Global guarantees from local knowledge: stable and robust recovery of sparse in levels vectors

by

Matthew King-Roskamp

B.Sc., Simon Fraser University, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Mathematics
Faculty of Science

Declaration of Committee

Name: **Matthew King-Roskamp**

Degree: **Master of Science**

Thesis title: **Global guarantees from local knowledge: stable and robust recovery of sparse in levels vectors**

Committee: **Chair:** Nils Bruin
Professor, Mathematics

Ben Adcock
Supervisor
Associate Professor, Mathematics

Simone Brugliapaglia
Committee Member
Assistant Professor, Mathematics and Statistics
Concordia University

Weiran Sun
Committee Member
Associate Professor, Mathematics

Paul Tupper
Examiner
Professor, Mathematics

Abstract

The model of sparse vectors has proven invaluable for compressive imaging, allowing for signal recovery from very few linear measurements. Recently however, the structured sparsity model of *sparsity in levels* has inspired a new generation of effective acquisition and reconstruction modalities. Moreover this local structure arises in various areas of signal processing such as parallel acquisition, radar, and the sparse corruptions problem. Reconstruction strategies for sparse in levels signals have previously relied on a suitable convex optimization program. While iterative and greedy algorithms can outperform convex optimization and have been studied extensively in the case of standard sparsity, little is known about their generalizations to the sparse in levels setting. We bridge this gap by showing new stable and robust recovery guarantees for sparse in level variants of the iterative hard thresholding and the compressive sampling matching pursuit algorithms. Our theoretical analysis generalizes recovery guarantees currently available in the case of standard sparsity and favorably compare to sparse in levels guarantees for weighted ℓ^1 minimization, both in accuracy and computational time. In addition, we propose and numerically test an extension of the orthogonal matching pursuit algorithm for sparse in levels signals.

Keywords: sparsity in levels, compressed sensing, iterative and greedy methods, stability and robustness

Table of Contents

Chapter 1

Introduction

The model of *sparse vectors* has proven to be extremely useful for many physical phenomena. A sparse vector has s nonzero entries whose locations may be arbitrary. The assumption of underlying low-dimensional structure allows for recovery techniques from *compressed sensing* to recover an s-sparse vector $x \in \mathbb{C}^N$ from noisy linear measurements $y = Ax + e \in \mathbb{C}^m$. These techniques allow for accurate recovery of x even in the case where the number of measurements m is proportional to s up to log factors in N, and thus much less than the underlying dimension N. However, many applications exhibit structure beyond classical sparsity. Hence, there has been study on more complex *structured sparsity* models such as group or block sparsity, joint sparsity, weighted sparsity, connected tree sparsity and numerous others. That is, assumptions on not just the total sparsity s, but on the distribution of these nonzero entries. In fact, many of these more sophisticated models can lead to boosted practical performance [10, 25, 37]. Thus the problem of identifying a useful structured sparsity model that models applications and also admits a useful, tractable theory, is one of significant importance.

The focus of this thesis is the so-called *sparsity in levels model*, which has been shown to provide significant theoretical and practical gains over the standard sparsity model [7, 11]. Sparse in levels vectors exhibit a local sparsity pattern, specified by a vector $(s_1, \dots, s_r)$, as opposed to a single sparsity s. This seemingly simple generalization allows specification of a local structure within a vector, and leads to a rich theory that extends from the sparse setting [7, 29]. Up to this point however, the recovery of sparse in levels vectors has been approached via convex optimization techniques. Parallel to this, and of equal interest, is that of greedy or iterative approaches – which generally use fewer computational resources.

We focus on three standard iterative and greedy algorithms: Iterative Hard Thresholding (IHT), Orthogonal Matching Pursuit (OMP) and Compressive Sampling Matching Pursuit (CoSaMP), which are important algorithms for compressed sensing with standard sparsity [13, 32]. We introduce new generalizations of these methods to the sparse in levels setting, the first iterative approaches for this sparsity class. We generalize the theory of these recovery algorithms for sparse vectors to the levels case, and provides theoretical guarantees for the recovery of sparse in levels vectors that are robust to terms of noise in the measurement device or perturbations in the solution x itself. Furthermore, we study in the practical behavior of these approaches in both speed and accuracy compared to their sparse counterparts, and contrast them with convex optimization.

1.1 Motivation

While the theory of this model is inherently interesting, sparsity in levels arises commonly in various applications. Some examples of interest include so-called *sparse and distributed* or *sparse and balanced* vectors, which occur in parallel acquisition problems [20, 19] and radar [24]. Notably, the specific case of two levels also arises in the *sparse corruptions* problem [2, 30], in which a small fraction of the measurements of a signal are substantially corrupted – in comparison to some assumption of bounded or Gaussian noise on the signal.

Sparsity in levels has hereto been exploited using optimization-based decoders. Yet, it is well known that such decoders are not without issues; for example being computationally intensive. More directly worrisome is the observation that a decoder based on minimizing a convex optimization problem is not a method *per se*, as it requires a secondary algorithm to actually compute a solution. Therefore, as noted in [3], there is a gap between compressed sensing theory based on minimizers of optimization problems, and the observed practical performance. With this in mind, a primary motivation for this work is to derive algorithms with both recovery guarantees and provable polynomial computational time bounds in m and N.

Finally this model is useful in the problem of compressive imaging, which is the most direct and important application of this work. In compressive imaging problems, an image is first written as Wx where the coefficient vector x is sparse given suitable W, such as W encoding a wavelet basis. Then, linear samples of this image are recovered as $y = AWx$ for some sampling scheme A – for example $A = PF$ where F is the the standard Fourier matrix, and P a row sumsampling matrix. Not only is this vector x sparse, it exhibits sparsity in levels structure with asymptotic decay in the nonzero entries. This can be taken advantage of in the design of sampling strategies – in the design of the matrix A – to outperform methods optimized for standard sparse models, and give enhanced recovery performance [7, 8, 36]. Coupling this with a computationally guaranteed iterative or greedy method to quickly recover x is an current inquiry for accelerating imaging problems, inspired by [3].

1.2 Contributions

In this work, we propose three new algorithms for sparse in levels recovery: Iterative Hard Thresholding in Levels (IHTL), Orthogonal Matching Pursuit in Levels (OMPL) and Compressive Sampling Matching Pursuit in Levels (CoSaMPL). In the analysis of these methods, the main results of this work are robustness and stability guarantees for the levels-based algorithms, IHTL and CoSaMPL. They directly generalize known results for the sparse case, and require no additional or stricter assumptions in the general setting of sparsity in levels. These are presented in Theorem 12 and Theorem 13 respectively. These results determine an error bound in certain weighted ℓ^1-norms and the ℓ^2-norm depending on the approximate sparsity in levels, and the noise level. Furthermore we present equivalent results for the optimization-based weighted Quadratically-Constrained Basis Pursuit (wQCBP) decoder with the sparsity in levels model (Theorem 3). In comparing these theoretical bounds, we show that iterative approaches have no more stringent requirements for recovery than wQCBP, and comparable stability and robustness guarantees.

Finally we contrast these approaches numerically. Generally, we find that the levels based generalizations IHTL, CoSaMPL improve over their non-local counterparts, whereas OMPL and wQCBP show situational improvement. Furthermore, comparison with optimization based approaches show the iterative methods - specifically CoSaMPL - have similar accuracy with much less computational time. These numerics provide some evidence of cases where approaches based on iterative approaches outperform, or compare similarly, to convex optimization.

1.3 Literature review

The IHT and CoSaMP algorithms were introduced to compressed sensing in [13] and [32] respectively. Their theoretical analysis can be found, for instance, in [27]. The *Iterated Shrinkage* methods [26] served as a precursor for IHT, which was introduced in the context of compressed sensing in the late 2000s [13, 14]. Accelerating IHT using variable stepsize was examined later [12, 15]. A generalization of IHT to the union of subspaces model was studied in the context of model-based compressed sensing [10, 28]. Extensions of CoSaMP to the union of subspaces model were developed and analyzed in the context of model-based compressed sensing [10, 28]. Matching pursuit approaches were first studied by Zhang and Mallat [31], with OMP formulated later in [21]. Basis pursuit (a special case of wQCBP) was introduced to compressed sensing at its onset, in the seminal paper by Candès, Romberg, and Tao [18].

The sparsity in levels model was introduced in [7, 11]. Nonuniform recovery guarantees wQCBP were proven first in [7], with uniform guarantees later in [29].

Weighted QCBP has been studied recently in the compressed sensing literature, and has shown theoretical and practical uses in this context [1, 5]. The IHTL and CoSaMPL algorithms were introduced, and examined numerically, in [4] by the author. However, this previous work contained no theoretical analysis, and did not consider OMPL. Much of this thesis is based on a combination of work done by the author in [3, 4].

1.4 Outline

Chapter 2 begins with an introduction to compressed sensing in the sparse case, developing the standard tools of compressed sensing. Using this as a guideline, Section 2.3 develops sparsity in levels in a way motivated by the sparse case. This contains important theoretical tools necessary for later proofs, foremost the RIPL and the wRSNPL. This chapter concludes with the application of this theory to the example of Fourier-Haar wavelet problem in compressive imaging – a sparse in levels recovery problem with practical value.

Chapter 3 first develops the sparse versions of the algorithms of interest, IHT, CoSaMP, and OMP, and the merits of each. These approaches are then generalized to the levels setting, leading to the new algorithms IHTL, CoSaMPL, and OMPL.

Chapter 4 begins with the statement of the main results of this work: the recovery guarantees for IHTL and CoSaMPL. The benefits of these results, as well as comparison to the sparse case, and to wQCBP are discussed at length in Section 4.1. These results are proven in Section 4.2, which begins with its own detailed outline.

