitation in MRI. It was shown that by designing an appropriate RF pulse sequence, one can correct inhomogeneous sensitivity patterns in MRI transmission coils. POMP has been shown to produce such a sequence as accurately as a conventional convex program, but in significantly less time. Since the calculation of a good RF pulse sequence must be done while the patient lies inside the MRI machine, it is clear than POMP is far more suitable for in vivo medical imaging than convex optimization.

CHAPTER V

Conclusion and Future Work

Although CT and MRI are relatively modern imaging modalities that are highly utilized by the medical community to generate highly informative cross-sectional images of patients, they still suffer from their inability to run quickly. For example, a typical MRI scan may take anywhere from a few minutes for a simple localized scan to an hour or more for a complete exam. One primary reason for this issue is that a large amount of Fourier data must be obtained in order to run an inverse Fast Fourier Transform to recover an image. By taking advantage of sparsity in some domain (e.g. gradient or wavelet), it is possible to sub-sample the entire Fourier space below the Nyquist rate and still obtain enough information for a high-quality image reconstruction. Conventional methods that transform these smaller data sets into "good images" consist of slow convex optimization routines. While unconstrained total variation minimization routines such as FTVd (see [64]) significantly speed things up, these algorithms often suffer poor convergence results which may leave unwanted artifacts such as blurriness in their final reconstructions. The net result is a tradeoff. Either one of two things can happen:

1. One may place a patient inside a CT or MRI machine for a long period of time in order to generate the large amount of Fourier data required to run an efficient

FFT routine effectively.

2. One may exploit sparsity to reduce the scan time; however, the patient might possibly have to wait a long period of time before a high quality image can be produced.

Either way, a significant amount of unnecessary waiting occurs. In an emergency room setting, neither case scenario is acceptable. This work attempts to change that by introducing the Gradient Matching Pursuit algorithm. This novel algorithm takes advantage of the fact that many medical images are nearly sparse in gradient. By converting raw Fourier data into Fourier observations of the vertical and horizontal edge images, the algorithm is able to enlist the help of fast greedy algorithms such as Orthogonal Matching Pursuit and its many variations in order to recover these edges. After the edges are approximated, a cleverly designed inverse filter is used to reconstruct the final image. In the process of developing the GRADIENTMP algorithm, several related contributions have been made.

First, it was shown that Orthogonal Matching Pursuit is a strong compressive sensing algorithm that possesses performance guarantees based on restricted isometry conditions. While these guarantees are not as strong as those previously derived for convex optimization, it was shown that by modifying OMP so that it acts less locally and selects multiple atoms per iteration, new performance guarantees can be derived that look more and more like the benchmark standard set by Basis Pursuit.

Secondly, this work proposed two methods to recover an image from approximations of its vertical and horizontal derivatives. Both approaches were discussed in great detail and were shown to perform much better than the naive solution consisting of integration in a single direction.

The first method involves integrating both edge approximations in their respective directions and then using post processing in the Haar Wavelet domain in order to eliminate the streakiness that results from summing over erroneous pixels in the edge estimates. The important observation is that in the vertically integrated image, only Haar wavelets corresponding to horizontal differences reflect any streakiness. On the other hand, in the horizontally integrated image, only Haar wavelets corresponding to vertical differences show streakiness. The result of this observation is that one may mix and match uncorrupted wavelet coefficients from both expansions to produce a high quality image reconstruction. The advantage of this approach is that it performs well asymptotically, i.e. the final error is only $O(\sqrt{T \log(N)})$ times worse than the maximum edge error. However, the downside is that the original errors in the edge images will manifest themselves as blockiness in the final image reconstruction.