Chapter 5 numerically compares all the approaches mentioned within this work - - IHTL, CoSaMPL, OMPL and wQCBP. We compare both accuracy and runtime of these methods, to evaluate use cases for each.

Lastly, we conclude and point to some interesting directions of future work.

Chapter 2

Developing compressed sensing

2.1 Notation

We first require some notation. We write $\{e_i\}_{i=1}^N$ for the canonical basis of $\mathbb{C}^N$. For $x \in \mathbb{C}^n, 1 \leq p < \infty$, the ℓ^p-norm of x is

$$\|x\|_{\ell^p} = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}.$$

For $\Delta \subset \{1, \ldots, N\}$ we write P_Δ for the matrix of the orthogonal projection with the range $\text{span}\{e_i : i \in \Delta\}$. Hence, for $x \in \mathbb{C}^N$

$$(P_\Delta x)_i = \begin{cases} x_j & j \in \Delta \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Furthermore, if $A \in \mathbb{C}^{N \times N}$ then $P_\Delta A$ is the $N \times N$ matrix with ith row equal to the ith row of A for $i \in \Delta$ and zero otherwise. Similarly, AP_Δ has j-th column equal to A for each $j \in \Delta$ and zero otherwise. Finally, we note that the complement $P_\Delta^\perp = P_{\Delta^c}$ is the projection onto those indices not in Δ. Generally we will use the latter notation. To state certain error bounds, we make use of the notation $C \lesssim D$ if there exists some universal constant c independent of any parameters such that $C \leq cD$. Similarly, we may write $w > 0$ for a real valued vector $w \in \mathbb{R}^n$. This is read component-wise, namely $w_i > 0, 1 \leq i \leq N$.

Finally, we remark on some abuse of notation we shall use throughout. $P_\Delta x$ is isomorphic to a vector in $\mathbb{C}^{|\Delta|}$, and on occasion we shall use $P_\Delta x$ to refer to this object. Similarly, $P_\Delta A$ may refer to the $\mathbb{C}^{|\Delta| \times N}$ matrix with the zero rows of $P_\Delta A \in \mathbb{C}^{N \times N}$ deleted. In either case, it should be clear from context – we will point out this where there is any ambiguity.

2.2 Sparsity and classical compressed sensing

Definition 1. *A vector $x \in \mathbb{C}^N$ is s-sparse if*

$$|\{x_i \,:\, x_i \neq 0\}| \leq s.$$

The set of all s-sparse vectors is denoted Σ_s.

The set $\{x_i \,:\, x_i \neq 0\}$ is referred to as the support of x. Further, the set of all index sets $\Delta \subset \{1, \ldots, N\}$ with $|\Delta| = s$ is denoted by D_s. We also remark that $|\mathrm{supp}(x)|$ is often denoted $\|x\|_{\ell^0}$, the 'ℓ^0-norm'. This is in fact not a norm, but proves useful for notational convenience.

Given that we wish to recover not only s-sparse vectors, it is important to develop some notion of the effectiveness of an s-sparse approximation. These desired approximation rates will be measured in the standard ℓ^p-norms.

Definition 2. *For $x \in \mathbb{C}^N$, the ℓ^p-norm best s-term approximation error is given by*

$$\sigma_s(x)_{\ell^p} = \inf_{z \in \Sigma_s} \|z - x\|_{\ell^p}.$$

Note that of course, if $x \in \Sigma_s$ we have $\sigma_s(x)_{\ell^p} = 0$. Vectors for which $\sigma_s(x)_{\ell^p}$ is sufficiently small are called *compressible* or *approximately sparse*. Of course, the tolerance for which vectors are 'sufficiently small' may depend on application or context.

With these in hand, we can formally state the problem of interest. Given some set of linear measurements $y = Ax + e \in \mathbb{C}^m$, where $e \in \mathbb{C}^N$ is some unknown noise, we wish to recover $x \in \mathbb{C}^N$, where x is sparse. Here the *measurement matrix* A is given, for example, by a physical measurement device. Furthermore, we wish to do so with the smallest number of measurements $m \ll N$. This problem models the situation where samples are very expensive to compute, such in a parametric PDEs [23, 41] or imaging problems [3]. Of course this is not generally possible as our system is highly undetermined, and so we make the addition assumption that x is sparse. With oracle knowledge of the nonzero entries of x, recovery could be done in $O(s)$ measurements. However, we seek to attain the more reasonable order $O(s \cdot L)$ for a factor L logarithmic in both s and N.

While exact recovery of sparse vectors is desirable, we wish to generally recover compressible vectors, or sparse approximations. More concretely, we wish to recover a vector x with an error controlled by its distance to the set of s-sparse vectors, $\sigma_s(x)_{\ell^p}$. This is referred to as *stability* of a scheme. Similarly, we cannot measure our original signal x with infinite precision. Thus, even in ideal circumstances, we should only seek to guarantee an approximation $\hat{x}$ to x that is close enough to x relative to $\|e\|_{\ell^2}$ - not an exact recovery to x. A result that gives recovery even under measurement error is known as a *robustness*

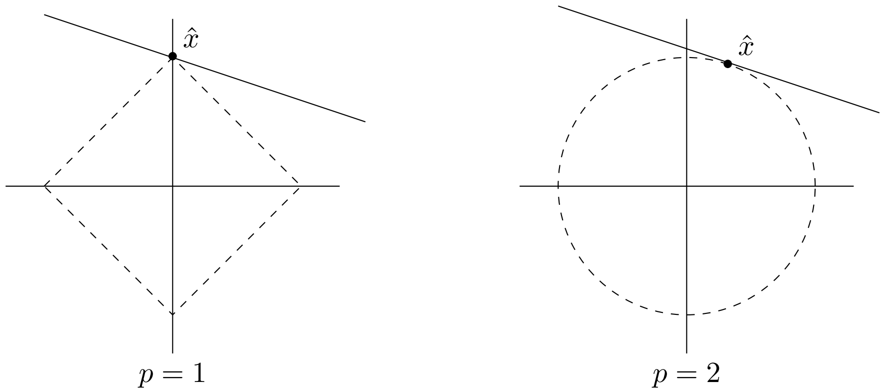


Figure 2.1: The ℓ^1-norm promotes sparsity. The solid line is the feasible set $\{z : Az = Ax\}$. The Dashed shape is the ℓ^1 (left) or ℓ^2 (right) ball, with the intersection point $\hat{x}$ being the minimal ℓ^1-norm (left) or ℓ^2-norm (right) solution. The ℓ^1-norm solution is generically sparse, whereas the ℓ^2-norm solution is generically nonsparse.

guarantee. With these ideas, we can restate our goal as follows: stable and robust recovery of x with as few measurements m as possible.

Let us consider the simplest case, when x is exactly sparse and there is no noise. This will serve to inform us about the difficulties that arise in the general case. One seemingly innocuous approach is to enforce the assumption of sparsity via a minimization approach, seeking to find the sparsest solution to $Ax = y$, that is,

$$\min_{z \in \mathbb{C}^N} \|z\|_{\ell^0} \text{ subject to } Az = y.$$

However, even this very basic idea immediately fails, as this problem is NP-hard and thus infeasible to solve even for moderately sized problems [27].

2.2.1 Basis pursuit and null space properties

Knowing we cannot solve the ℓ^0 minimization problem, even for exactly sparse vectors, we need some computationally tractable approach. One natural idea is to use a surrogate or relaxation of $\|\cdot\|_0$. In the sparse, noiseless case, the *Basis Pursuit* (BP) problem is a natural approach of interest

$$\min_{z \in \mathbb{C}^N} \|z\|_{\ell^1} \text{ subject to } Az = y. \tag{BP}$$

To see why this approach is reasonable, consider Fig. 2.1 in $\mathbb{R}^2$, reproduced with the authors permission from [6].

The line pictured is the feasible set, with the unit ℓ^1 ball. By finding the solution that minimizes the ℓ^1-norm, we see that we coincide with an axis, namely the ℓ^0 ball! Thus based on this intuition, we expect the solution of the basis pursuit problem to coincide with the sparse solution in most cases, with the exception being when the feasible set pictured here is tangent to the ℓ^1 ball. Or in more precise terms, when there are vectors v in the null space of the matrix A such that $\|x+v\|_{\ell^1} < \|x\|_{\ell^1}$. This leads to the following definition:

Definition 3. *A matrix $A \in \mathbb{C}^{m\times N}$ satisfies the Null Space Property (NSP) of order s if for each set $\Delta \in D_s$ we have*

$$\|P_\Delta v\|_{\ell^1} < \|P_{\Delta^c} v\|_{\ell^1} \text{ for all } v \in Null(A)\backslash\{0\} \tag{2.2}$$

This property gives rise to the following recovery result [27, Theorem 4.5]:

Theorem 1. *Given a matrix $A \in \mathbb{C}^{m\times N}$, every s-sparse vector $x \in \mathbb{C}^n$ is the unique solution (BP) with $y = Ax$ if and only if A satisfies the NSP of order s.*

Proof. Let us take some arbitrary Δ, and assume any vector x supported on Δ is the unique minimizer of the problem

$$\min_{z\in\mathbb{C}^m} \text{ subject to } Az = Ax. \tag{2.3}$$

Then, taking arbitrary v in the null space of A this holds in particular for $Az = AP_\Delta v$. But noting that $0 = Av = A(P_\Delta v + P_{\Delta^c} v)$ we have $A(-P_{\Delta^c} v) = AP_\Delta v$ and thus by assumption $\|P_\Delta v\|_{\ell^1} < \|P_{\Delta^c} v\|_{\ell^1}$, thus establishing the NSP.

Conversely, let us assume the NSP holds. Then, given some x supported on Δ and any $z \neq x$ with $Ax = Az$ we construct $v = x - z$, which is nonzero and in the null space of A. Then using the NSP

$$\begin{aligned}\|x\|_{\ell^1} \leq \|x - P_\Delta z\|_{\ell^1} + \|P_\Delta z\|_{\ell^1} &= \|P_\Delta v\|_{\ell^1} + \|P_\Delta z\|_{\ell^1} \\ &< \|P_{\Delta^c} v\|_{\ell^1} + \|P_\Delta z\|_{\ell^1} = \|P_{\Delta^c} z\|_{\ell^1} + \|P_\Delta z\|_{\ell^1} = \|z\|_{\ell^1}.\end{aligned}$$

As the choice of z was arbitrary we have thus established the minimality of x, as was to be shown. □

Now let us move to the more general context where the noise level e is nonzero. In this setting, solving the BP problem is not a good idea. To see why, consider solving the BP problem with $y = Ax + e$. A solution $\hat{x}$, for which $A\hat{x} = y$, generally will be non-sparse. In fact, the feasible set, the set of *sparse* vectors satisfying $A\hat{x} = y$, may simply be empty. However, given some estimate of $\|e\|_{\ell^2} \leq \eta$, we know there is a sparse solution x satisfying $\|y - Ax\|_{\ell^2} \leq \eta$. Replacing our equality constraint with this inequality constraint then allows for minimization of the ℓ^1 norm as before. This approach is Quadratically Constrained Basis Pursuit (QCBP), which allows noise to be addressed in the recovery process:

$$\min_{z\in\mathbb{C}^N} \|z\|_{\ell^1} \text{ subject to } \|Az - y\|_{\ell^2} \leq \eta. \tag{QCBP}$$

We shall see the following null space property gives guarantees useful for QCBP:

Definition 4. *A matrix $A \in \mathbb{C}^{m\times N}$ has the robust Null Space Property (rNSP) of order s with constants $0 < \rho < 1$ and $\tau > 0$ if*

$$\|P_\Delta v\|_{\ell^2} \leq \frac{\rho\|P_{\Delta^c} v\|_{\ell^1}}{\sqrt{s}} + \tau\|Av\|_{\ell^2},$$

for all $v \in \mathbb{C}^N$ and $\Delta \in D_s$.

Note that this definition makes no assumptions on the vector v. However, if $v \in \mathrm{Null}(A)$ and $\rho = 1$, we recover a NSP-like property as the second term is simply $\|Av\|_2 = 0$, with a stronger requirement that $\|P_\Delta v\|_2 \leq \frac{\rho\|P_{\Delta^c} v\|_{\ell^1}}{\sqrt{s}}$. However, this second term gives us greater control over the norm of Ax even for those vectors not in the null space. As with the NSP, this rNSP provides stable and robust recovery, formalized in the following result

Theorem 2. *(See e.g. [6, Theorem 5.14]) Suppose that $A \in \mathbb{C}^{m\times N}$ satisfies the rNSP of order s with constants $\rho \in (0, 1)$ and $\tau > 0$. Then for any $x \in \mathbb{C}^N$, a solution $\hat{x}$ of QCBP with $y = Ax + e$ and $\|e\|_{\ell^2} \leq \eta$ approximates x with error*

$$\|\hat{x} - x\|_{\ell^1} \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_{\ell^1} + \frac{4\tau}{1-\rho}\sqrt{s}\eta \tag{2.4}$$

$$\|\hat{x} - x\|_{\ell^2} \leq \frac{(3\rho+1)(\rho+1)}{(1-\rho)}\frac{\sigma_s(x)_{\ell^1}}{\sqrt{s}} + \frac{(3\rho+5)\tau}{(1-\rho)}\tau\eta \tag{2.5}$$

It is important to pause here and emphasize that this result not only gives the robustness we desired as the right hand side depends on η, but the stability as well, as the bound depends on $\sigma_s(x)_{\ell^1}$. Results of this flavor are of the utmost interest for this thesis, with the rNSP implying stable and robust recovery for some procedure – in this case QCBP.

2.2.2 From null space to restricted isometry property

While the rNSP gives a very useful theory, even in this sparse case it can be difficult to establish even for a given matrix A. Instead we develop a related definition, the Restricted Isometry Property (RIP).

Definition 5. *Let $1 \leq s \leq N$. The s-th Restricted Isometry Constant (RIC) δ_s of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest $\delta \geq 0$ such that*

$$(1-\delta)\|x\|_{\ell^2}^2 \leq \|Ax\|_{\ell^2}^2 \leq (1+\delta)\|x\|_{\ell^2}^2, \quad \forall x \in \Sigma_s. \tag{2.6}$$

If $0 < \delta_s < 1$ then the matrix is said to have the Restricted Isometry Property (RIP) of order s.

This, as the name suggest, the RIP measures how close A is to an isometry – a distance preserving map – on the set of s-sparse vectors. A useful characterization of the RIC is the following:

Lemma 1. *(See e.g. [6, Lemma 5.16]) The s-th RIC constant of a matrix satisfies*

$$\delta_s = \sup_{\Delta \in D_s} \|P_\Delta A^* A P_\Delta - P_\Delta\|_{\ell^2}. \tag{2.7}$$

Now, one may wonder why this property is useful in relation to sparse recovery – or to the rNSP previously developed. The next result gives us precisely this link in the sparse case:

Theorem 3. *(See e.g. [17]) Suppose that A satisfies the RIP of order $2s$ with constant $\delta_{2s} = \delta < \sqrt{2} - 1$. Then A satisfies the rNSP of order s with constants*

$$\rho = \frac{\sqrt{2}\delta}{1-\delta}, \qquad \tau = \frac{\sqrt{1+\delta}}{1-\delta}. \tag{2.8}$$

Firstly, and perhaps most importantly, as the RIP implies the rNSP, and the rNSP implies stable and robust recovery, we can attack the recovery problem – in theory – by finding those matrices which satisfy the RIP. It is also important to note that the condition $\delta < \sqrt{2} - 1$ is not optimal. In fact, it has been shown $\delta < 1/\sqrt{2}$ suffices, and that this is sharp [16].

From this follows a recovery result:

Theorem 4. *(See e.g. [6, Theorem 5.14]) Suppose that $A \in \mathbb{C}^{m \times N}$ satisfies the rNSP of order s with constants $\rho \in (0,1)$ and $\tau > 0$. Then for any $x \in \mathbb{C}^N$, a solution $\hat{x}$ of QCBP with $y = Ax + e$ and $\|e\| \leq \eta$ approximates x with error*

$$\|\hat{x} - x\|_{\ell^1} \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(x)_{\ell^1} + \frac{4\tau}{1-\rho}\sqrt{s}\eta, \tag{2.9}$$

$$\|\hat{x} - x\|_{\ell^2} \leq \frac{(3\rho+1)(\rho+1)}{(1-\rho)}\frac{\sigma_s(x)_{\ell^1}}{\sqrt{s}} + \frac{(3\rho+5)\tau}{(1-\rho)}\tau\eta. \tag{2.10}$$

2.2.3 Matrices satisfying the RIP

Until this point, we have talked about the properties measurement matrices A should have – but we have not examined any particular matrices. So it remains to see what measurement matrices are used in practice. Notably, we now have several desirable properties we wish for our matrices to satisfy – foremost being the RIP. With this in mind we return to one of the original problems set out in this section - reducing the number of measurements m needed. It is important to note that constructing deterministic matrices that give stable recovery *and* optimal or near-optimal measurement complexity – $m = s$ or $m = O(s \log(s))$ – is an open problem. Thus, all the matrices discussed here must naturally have some random element in their construction.

The first class of standard measurement matrices is a Gaussian random matrix

Definition 6. *A matrix $A \in \mathbb{R}^{m \times N}$ is a Gaussian random matrix if its entries are independent normal random variables with mean zero and variance one.*

Analyzing these types of matrices with tools from random matrix theory gives useful recovery guarantees, that explicitly bound the number of measurements m needed.

Theorem 5. *(See e.g. [6, Theorem 5.19]) Let $\epsilon \in (0,1)$, $s \in \{1, \dots, N\}$ and $A = \sqrt{\frac{1}{m}}\tilde{A}$ where $\tilde{A} \in \mathbb{R}^{m \times N}$ is a Gaussian random matrix. Then, if*

$$m \gtrsim s \log(eN/s) + \log(2\epsilon^{-1}), \tag{2.11}$$

the matrix A satisfies the RIP of order s with probability $1 - \epsilon$.

From the previous discussions of the RIP and rNSP we have that

Corollary 1. *(e.g. [6, Theorem 5.19]) Let $A = \frac{1}{\sqrt{m}}\tilde{A}$ satisfy the RIP of order s, where $\tilde{A} \in \mathbb{R}^{m \times N}$ is a random Gaussian matrix. Then, for every $x \in \mathbb{C}^{m \times N}$ and $y = Ax + e$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, any minimizer $\hat{x}$ of QCBP satisfies*

$$\|x - \hat{x}\|_{\ell^1} \lesssim \sigma_s(x)_{\ell^1} + \sqrt{s}\eta,$$
$$\|x - \hat{x}\|_{\ell^2} \lesssim \frac{1}{\sqrt{s}}\sigma_s(x)_{\ell^1} + \eta.$$

While this result is extremely promising, issues arise applying this practically. Generally, Gaussian random matrices are dense with no structure – and thus cannot be efficiently stored. Thus computing with these can be infeasible for moderately large problems. Another issue is simply being able to sample the underlying signal with a Gaussian matrix. For example in compressive imaging, samples are often prescribed or given from the application – such as Fourier samples in Magnetic Resonance Imaging.