The second method involves modifying a convex optimization procedure first suggested by [19] so that it effectively becomes an ℓ_2 minimization problem that is pointwise closed-form solvable in the Fourier Domain. Asymptotically, the performance of this method is not wonderful: the final error is $O(N/r)$ times worse than the maximum edge error where r is the radius of the maximum filled circle inscribable in the 2D frequency set Ω. This is not much better than the $O(N/\sqrt{2})$ attainable for naive integration. However, in practice, this method does seem to perform the best. In addition, it works especially well in situations, e.g. parallel-beam CT, where the frequencies in Ω are heavily distributed in low frequency regions. By utilizing this method, it is possible to generate final reconstructions that are of the same quality as those generated by total variation minimization or, in some cases, even slightly better. Thus, utilizing this approach demonstrates that is possible to match the accuracy of convex optimization in much less time.

Another major contribution of this work was the theoretical and practical development of the Parallel Orthogonal Matching Pursuit algorithm and its many variants. Performance guarantees based on restricted isometry and cross-isometry numbers were developed. While [67] independently derived an algorithm equivalent to $POMP_2$ for the purposes of solving inverse problems with multiple inputs and a single output, it does not prove any theoretical results pertaining thereto nor does it present any of the other variations discussed here. The POMP algorithm was motivated by the necessity for fast algorithms that can design a short RF pulse sequence for MRI Parallel Excitation. Previous approaches ([69], [27], [68], and others) either did not take advantage of sparsity or used slow computational methods, e.g. convex optimization, in order to determine a pulse design. It was shown empirically that while both convex optimization and POMP perform similarly in terms of accuracy, POMP takes a few seconds to run whereas its counterpart has a runtime on the order of approximately a half hour. Because the excitation process must take place while a patient sits inside the MRI scanning tube, it makes sense to be able to compute a good pulse quickly. POMP accomplishes this objective.

While much progress was made through this work, there are some remaining open questions that remain to be answered. What follows is an outline of just a few of these.

The Gradient Matching Pursuit algorithm was not developed overnight. It slowly improved over the last several years as more sophisticated techniques were developed to improve the quality of the edge detection and inverse filtering steps. While the algorithm currently performs very well and produces results as accurate as those produced by Total Variation Minimization routines. What remains to be answered is whether GRADIENTMP can be optimized any further so that it can develop even

better scans? This may involve noise reduction, post-processing, or some other operation not considered here. As an even deeper question, what is the best that one can do? From an information theoretic point of view, given a small set of Fourier coefficients, how much information does one actually have about a sparse Gradient image? Answering this question provides a measure of the optimality of both the Gradient Matching Pursuit algorithm and total variation minimization techniques.

What poses a great challenge to Gradient Matching Pursuit and the underlying greedy edge recovery algorithm (e.g. OMP, KOMP, etc.) is the fact that most useful medical images are not extremely sparse in gradient. While super-linear time algorithms such as OMP and its variants can still handle this lack of sparsity reasonably well, there are other sub-linear time combinatorial compressive sensing algorithms that can't yet. Examples of these include Chaining Pursuit [24], HHS Pursuit [25], sublinear wavelet methods ([21] and [38]), Fourier methods ([20], [23], [31]), and others. The underlying principles behind all of these algorithms are those of isolation and group testing which are discussed thoroughly in [16]. The sublinear Fourier methods were designed to recover a sparse spectrum signal using a small set of structured samples in the Euclidean domain. The imaging problem of Chapter III is exactly the dual of this problem and therefore the same algorithm can be used to solve it. Theoretically, given an $N \times N$ image X that is T-sparse in gradient, the Ann Arbor Fast Fourier Transform (AAFFT) [31], is able to recover a list of the non-zero pixels of the edge images and their values using $O(T\,polylog(N))$ Fourier measurements in runtime $O(T\,polylog(N))$. Unfortunately, due to the nature of real-life medical images, no known isolation schemes effectively single out edge locations for group testing purposes. Instead, research related to this algorithm has taken a different direction: More stringent isolation schemes that require significantly more

measurements have been designed in order to construct deterministic (versus "with high probability") performance guarantees [30]. While this is a significant advance for those interested in processing extremely sparse signals, it does not help the medical imaging community much. Thus, one may ask: how can better isolation and group testing schemes be designed that work well with data that is not "that" sparse? The researchers of tomorrow need to derive better strategies in order for the various combinatorial compressive sensing algorithms to be effective in medical imaging.