These can both be tackled by the following ideas.

Definition 7. *A matrix $U \in \mathbb{C}^{N \times N}$ is unitary if $UU^* = U^*U = I$.*

Unitary matrices are interesting for many reasons, but one property of note is that they are distance preserving, $\|Ux\|_{\ell^2} = \|x\|_{\ell^2}$.

Definition 8. *Suppose $U \in \mathbb{C}^{N \times N}$ is unitary. A randomly-subsampled unitary matrix is a matrix $A \in \mathbb{C}^{m \times N}$ of the form*

$$A = \sqrt{\frac{N}{m}} P_\Omega U, \tag{2.12}$$

where $\Omega = \{\omega_1, \dots, \omega_m\}$ and the ω_i are chosen randomly and independently from $\{1, \dots, N\}$.

Under the umbrella of subsampled unitary matrices are methods permitting fast transforms. A particularly useful example of this is when A is the Fourier matrix, which allows for fast matrix-vector products via the FFT. To recall, the Fourier matrix F is

$$F = \frac{1}{\sqrt{N}} \left(e^{\frac{(-2\pi i)jk}{N}} \right)_{j=0,k=0}^{N-1,N-1} \tag{2.13}$$

The next question is which unitary matrices U will give good measurement conditions. For example, a terrible choice (but a valid one in the definition above) would be that of $U = I$ - measuring the entries of x directly. Without any knowledge of the support of x *a priori*, poor choices of indices can easily result in $P_\Omega I x = 0$, and generally we would need $m = N$ total measurements. Inspired by this, a good choice of U would be one that takes sparse x and transforms to nonsparse $z = Ux$, as then each measurement gives additional information about x. Concretely, we measure this by the coherence of a matrix:

Definition 9. *The coherence of a matrix $U \in \mathbb{C}^{N\times M}$ is*

$$\mu(U) = N \max_{i,j} |u_{i,j}|^2, \qquad U = (u_{ij})_{i,j=1}^{N,M}. \tag{2.14}$$

Observe for a unitary matrix $1 \leq \mu(U) \leq N$. We then shall say U is incoherent if $\mu(U) \approx 1$ and coherent if $\mu(U) \approx N$. The extreme undesirable coherent case is $\mu(I) = N$, whereas an example of a maximally incoherent matrix is the normalized DFT matrix, for which $\mu(U) = 1$. As we expect intuitively, coherence measures how the map U transforms sparse vectors to nonsparse ones:

Lemma 2. *Let $U \in \mathbb{C}^{N\times N}$ be unitary. Then, for any $x \in \mathbb{C}^N$,*

$$\|x\|_{\ell^0} + \|Ux\|_{\ell^0} \geq 2\sqrt{\frac{N}{\mu(u)}}. \tag{2.15}$$

In the sparse case where $\|x\|_{\ell_0} \ll N$, we require $2\sqrt{\frac{N}{\mu(u)}}$ to be as large as possible, as then it follows $\|Ux\|_{\ell^0}$ must be large, we have "spread out" the nonzero entries. Thus, small coherence is precisely the correct measure. This also relates back to our previous definitions in the following way:

Theorem 6. *([6, Theorem 5.22]) Let $\epsilon \in (0,1)$, $2 \leq s \leq N$. Let $U \in \mathbb{C}^{N\times N}$ be unitary, and A be a randomly-subsampled unitary matrix based on U. Suppose that*

$$m \gtrsim s \left(\log(s)\log^2(s)\log(N) + \log(\epsilon^{-1}) \right). \tag{2.16}$$

Then, with probability $1-\epsilon$, A satisfies the RIP of order s.

Which gives the corresponding recovery result.

Corollary 2. *([6, Theorem 5.22]) Let $\epsilon \in (0,1)$, $2 \leq s \leq N$. Let $U \in \mathbb{C}^{N\times N}$ be unitary, and A be a randomly-subsampled unitary matrix based on U. Suppose that*

$$m \gtrsim s\mu(U)\left(\log(\mu(U)s)\log^2(s)\log(N) + \log(\epsilon^{-1})\right). \tag{2.17}$$

Then, for every $x \in \mathbb{C}^{m\times N}$ and $y = Ax + e$, where $\|e\|_{\ell^2} \leq \eta$ for some $\eta \geq 0$, any minimizer $\hat{x}$ of QCBP satisfies

$$\|x - \hat{x}\|_{\ell^1} \lesssim \sigma_s(x)_{\ell^1} + \sqrt{s}\eta,$$
$$\|x - \hat{x}\|_{\ell^2} \lesssim \frac{1}{\sqrt{s}}\sigma_s(x)_{\ell^1} + \eta.$$

For incoherent U, we can simplify the measurement condition to be of the form

$$m \gtrsim s \cdot p(\log(s), \log(N)),$$

for some polynomial p of log factors in s, N. This generally means that the log factor is worse than the Gaussian case – intuitively a consequence of Gaussian matrices being more random. However, the amenability to fast computation makes subsampled unitary transforms desirable nonetheless.

2.3 Sparsity in levels

We now introduce sparsity in levels, a natural generalization of sparsity. This and the previous section proceed exactly in parallel, solving the same underlying problem of solving the linear system $Ax + e$, now for sparse in levels x.

Definition 10. *Let $r \geq 1$, $\boldsymbol{M} = (M_1, \ldots, M_r)$, where $1 \leq M_1 < M_2 < \ldots < M_r = N$ and $\boldsymbol{s} = (s_1, \ldots, s_r)$, where $s_k \leq M_k - M_{k-1}$ for $k = 1, \ldots, r$, with $M_0 = 0$. A vector $x = (x_i)_{i=1}^{M} \in \mathbb{C}^N$ is $(\boldsymbol{s}, \boldsymbol{M})$-sparse if*

$$|supp(x) \cap \{M_{k-1}+1, \ldots, M_k\}| \leq s_k, \quad k = 1, \ldots, r.$$

We write $\Sigma_{\boldsymbol{s},\boldsymbol{M}} \subseteq \mathbb{C}^N$ for the set of $(\boldsymbol{s}, \boldsymbol{M})$-sparse vectors.

This model was first introduced in [7]. We refer to $s = s_1 + \ldots + s_r$ for the total sparsity, and we denote by $D_{\boldsymbol{s},\boldsymbol{M}} \subset \{1, \ldots M\}$ the set of all index sets which are the support an $(\boldsymbol{s}, \boldsymbol{M})$-sparse vector. We refer to $\boldsymbol{M}$ as *sparsity levels* and $\boldsymbol{s}$ as *local sparsities*. Moreover, any index set of the form $\{M_{k-1}+1, \ldots, M_k\}$ for some $k = 1, \ldots, r$ is said to be a *level*. Of further use are the projection operators onto a level. Given $x \in \mathbb{C}^N$, these are defined as

$$(P_{M_{k+1}}^{M_k} x)_i = P_{M_{k+1}} P_{M_k}^{\perp} = \begin{cases} x_i & i \in \{M_k+1, \ldots M_{k+1}\} \\ 0 & \text{otherwise} \end{cases}.$$

While this is obviously a special case of the projection onto an index set, it is used with enough frequency to warrant special notation.

As in the sparse case it is important to develop some notion of the effectiveness of an approximation. These desired approximation rates will be measured in the *weighted* ℓ^p-norms

Definition 11. *Let $w = (w_1, \ldots, w_N) \in \mathbb{R}^N$ be a set of weights, with $w_i > 0$ for all i. For $0 < p \leq 2$, the weighted ℓ^p norm of a vector $x \in \mathbb{C}^N$ is*

$$\|x\|_{\ell^p_w} = \left(\sum_{i=1}^{N} w_i^{2-p)} |x_i|^p \right)^{1/p}. \tag{2.18}$$

Past works on convex optimization-based decoders for the sparsity in levels model have found that better uniform recovery guarantees can be obtained by replacing a usual ℓ^1-norm with a suitable weighted ℓ^1-norm [9, 37]. We shall find a similar phenomenon occurs in the case of iterative and greedy methods. As in these previous works, we shall suppose that the weights are constant on each level

$$w_i = w^{(k)}, \quad M_{k-1} < i \leq M_k,\ 1 \leq k \leq r, \tag{2.19}$$

for some $w^{(k)} > 0$. This is natural, as the assumption of levels structure is a prior on the nonzero entries. Assuming a sparse in levels structure assumes a s_i nonzero entries in a level – which naturally leads to having a constant weight on that level, expressing the size of the prior s_i. For example, a prior on the local sparsity s_i being large is expressed by having small weighting on this level.

Definition 12. *Given a vector of weights $w \in \mathbb{R}^N$ with $w > 0$, the best $(\boldsymbol{s}, \boldsymbol{M})$-term approximation error of $x \in \mathbb{C}^N$ (with respect to the weighted ℓ^p-norm) is defined as*

$$\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^p_w} = \inf_{z \in \mathbb{C}^N} \{ \|x - z\|_{\ell^p_w} : z \in \Sigma_{\boldsymbol{s},\boldsymbol{M}} \}.$$

This directly generalizes the desired approximation error for the sparse case: if $r = 1$ and $w_1 = \ldots = w_r = 1$, we recover $\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^p_w} = \sigma_s(x)_{\ell^p}$. As before, vectors for which $\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^p_w}$ is sufficiently small are called *compressible* or *approximately sparse* in levels.

2.3.1 Null space property in levels

Inspired by (QCBP), we examine the weighted Quadratically Constrained Basis Pursuit (wQCBP) decoder

$$\min_{z \in \mathbb{C}^N} \|z\|_{\ell^1_w} \text{ subject to } \|Az - y\|_{\ell^2} \leq \eta. \tag{wQCBP}$$

In the case of constant weights $w_1 = \ldots = w_r$, this is exactly QCBP. In the sparse case, we had strong intuition building towards a robust null space property. We continue this thread in the levels case. The following null space property gives guarantees useful for wQCBP, and will also appear later in recovery guarantees for other methods. Before stating this property, we define following key quantities:

$$\zeta = \sum_{i=1}^{r}(w^{(i)})^2 s_i, \qquad \xi = \min_{i=1,\dots,r}(w^{(i)})^2 s_i. \tag{2.20}$$

Definition 13. *A matrix $A \in \mathbb{C}^{m\times N}$ has the weighted rNSP (wrNSPL) in levels of order $(\boldsymbol{s},\boldsymbol{M})$ with constants $0<\rho<1$ and $\tau>0$ if*

$$\|P_\Delta v\|_{\ell^2} \le \frac{\rho\|P_{\Delta^c}v\|_{\ell^1_w}}{\sqrt{\zeta}} + \tau\|Av\|_{\ell^2},$$

for all $v\in\mathbb{C}^n$ and $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$.

This definition is due to see [9, Defn. 5.1]. In the case one level, unweighted case, this definition simplifies to the rNSP for sparse vectors (2.2). As we had hoped, this definition is sufficient to give a generalized recovery result.

Theorem 7. *(derived from [9, Theorem 5.4]) Suppose A has the wrNSP in levels of order $(\boldsymbol{s},\boldsymbol{M})$ with constants $0<\rho<1$ and $\tau>0$. Then for any $x\in\mathbb{C}^N$, a solution $\hat{x}$ of wQCBP with $y=Ax+e$, and $\|e\|_{\ell^2}\le\eta$, satisfies*

$$\|x-\hat{x}\|_{\ell^1_w} \le \frac{2(1+\rho)}{(1-\rho)}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \frac{4}{1-\rho}\tau\sqrt{\zeta}\eta,$$
$$\|x-\hat{x}\|_{\ell^2} \le \left(1+(\zeta/\xi)^{1/4}\right)\left(\frac{1}{\sqrt{\xi}}\frac{2(1+\rho)}{(1-\rho)}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \frac{4}{1-\rho}\tau\eta\right).$$

As before, this result not only gives the robustness we desired, but the stability as well.

2.3.2 Restricted isometry in levels

Proceeding exactly as the sparse case, it can be difficult to verify the wrNSPL for a given matrix A. So, we extend the RIP to the Restricted Isometry Property in Levels (RIPL).

Definition 14. *Let $\boldsymbol{M}=(M_1,\dots,M_r)$ be sparsity levels and $\boldsymbol{s}=(s_1,\dots,s_r)$ be local sparsities. The $(\boldsymbol{s},\boldsymbol{M})$-th Restricted Isometry Constant in Levels (RICL) $\delta_{\boldsymbol{s},\boldsymbol{M}}$ of a matrix $A\in\mathbb{C}^{m\times N}$ is the smallest $\delta\ge 0$ such that*

$$(1-\delta)\|x\|_{\ell^2}^2 \le \|Ax\|_{\ell^2}^2 \le (1+\delta)\|x\|_{\ell^2}^2, \quad \forall x\in\Sigma_{\boldsymbol{s},\boldsymbol{M}}. \tag{2.21}$$

If $0<\delta_{\boldsymbol{s},\boldsymbol{M}}<1$ then the matrix is said to have the Restricted Isometry Property in Levels (RIPL) of order $(\boldsymbol{s},\boldsymbol{M})$.

The RIPL measures how close A is to an isometry on the set $\Sigma_{\boldsymbol{s},\boldsymbol{M}}$, much how the RIP measured this for sparse vectors. An alternate characterization of the RICL is the following.

Lemma 3. *The* $(\boldsymbol{s}, \boldsymbol{M})$*-th RICL constant of a matrix* A *satisfies*

$$\delta_{\boldsymbol{s},\boldsymbol{M}} = \sup_{\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}} \|P_\Delta A^* A P_\Delta - P_\Delta\|_{\ell^2}. \tag{2.22}$$

Here, we will prove this lemma. Note that this proof also shows the result for the sparse case, Lemma 1.

Proof. By inspection, Eq. (2.21) is equivalent to

$$\left| \|Ax\|_{\ell^2}^2 - \|x\|_{\ell^2}^2 \right| \leq \delta \|x\|_{\ell_2}^2, \qquad \forall x \in \Sigma_{\boldsymbol{s},\boldsymbol{M}}, \tag{2.23}$$

and thus

$$\delta_{\boldsymbol{s},\boldsymbol{M}} = \sup_{\substack{x \in \Sigma_{\boldsymbol{s},\boldsymbol{M}} \\ \|x\|_{\ell^2}=1}} \left| \|Ax\|_{\ell^2}^2 - \|x\|_{\ell^2}^2 \right| = \sup_{\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}} \sup_{\substack{\operatorname{supp}(x) \subset \Delta \\ \|x\|_{\ell^2}=1}} \left| \|Ax\|_{\ell^2}^2 - \|x\|_{\ell^2}^2 \right|. \tag{2.24}$$

Thus consider arbitrary $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$ with $\operatorname{supp}(x) \subset \Delta$. Then

$$\left| \|Ax\|_{\ell^2}^2 - \|x\|_{\ell^2}^2 \right| = \left| \|AP_\Delta x\|_{\ell^2}^2 - \|P_\Delta x\|_{\ell^2}^2 \right| = \left| \langle (P_\Delta A^* A P_\Delta - P_\Delta) x, x \rangle \right|.$$

And, noting that the matrix $Q = P_\Delta A^* A P_\Delta - P_\Delta$ is self adjoint, we have $\|Q\|_{\ell^2} = \sup_{\|x\|_{\ell^2}=1} |\langle Qx, x \rangle\rangle|$, giving precisely

$$\sup_{\substack{\operatorname{supp}(x) \subset \Delta \\ \|x\|_{\ell^2}=1}} \left| \|Ax\|_{\ell^2}^2 - \|x\|_{\ell^2}^2 \right| = \sup_{\substack{\operatorname{supp}(x) \subset \Delta \\ \|x\|_{\ell^2}=1}} |\langle (P_\Delta A^* A P_\Delta - P_\Delta) x, x \rangle| = \|P_\Delta A^* A P_\Delta - P_\Delta\|_{\ell^2},$$

and thus the result follows immediately. ☐

Both of these results, again, directly generalize theory from the sparse case. As before, we now need the crucial ingredient: to show the RIPL implies the wRNSP, and thus stable and robust recovery.

Theorem 8. *Suppose that the* $(2\boldsymbol{s}, \boldsymbol{M})$*-th RICL constant of* $A \in \mathbb{C}^{m \times N}$ *satisfies*

$$\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{2\zeta/\xi} + 1}, \tag{2.25}$$

where ζ *and* ξ *are as in* (2.20)*, for* $w \in \mathbb{R}^N$*, with* $w > 0$*, be a set of weights as in* (2.19)*. Then* A *satisfies the wrNSPL of order* $(\boldsymbol{s}, \boldsymbol{M})$ *with constants*

$$\rho = \sqrt{2} \frac{\delta_{2\boldsymbol{s},\boldsymbol{M}}}{1 - \delta_{2\boldsymbol{s},\boldsymbol{M}}}, \qquad \tau = \frac{\sqrt{1 + \delta_{2\boldsymbol{s},\boldsymbol{M}}}}{1 - \delta_{2\boldsymbol{s},\boldsymbol{M}}}. \tag{2.26}$$

And following from the previous discussion of the wRNSPL, we have recovery guarantees from the RIPL:

Corollary 3. *Suppose that the* $(2\boldsymbol{s},\boldsymbol{M})$*-th RICL constant of* $A \in \mathbb{C}^{m\times N}$ *satisfies*

$$\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{2\zeta/\xi}+1}, \tag{2.27}$$

where ζ *and* ξ *are as in* (2.20)*, and let* $\eta \geq 0$ *and* $w \in \mathbb{R}^N$*, with* $w > 0$*, be a set of weights as in* (2.19)*. Then, for all* $x \in \mathbb{C}^N$ *any minimizer of* $\hat{x}$ *of* (wQCBP) *for* $y = Ax + e$ *and* $\|e\|_{\ell^2} \leq \eta$ *satisfies*

$$\begin{aligned} \|x-\hat{x}\|_{\ell^1_w} &\leq C\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D\sqrt{\zeta}\eta, \\ \|x-\hat{x}\|_{\ell^2} &\leq \left(1+(\zeta/\xi)^{1/4}\right)\left(\frac{E}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + F\eta\right), \end{aligned} \tag{2.28}$$

where C, D, E, F *depend on* $\delta_{2\boldsymbol{s},\boldsymbol{M}}$ *only.*

The proof of Theorem 8 will conclude this chapter.

2.3.3 Constructing matrices satisfying the RIPL

Even in the general levels setting, Gaussian random matrices satisfy the RIPL. In fact, a Gaussian matrix satisfies the RIPL of order $\delta_{\boldsymbol{s},\boldsymbol{M}} \leq \delta$ with probability at least $1-\epsilon$, provided

$$m \geq C\delta^{-2}\left(\sum_{k=1}^{r} s_k \log\left(\frac{\mathrm{e}(M_k - M_{k-1})}{s_k}\right) + \log(\epsilon^{-1})\right), \tag{2.29}$$

which follows as a corollary of [22] as noted in [29]. This is promising for the levels model, as it reduces the number of measurements needed. More concretely, for the sparse case we required $s\log(eN/s)$ measurements. In the levels case, this translates to $s\log(eN/s) = \sum_{k=1}^{r} s_k \log(eN/s)$. To see if there is any gain to be had, we compare the terms of this previous result with the sum above. For example, if some level satisfies $s_k = M_k - M_{k-1}$, this contributes $O(s_k)$ to the sum above, as compared to the usual $O(s_k\log(eN/s))$. Thus, the measurement condition for the RIPL here requires a smaller m than was required for the RIP.