In Chapter III, Gradient Matching Pursuit was developed as a tool for obtaining a reconstruction of a single slice of the cross-section of a patient from Fourier measurements. In practice, doctors often need to image several slices of the human body. Now one may wonder if there is any way to exploit the high correlation between two consecutive slices (or two consecutive frames in a dynamic scan) in order to perform simultaneous processing utilizing some variation of GRADIENTMP. The work [37] presents several applications of Simultaneous Orthogonal Matching Pursuit (SOMP) within the basic GRADIENTMP framework. Unfortunately, attempts to apply SOMP to this particular problem have not been very successful. To complicate matters worse, one may consider the case of an MRI setup with multiple imperfect receiver coils. Lustig et. al. recently presented some promising work that utilizes an iterative procedure tag-teamed with convex optimization known as SPIR-ıT [35]. Perhaps this convex optimization can be replaced with some sort of greedy algorithm to speed things up. However, even if successful, this does not address the multiple slice processing issue. It would behoove the medical community to find some efficient way of simultaneously processing data pertaining to multiple slices that were generated from multiple receiver coils. This may require a unification of Gradient Matching Pursuit, Simultaneous Orthogonal Matching Pursuit, and

Parallel Orthogonal Matching Pursuit.

The fourth and final question offered for future research pertains to the Parallel Excitation application of Sparse Approximation theory. In Chapter IV Section 4.3, it was stated as an assumption that when designing an optimal RF pulse sequence, one can consider the time-delay between transmitted pulse segments to be negligible. The result of this assumption is that the order in which the pulses are transmitted is unimportant and therefore, POMP becomes an appropriate tool for solving the Parallel Excitation problem. Unfortunately, even though the time delay between pulses is relatively small, the gap is still large enough to induce some decay in the tip angle of the magnetization which will effectively modify the weight of each pulse. This by itself is not a huge problem; rather, it is the phase distortion induced by these delays that is detrimental to the final image quality [66]. In the case of a single coil MRI setup, this problem can be modeled by

$$b = D^{T-1}\phi_{\lambda_1} + D^{T-2}\phi_{\lambda_2} + \cdots D^1\phi_{\lambda_{T-1}} + \cdots + \phi_{\lambda_T}$$

where ϕ_{λ_k} is the kth transmitted RF pulse and D is a time-delay operator that incorporates a tip angle decay and phase distortion. While this can be recast into a traditional sparse approximation problem, the high coherence between the slightly time-delayed atoms will make algorithms such as POMP ineffective without any additional work. Two suggestions for solving this problem are as follows: The first approach would be to attempt to design a new sparse approximation setup that reconciles the high coherences that otherwise arise from the time gaps. The second approach involves solving the parallel excitation problem assuming no delays and then optimizing the order in which the pulses are transmitted. In either case, many unknowns still exist, and therefore, this is definitely an area for significant future research.

Even though much work still remains to be done, this work hopes to make a significant impact on the medical imaging community in terms of demonstrating the applicability of sparse approximation methods to this field. This may not have been possible in the past given the fact that most sparse approximation algorithms of the past relied on slow convex programming techniques. However, with the introduction of the faster greedy algorithms introduced here, this goal may soon become a reality.

BIBLIOGRAPHY

[1] F. Bloch. Nuclear induction. *Physics Review*, 70:460–473, 1946.

[2] T. Blumensath. Sparsify 0.4.

[3] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Preprint*, 2008.

[4] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences*, 346:589–592, 2008.

[5] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Submitted*, 2004.

[6] E. J. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and innacurate measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

[7] E. J. Candes, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. *Proc. of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 295–308, 2005.

[8] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info. Theory*, 51:4203–4215, 2004.

[9] E. J. Candes and M. Wakin. An introduction to compressive sensing. *IEEE Signal Processing Magazine*, pages 21–30, Mar 2008.