As before, this approach has two fundamental flaws. Gaussian matrices are infeasible in large problems to efficiently store or multiply, and real applications often do not permit Gausssian sampling. Thus, we need to develop some new notion of a subsampled isometry. For this, the notion of a multilevel sampling scheme is required:

Definition 15. *Let* $\boldsymbol{N} = (N_1,\ldots,N_r)$*, where* $1 \leq N_1 \leq \ldots < N_r = N$ *and* $\boldsymbol{m} = (m_1,\ldots m_r)$*, where* $m_k \leq N_k - N_{k-1}$ *for* $k = 1,\ldots,r$ *and* $N_0 = 0$*. Nn* $(\boldsymbol{m},\boldsymbol{N})$*-multilevel sampling scheme is a set* $\Omega = \Omega_1 \cup \cdots \cup \Omega_r$ *of* $m = m_1 + \ldots + m_r$ *indices, where for* *each k the following holds. If $m_k = N_k - N_{k-1}$ then $\Omega_k = \{N_{k-1}+1, \dots, N_k\}$, and otherwise Ω_k consists of m_k indices chosen independently and uniformly at random from the set $\{N_{k-1}+1, \dots, N_k\}$.*

How does this relate to our problem of recovering sparse in levels vectors? In the case where $\boldsymbol{N} = \boldsymbol{M}$, we hope (intuitively) that we can spend m_k measurements in the k-th sampling level $\{M_{k-1}+1, \dots, M_k\}$ to recover the s_k nonzero entries of an $(\boldsymbol{s}, \boldsymbol{M})$-sparse vector x. This leads to the extension of a multilevel subsampled unitary matrix:

Definition 16. *A matrix $A \in \mathbb{C}^{m \times N}$ is an $(\boldsymbol{m}, \boldsymbol{N})$-multilevel subsampled unitary matrix if $A = P_\Omega D U$ for a unitary matrix $U \in \mathbb{C}^{N \times N}$ and $(\boldsymbol{m}, \boldsymbol{N})$ multilevel random sampling scheme Ω. Here, D is a diagonal scaling matrix with entries*

$$d_{ii} = \sqrt{\frac{N_k - N_{k-1}}{m_k}}. \tag{2.30}$$

An important remark on Definition 16 is that there are two rather different types of levels: those which have every index included, the so-called *saturated* level, and those which are randomly sampled and are *unsaturated*. This is an important distinction, as the former is explicitly not random in any sense. Otherwise, randomly sampling until we fully saturated a level would fall afoul of the coupon collectors effect – requiring $(N_k - N_{k-1})\log((N_k - N_{k-1}))$ measurements.

While coherence is the correct tool for sparse recovery, it does not capture the local behavior of our multilevel scheme, leading to the more general definition:

Definition 17. *The (k,l)-th local coherence of the matrix U is the coherence of the block defined by the k-th sampling level and the l-th sparsity level. That is*

$$\mu_{kl}(U) := \mu(P_{N_k}^{N_{k-1}} U P_{M_l}^{M_{l-1}}).$$

The block $P_{N_k}^{N_{k-1}} U P_{M_l}^{M_{l-1}}$ of the matrix U is for convenience denoted $U^{(k,l)}$, so that we may write

$$U = \begin{pmatrix} U^{(1,1)} & U^{(1,2)} & \dots & U^{(1,r)} \\ U^{(2,1)} & U^{(2,2)} & \dots & U^{(2,r)} \\ \vdots & \vdots & \ddots & \vdots \\ U^{(r,1)} & U^{(r,2)} & \dots & U^{(r,r)} \end{pmatrix}$$

This notion of local coherence is precisely the correct idea to imply the RIPL. As a result, this coherence gives recovery guarantees for wQCBP with a well chosen set of weights. In the general multilevel subsampled unitary matrix case, we have

Theorem 9. *([29, Theorem 3.1]) Let $\epsilon, \delta \in (0,1)$, $x \in \mathbb{C}^N, U \in \mathbb{C}^{N\times N}$ be unitary. Take A to be a multilevel subsampled unitary matrix as in Definition 16, and suppose*

$$m_k \gtrsim \delta^{-2}(N_k - N_{k-1}) \left(\sum_{i=1}^{r} s_l \mu_{kl}(U)\right) \cdot L, \qquad k = 1, \ldots, r, \tag{2.31}$$

where $L = r\log(2m)\log(2N)\log^2(2s) + \log(\epsilon^{-1})$ and $m = m_1 + \ldots + m_r$. Then, with probability $1-\epsilon$, The matrix A satisfies the RIPL of order $(\boldsymbol{s}, \boldsymbol{M})$ with RICL constant $\delta_{\boldsymbol{s},\boldsymbol{M}} \le \delta$.

2.3.4 The Fourier-Haar example

To illustrate, we now consider a concrete example of this setup. This example will be rather terse, as to avoid developing the full tools of wavelet theory, but should serve as an inspiration as to why these levels-based approaches are useful.

Consider a one dimensional function f that is smooth on a finite number of subintervals of $[0,1]$. This is a very simplistic model of a 1-D image – with a finite number of edges. The first task is to find some basis in which f is approximately sparse – and wavelets are a perfect candidate. The so called *Haar basis* or *Haar wavelet basis* is a basis of $L^2([0,1])$ given by shifts and scalings of the Haar function

$$h(x) = \begin{cases} 1 & 0 < x \le 1/2 \\ -1 & 1/2 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

This gives a basis of $L^2([0,1])$ indexed by $j \in \mathbb{Z}$ and k as $h_{j,k}(x) = 2^{j/2}h(2^j x - k)$ for $k = 0, \ldots 2^j - 1$, with the additional basis element $\phi(x) = 1$, called the scaling function.

Here the choice of index with two parameters is an instructive one. The parameter j is the *scale*: the haar wavelets at scale j are supported on intervals of size 2^{-j}. Thus, this may be also referred to as the resolution of the approximation. This basis having localized support implies the expansion of f in this basis reveals local behavior – wavelet coefficients are large if their support intersects some local change in f, such as an edge. This is interesting in particular our model of images, as there will be many small wavelet coefficients, capturing the smoothly varying regions of f, and much fewer large coefficients capturing the edges. For example, consider Fig. 2.2, plotting the percentage of wavelet coefficients per scale greater than threshold 10^{-3} for a natural image f. As the scale increases, the number of non-negligible coefficients decreases: these are only the wavelets with support intersecting the edges of f.

In fact, the coefficients of f in a wavelet basis decay asymptotically like $2^{-j/2}$ if the corresponding wavelet has suppoer intersecting an edge, and $2^{-3j/2}$ otherwise. This means

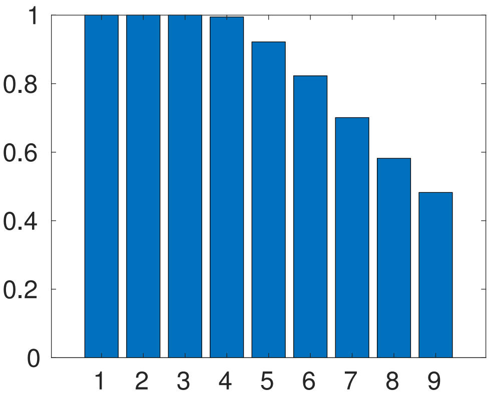


Figure 2.2: Percentage of wavelet coefficients above $\epsilon = 10^{-3}$ per level for image f. Reproduced with the authors permission from [6]

asymptotically, there are very few non-negligible coefficients. This is, the coefficient vector of f is asymptotically compressible in levels, in a naturally occurring levels structure given by the wavelet scales!

To phrase this as a sparse in levels recovery problem, let us truncate the series representation of f at some scale r to write

$$f(x) = d\phi + \sum_{j=0}^{r-1} \sum_{k=0}^{2^{j-1}} c_{jk} h_{j,k}(x). \tag{2.32}$$

Putting aside indexing for a moment, we have $f = \sum_{i=1}^{2^r} c_i h_i(x)$. Sampling f on a uniform grid of 2^r points gives an equation for each gridpoint, and thus a system of equations $f = Hc$. Here each row is simply the evaluation of the expansion of f at each gridpoint, and each column of H corresponds to a specific wavelet.

Now we consider attempting to recover this c from some set of linear measurements. In the case of MRI for example, this is prescribed as Fourier samples. However, this approach is also theoretically sound – Fourier sampling performs well even when not enforced by application. This comes from the observation that the DFT of a wavelet decomposition has, at the wavelet scale j has frequency support essentially contained in the j-th frequency band $B_j = \{-2^j + 1, \dots, -2^{j-1}\} \cup \{2^{j-1}, \dots, -2^j\}$. Phrased intuitively, the wavelets at high scales capture high frequency behavior – the finer the resolution; the better objects like edges are characterized.

By having sampling levels corresponding to these dyadic frequency bands, and sparsity levels corresponding to the wavelet scales, we recover an $(\boldsymbol{N}, \boldsymbol{M})$-multilevel sampling scheme. Concretely, $M_i = N_i = 2^i$, with corresponding matrix $U = FH$, for F the DFT matrix and H the Haar matrix above.

It can be shown that the local coherences of this matrix satisfy [9]

$$\mu(U^{(k,l)}) \lesssim \begin{cases} 2^{-(k-l)} & l \leq k \\ 2^{-3(l-k)} & l > k \end{cases},$$

which by substitution into our previous theory gives a measurement condition

$$m_k \gtrsim \left(s_k + \sum_{l=1}^{k-1} s_l 2^{-(k-1)} + \sum_{l=k+1}^{r} s_l 2^{-3(l-k)} \right) \cdot L \text{ for } k = 1, \ldots, r,$$

where $L = \log(N/\epsilon) + \log(s) + \log(s/\epsilon)$.

This condition is quite useful, as it specifies how to choose m bases on the local sparsities. Particularly in the natural imaging context where one has upper bounds on the s_k, this provides a simple recipe for m_k. As one might expect, tighter bounds on the local sparsities allow for reducing of the number of measurements m.

To remark on this more directly, this is a crucial step for approximating natural images with a number of samples $m = s + o(s)$, saving a log factor over what was previous considered optimal. This approach is twofold. First, the coarse, saturated, levels are sampled deterministically. Further, the fine scales are sampled at a rate proportional to the number of discontinuities of f times $\log(s_k)$, seeking to only recover those fine scale coefficients corresponding to wavelets intersecting the discontinuities. Overall, this leads to an intuitive sampling complexity $s + o(s)$. Proving this formally recovers f requires significant mathematical technique beyond the scope of this thesis, but was previously tackled by the Author, Adcock, and Brugiapaglia in [3], wherein the author of this thesis introduced much of the groundwork for the approach.

2.3.5 Proof of Theorem 8

Proof. We begin by performing a decomposition in each level. For $l = 1, \ldots r$, let $\Xi_{0,l}$ be the index set to the largest s_l entries of $P_{M_l}^{M_{l-1}} x$ in absolute value. Then, define $\Xi^{(0)} = \Xi_{0,1} \cup \ldots \cup \Xi_{0,r}$. For such index set, we decompose $(\Xi^{(0)})^c = \{1, \ldots, N\} \setminus \Xi^{(0)}$, letting

$$\Xi_{0,l}^c = \Xi_{1,l} \cup \Xi_{2,l} \cup \ldots,$$

where $\Xi_{1,l}$ is index set of the s_l largest entries of $P_{\Xi_{0,l}^c} x$, $\Xi_{2,l}$ is the index set of the largest s_l entries of $P_{(\Xi_{0,l} \cup \Xi_{1,l})^c} x$, and so on, letting $\Xi_{i,l} = \emptyset$ as needed for sufficiently large i. (Note that $\Xi_{0,l}^c$, $(\Xi_{0,l} \cup \Xi_{1,l})^c$, etc. are relative complements with respect to the level l). Finally we define $\Xi^{(i)} = \Xi_{i,1} \cup \ldots \cup \Xi_{i,r}$ for each $i = 1, 2, \ldots$. Then

$$\|P_{\Xi^{(0)} \cup \Xi^{(1)}} x\|_{\ell^2}^2 \leq \frac{1}{1 - \delta_{2\boldsymbol{s},\boldsymbol{M}}} \|A P_{\Xi^{(0)} \cup \Xi^{(1)}} x\|_{\ell^2}^2, \tag{2.33}$$

as by assumption A has the RIPL of order $(2\boldsymbol{s}, \boldsymbol{M})$. Then expanding according to the partition as in [6, Theorem 5.13]

$$\|AP_{\Xi^{(0)}\cup\Xi^{(1)}}x\|_{\ell^2}^2 \leq \sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}\|P_{\Xi^{(0)}\cup\Xi^{(1)}}x\|_{\ell^2}^2\|Ax\|_{\ell^2}+\sum_{i\geq 2}|\langle AP_{\Xi^{(0)}\cup\Xi^{(1)}}x, AP_{\Xi^{(i)}}x\rangle|. \quad (2.34)$$

Now using that $|\langle AP_{\Xi^{(0)}\cup\Xi^{(1)}}x, P_{\Xi^{(i)}}x\rangle| = 0$ in tandem with Part (i) of Lemma 4 – which here may be found in Chapter 4 – we see

$$\begin{aligned}|\langle AP_{\Xi^{(0)}\cup\Xi^{(1)}}x, AP_{\Xi^{(i)}}x\rangle| &\leq \delta_{2\boldsymbol{s},\boldsymbol{M}}(\|P_{\Xi^{(0)}}x\|_{\ell^2}+\|P_{\Xi^{(1)}}x\|_{\ell^2})\|P_{\Xi^{(i)}}x\|_{\ell^2}\\ &\leq \sqrt{2}\delta_{2\boldsymbol{s},\boldsymbol{M}}\|P_{\Xi^{(0)}\cup\Xi^{(1)}}x\|_{\ell^2}\|P_{\Xi^{(i)}}x\|_{\ell^2}.\end{aligned}$$

Furthermore using bounds (2.33) and (2.34) and the RIPL we have

$$\|P_{\Xi^{(0)}\cup\Xi^{(0)}}x\|_{\ell^2}^2 \leq \frac{\sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\|Ax\|_{\ell^2}+\sqrt{2}\frac{\delta_{2\boldsymbol{s},\boldsymbol{M}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\sum_{i\geq 2}\|P_{\Xi^{(i)}}x\|_{\ell^2}. \quad (2.35)$$

Recalling our goal is the show the weighted robust null space property, we need to relate the summation in the latter term to $\|P_{\Delta}^{c}x\|_{\ell^1_w}$, where $\Delta = \Xi^{(0)}$. But by construction

$$\|P_{\Xi_{i,l}}x\|_{\ell^2} \leq \sqrt{s_l}\|P_{\Xi_{i,l}}x\|_{\ell^\infty} \leq \sqrt{s_l}\min_{j\in\Xi_{i-1,l}}|x_j| \leq \frac{\|P_{\Xi_{i-1,j}}x\|_{\ell^1}}{\sqrt{s_l}} = \frac{\|P_{\Xi_{i-1,l}}x\|_{\ell^1_w}}{w_l\sqrt{s_l}}.$$

Thus overall, we have

$$\|P_{\Xi^{(i)}}\|_{\ell^2}^2 = \sum_{l=1}^{r}\|P_{\Xi_{i,l}}x\|_{\ell^2}^2 \leq \sum_{l=1}^{r}\left(\frac{\|P_{\Xi_{i-1,l}}x\|_{\ell^1_w}}{w_l\sqrt{s_l}}\right)^2 \leq \frac{1}{\xi}\|P_{\Xi^{(i-1)}}x\|_{\ell^1_w}^2.$$

And hence,

$$\sum_{i\geq 2}\|P_{\Xi_{i,l}}x\|_{\ell^2} \leq \frac{1}{\sqrt{\xi}}\sum_{i\geq 2}|P_{\xi^{(i-1)}}x\|_{\ell^1_w} = \frac{1}{\sqrt{\xi}}\|P_{\Delta^c}x\|_{\ell^1_w}.$$

Combining this with gives that

$$\begin{aligned}\|P_{\Delta}x\|_{\ell^2} \leq \|P_{\Xi^{(0)}\cup\Xi^{(0)}}x\|_{\ell^2} &\leq \frac{\sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\|Ax\|_{\ell^2}+\sqrt{2}\frac{\delta_{2\boldsymbol{s},\boldsymbol{M}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\frac{1}{\sqrt{\xi}}\|P_{\Delta^c}x\|_{\ell^1_w}\\ &= \frac{\sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\|Ax\|_{\ell^2}+\sqrt{2}\frac{\delta_{2\boldsymbol{s},\boldsymbol{M}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\frac{\sqrt{\zeta}}{\sqrt{\xi}}\frac{\|P_{\Delta^c}x\|_{\ell^1_w}}{\sqrt{\zeta}}.\end{aligned}$$

Hence, A has the wrNSPL provided

$$\sqrt{2}\frac{\delta_{2\boldsymbol{s},\boldsymbol{M}}}{1-\delta_{2\boldsymbol{s},\boldsymbol{M}}}\frac{\sqrt{\zeta}}{\sqrt{\xi}} < 1,$$

or namely $\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{\frac{2\zeta}{\xi}}+1}$, as required. □

Chapter 3

Iterative and greedy methods

3.1 Outline

In this chapter, we will develop the key objects of interest of this work. In Section 3.1, we review and develop the methods of interest in the sparse case, building intuition and contextualizing the different approaches. This will culminate in the standard recovery results for the sparse case, Theorem 10 and Theorem 11. Next, the extension to the levels case will proceed in Section 3.2. This will lead to the formulations of iterative hard thresholding in levels, compressive sampling pursuit in levels, and orthogonal matching pursuit in levels. The analysis of these methods is found in Chapter 4.

3.2 The sparse case

Until this point we have developed our theory in tandem with QCBP. However, any optimization based recovery suffers from the same fundamental issue: we need to choose some algorithm to solve the underlying optimization problem. What we would instead prefer is an algorithm with stability and robustness results itself - thus skipping over this "middleman" of an optimization scheme. Thus we turn to iterative and greedy methods to solve our sparse recovery problem. Much like the optimization arena, there are many potential directions in which to proceed.

3.2.1 Developing OMP and CoSaMP

A first, and most classical, approach is that of *Orthogonal Matching Pursuit* (OMP), introduced in [35], with recovery results proven in [38]. This method is inspired by the idea of sparse approximation: if $y = Ax$ for s-sparse x, y is a linear combination of exactly s columns of A. So we need to determine which columns these are, and a first idea is to find the support of x index by index. That is, to construct approximations $x^{(n)}$ to x at step n, which have exactly n nonzero entries. While one can imagine many ways of doing this, a reasonable idea is to find the column of A most strongly correlated with the residual $y - Ax^{(n)}$, in the hopes that this column gives a new index from the support of x. Finally after finding the candidate index set, a least squares solution on this support set can be performed to find the current best sparse approximation. Ideally at the end of s iterations, the algorithm will have identified first the support of x, and then its entries via least squares.