[10] T. F. Chan, S. Esedoglu, F. Park, and M. H. Yip. *Recent Developments in Total Variation Image Restoration*. Springer Verlag, 2005.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scientific Computing*, 20:33–61, 1999.

[12] D. L. Donoho. Denoising via soft thresholding. *IEEE Transactions on Information Theory*, 44:613–627, May 1995.

[13] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. 2004.

[14] D. L. Donoho and I. M. Johnstone. Ideal spatial adaption via wavelet shrinkage. *Biometrika*, 81:425–455, Sep 1994.

[15] D. L. Donoho, Y. Tsiang, I. Drori, and J. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. 2006.

[16] D. Z. Du and F. K. Hwang. *Combinatorial Group Testing and its Applications*. World Scientific, New Jersey, 1993.

[17] Y. Eldar and H. Rauhut. Average case analysis of multichannel basis pursuit. *Proc. SampTA09*, 2009.

[18] C. L. Epstein. Introduction to magnetic resonance imaging for mathematicians. *Ann. Inst. Fourier, Grenoble*, 54(5):1697–1716, 2004.

[19] J. Fessler. Personal conversation.

[20] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. J. Strauss. Near-optimal sparse fourier representations via sampling. *Proc. of the 2002 ACM Symposium on Theory of Computing STOC*, pages 152–161, 2002.

[21] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. One-pass wavelet decompositions of data streams. *IEEE Trans. on Knowledge and Data Engineering*, 15:541–554, 2003.

[22] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. *Proc. of 2003 SIAM Symposium on Discrete Algorithms (SODA)*, pages 243–252, 2003.

[23] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Improved time bounds for near-optimal sparse fourier representation via sampling. *Proceedings of SPIE Wavelets XI*, 2005.

[24] A. C. Gilbert, M. J. Strauss, J. Tropp, and R. Vershynin. Algorithmic linear dimension reduction in the ℓ_1 norm for sparse vectors. submitted for publication. 2006.

[25] A. C. Gilbert, M. J. Strauss, J. Tropp, and R. Vershynin. One sketch for all: Fast algorithms for compressed sensing. *39th ACM Symposium Theory of Computing (STOC)*, Jun 2007.

[26] R. Gribonval, H. Rauhut, K. Schnass, and P. Vendergheynst. Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. *Submitted*.

[27] W. Grissom, C. Y. Yip, Z. Zhang, V. A. Stenger, J. A. Fessler, and D. C. Noll. Spatial domain method for the design of rf pulses in multicoil parallel excitation. *Magnetic Resonance Medicine*, 56:620–629, 2006.

[28] Rice University DSP Group. Rice single-pixel camera project. http://www.dsp.rice.edu/cscamera.

[29] E. K. Insko and L. Bolinger. B1 mapping. *Proceedings of the 11th Annual Meeting of SMRM*, page 4302, 1992.

[30] M. Iwen. A deterministic sub-linear time sparse fourier algorithm via non-adaptive compressed sensing methods. *ACM-SIAM Symposium of Discrete Algorithms (SODA)*, 2008.

[31] M. Iwen, A. C. Gilbert, and M. J. Strauss. Empirical evaluation of a sub-linear time sparse dft algorithm. *Communications in Mathematical Sciences*, 5(4), Dec 2007.

[32] J. Liao, J. M. Pauly, and N. J. Pelc. Mri using piecewise-linear spiral trajectory. *Magnetic Resonance in Medicine*, 38:246–252, 2005.

[33] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. Compresses sensing mri. *Signal Processing Magazine, IEEE*, 25:72–82, Mar 2008.

[34] M. Lustig, S. Kim, and J. M. Pauly. A fast method for designing time-optimal gradient waveforms for k-space trajectories. *IEEE Trans. on Medical Imaging*, 27:866–873, Jun 2008.

[35] M. Lustig and J. M. Pauly. Spir-it: Iterative self consistent parallel imaging reconstruction from arbitrary k-space. *Magnetic Resonance in Medicine*, 2009.

[36] R. Maleh. Sparse gradient image reconstruction via matching pursuit based algorithms. *Presentation at Sparse Approximation Workshop, Princeton University*, 2006.

[37] R. Maleh and A. C. Gilbert. Multichannel image estimation via simultaneous orthogonal matching pursuit. *SSP Proceedings*, 2007.

[38] R. Maleh and A. C. Gilbert. Sublinear recovery of sparse wavelet signals. *DDC Proceedings*, 2008.

[39] R. Maleh, A. C. Gilbert, and M. J. Strauss. Sparse gradient image reconstruction done faster. *ICIP Proc.*, 2007.

[40] R. Maleh, D. Yoon, and A. C. Gilbert. Fast algorithm for sparse signal approximation using multiple additive dictionaries. *SPARS Proceedings*, 2009.

[41] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1999.

[42] S. Mallat and A. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 1993.

[43] A. Marquina and S. Osher. Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. *SIAM J. Sci. Computing*, 1999.

[44] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and sub-gaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17:1248–1282, Nov 2007.

[45] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and innacurate samples. *Submitted*, 2008.

[46] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*. DOI: 10.1007/s10208-008-9031-3.

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

[48] J. Romberg. *L1-Magic*. http://www.acm.caltech.edu/ l1magic/.

[49] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. *International Mathematical Research Notices*, 64:4019–4041, 2005.

[50] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:1025–1045, 2008.

[51] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[52] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14:828–831, 2007.

[53] Salman Siddiqui, Stefan Robilia, and Jing Peng. Sparse representation for hyperspectral data classification. *IEEE International Geoscience and Remote Sensing Symposium*, 2008.

[54] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *Submitted*.

[55] R. Stollberger, P. Wach, G. McKinnon, E. Justich, and F. Ebner. Rf-field mapping in vivo. *Proceedings of the 7th Annual Meeting of SMRM*, page 106, 1988.

[56] J. F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones (updated for version 1.05). Oct 2001.

[57] H. Tan and C. H. Meyer. Estimation of k-space trajectories in spiral mri. *Magnetic Resonance in Medicine*, 61:1396–1404, Jun 2009.

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

[59] J. Tropp. Average-case analysis of greedy pursuit. *SPIE Wavelets XI*, pages 590401.01–11, Aug 2005.

[60] J. Tropp. Algorithms for simultaneous sparse approximation (parts i and ii). *Signal Processing, special issue "Sparse approximations in signal and image processing"*, 86:589–602, Apr 2006.

[61] J. Tropp and A. C. Gilbert. Signal recovery from partial information via orthogonal matching pursuit. Apr 2005.

[62] J. Tropp, A. C. Gilbert, and M. J. Strauss. Simultaneous sparse approximation via greedy pursuit. *Proc. 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, Philadelphia, PA*, Mar 2005. Invited paper, special session on "Sparse representations in signal processing".

[63] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimization Theory and Applications*, 109(3):475–494, Jun 2006.

[64] Y. Wang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 2008.

[65] R. Willett, M. Gehm, and D. Brady. Multiscale reconstruction for computational spectral imaging. *Computational Imaging V at SPIE Electronic Imaging*, Jan 2007.

[66] D. Yoon. Personal conversation and parallel excitation matlab code.

[67] A. C. Zelinski, V. K. Goyal, and E. Adalsteinsson. Simultaneously sparse solutions to linear inverse problems with multiple system matrices and a single observation vector. *SIAM Journal of Scientific Computing, Under Review*.

[68] A. C. Zelinski, L. L. Wald, K. Setsompop, V. K. Goyal, and E. Adalsteinsson. Sparsity-enforced slice-selective mri rf excitation pulse design. *IEEE Transactions on Medical Imaging*, 27:1213–1229, Sep 2008.

[69] Z. Zhang. Reduction of transmitted b1 inhomogeneity with transmit sense slice-select pulses. *Magnetic Resonance in Medicine*, 57(5):842–847, Apr 2007.