This approach is very different from optimization schemes such as wQCBP, in that it always terminates after a fixed number of iterations s. Notably then these intermediate least squares problems never exceed size $m \times s$. These together can, in certain cases (e.g., when the target sparsity s is very small), save significant computational time. For example, in the case where A is dense and unstructured, OMP has a running time proportional to $O(mNs)$ whereas basis pursuit requires $O(N^2 s^{3/2})$ operations [21].

Function $\hat{x} = \mathrm{OMP}(A, y, s)$
Inputs: $A \in \mathbb{C}^{m\times N}$, $y \in \mathbb{C}^m$, sparsity s
Initialization: $x^{(0)} \in \mathbb{C}^N$ (e.g. $x^{(0)} = 0$) , $S^{(0)} = \emptyset$
Iterate: For each $k = 1, \dots, s$, set

$$j_k \in \operatorname*{argmax}_{j=1,\dots,N} |(A^*(y - Ax^{(k-1)}))_j|$$

Update $S^{(k)} = S^{(k-1)} \cup \{j_k\}$
Set $x^{(k)} \in \operatorname{argmin} \left\{ \|y - Az\|_{\ell^2} \;:\; z \in \mathbb{C}^N \text{ s.t. } \operatorname{supp}(z) \subseteq S^{(k)} \right\}$
Output: $\hat{x} = x^{(s)}$

A final interesting property of OMP is that we reduce the size of the residual at each iteration. Namely for all k,

$$\|y - Ax^{(k)}\|_{\ell^2}^2 \leq \|y - Ax^{(k-1)}\|_2^2 - |(A^*(y - Ax^{(k-1)}))_j|, \tag{3.1}$$

which can be found for example in [27, Lemma 3.3]. Thus one can see OMP in a slightly different lens: this is a greedy selection which reduces the ℓ^2 norm of the residual $y - Ax^{(k)}$ as much as possible per iteration, in the sense above.

A next approach is inspired again by an iteration which also seeks minimizes to the residual $\|y - Ax\|_{\ell^2}$. One obvious weakness of OMP is that if an incorrect choice of index is made at some step, it remains in the approximation for all future steps. Compressive sampling matching pursuit (CoSaMP) attempts to tackle this issue. As before, the indices columns most correlated with the residual are chosen as candidates for the support of x. But instead of simply taking a single index at each step, $2s$ are added to the running support set before least squares fitting. Then, a thresholding is performed to ensure this vector is s-sparse. This algorithm was proposed in the late 2000's [32], inspired by the so-called *regularized orthogonal matching pursuit* algorithm[33, 34]. In what follows, we use the index set $L_{2s}(x)$ to denote the indices of the $2s$ largest entries of x, in absolute value.

To enforce sparsity, we use the *hard thresholding* operator. For a vector $x \in \mathbb{C}^N$ (not necessarily sparse), let $L_s(x)$ be the index set of its s largest entries in absolute value. The *hard thresholding* operator $H_s : \mathbb{C}^N \to \mathbb{C}^N$ is, for $x = (x_i)_{i=1}^N \in \mathbb{C}^N$, is defined by

$$H_s(x) = (H_s(x)_i)_{i=1}^N, \qquad H_s(x)_i = \begin{cases} x_i & i \in L_s(x) \\ 0 & \text{otherwise} \end{cases}.$$

Function CoSaMP(A, y, s)
Inputs: $A \in \mathbb{C}^{m\times N}$, $y \in \mathbb{C}^m$, sparsity s
Initialization: $x^{(0)} \in \mathbb{C}^N$ (e.g. $x^{(0)} = 0$)
Iterate: Until some stopping criterion is met at $n = \overline{n}$, set

$$\begin{aligned} U^{(n+1)} &= \operatorname{supp}(x^{(n)}) \cup L_{2s}(A^*(y - Ax^{(n)})) \\ u^{(n+1)} &\in \underset{z\in\mathbb{C}^N}{\operatorname{argmin}}\{\|y - Az\|_2 \,:\, \operatorname{supp}(z) \subset U^{(n+1)}\} \\ x^{(n+1)} &= H_s(u^{(n+1)}) \end{aligned}$$

Output: $\hat{x} = x^{(\overline{n})}$

3.2.2 Developing IHT

One of the most simplistic ideas is perhaps that of *Iterative hard thresholding* (IHT). There are approaches to develop IHT, but one simple way is through the lens of a fixed point iteration. Suppose $y = Ax$. Then, we can make the simple rearrangement of $0 = y - Ax = A^*(y - Ax)$. And, adding x to both sides, we have that a solution to our linear system should satisfy the fixed point equation

$$x = x + A^*(y - Ax).$$

Then, from this argument we know a sparse solution satisfies

$$x = H_s(x + A^*(y - Ax)).$$

And, to inspire an iteration scheme, we use this as a fixed point iteration for x, namely

$$x^{(n+1)} = H_s(x^{(n)} + A^*(y - Ax^{(n)})).$$

This is a remarkably simple idea - and is perhaps surprising that such a strategy could work well. We can define the algorithm concretely as follows:

Function $\hat{x} = \mathrm{IHT}(A, y, s)$

Inputs: $A \in \mathbb{C}^{m \times N}$, $y \in \mathbb{C}^m$, sparsity s

Initialization: $x^{(0)} \in \mathbb{C}^N$ (e.g. $x^{(0)} = 0$)

Iterate: Until some stopping criterion is met at $n = \overline{n}$, set

$$x^{(n+1)} = H_s(x^{(n)} + A^*(y - Ax^{(n)}))$$

Output: $\hat{x} = x^{(\overline{n})}$

IHT was introduced in the context of compressed sensing in the late 2000s [13, 14]. Improvements to IHT with variable stepsizes have also been introduced, [12, 15], as have been generalizations to other structured sparsity models [10, 28]. It is important to note that IHT requires some estimation of the sparsity s, but does not reference any noise level η, the exact opposite of the QCBP decoder. Having a reasonable estimate of either of these parameters depends on the problem one wishes to solve. More strikingly for QCBP, the optimal choice of parameter η is exactly the noise level [40].

3.2.3 Recovery guarantees for IHT and CoSaMP.

The analysis of these algorithms hinges on the RIP. With suitable assumptions, each of these exhibit stable and robust recovery guarantees.

Theorem 10. *(E.g. [27, Theorem 6.21]) Suppose that the* $6s$*-th RIC constant of* $A \in \mathbb{C}^{m \times N}$ *satisfies* $\delta_{6s} < \frac{1}{\sqrt{3}}$*. Then, for all* $x \in \mathbb{C}^N$ *and* $e \in \mathbb{C}^m$*, the sequence* $(x^{(n)})_{n \geq 0}$ *defined by* $\mathrm{IHT}(A, y, 2s)$ *with* $y = Ax + e$ *and* $x^{(0)} = 0$ *satisfies, for any* $n \geq 0$*,*

$$\|x - x^{(n)}\|_{\ell^1} \leq C\sigma_s(x)_{\ell^1} + D\sqrt{s}\|e\|_{\ell^2} + 2\sqrt{s}\rho^n\|x\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \leq \frac{C}{\sqrt{s}}\sigma_s(x)_{\ell^1} + D\|e\|_{\ell^2} + \rho^n\|x\|_{\ell^2},$$

where $\rho = \sqrt{3}\delta_{6s} < 1$*, and* $C, D > 0$ *are constants only depending on* δ_{6s}*.*

Theorem 11. *(E.g. [27, Theorem 6.28]) Suppose that the* $8s$*-th RIC constant of* A *satisfies*

$$\delta_{8s} < \frac{\sqrt{\frac{11}{3}} - 1}{4} \approx 0.478.$$

Then, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$ the sequence $(x^{(n)})_{n\geq 0}$ defined by $\mathrm{CoSaMP}(A, y, 2s)$ *with $y = Ax + e$ and $x^{(0)} = 0$, satisfies for any $n \geq 0$,*

$$\|x - x^{(n)}\|_{\ell^1} \leq C\sigma_s(x)_{\ell^1} + D\sqrt{s}\|e\|_{\ell^2} + 2\sqrt{s}\rho^n\|x\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \leq \frac{C}{\sqrt{s}}\sigma_s(x)_{\ell^1} + D\|e\|_{\ell^2} + 2\rho^n\|x\|_{\ell^2},$$

where $\rho = \sqrt{\frac{2\delta_{8s}^2(1+3\delta_{8s}^2)}{1-\delta_{4s}^2}} < 1$ and $C, D > 0$ are constants only depending on δ_{8s}.

A similar result for OMP holds. We will not discuss it here in detail, as we will not extend the recovery result to the levels case. The difficulties in generalizing OMP to the levels setting will be discussed at length in the following sections.

3.3 Extension to the levels case

It is natural to develop generalization to the iterative and greedy methods for the levels case, as we still wish to avoid the inherent computational difficulties of optimization approaches.

We begin with IHT and CoSaMP, noting that the *only* explicit reference to the sparsity in each algorithms simply appears in the thresholding operator H_s, or set of largest entries L_s. Thus, fix sparsity levels $\boldsymbol{M} = (M_1, \ldots, M_r)$. Note that any vector $x \in \mathbb{C}^N$ can be written uniquely as $x = \sum_{k=1}^r x_k$, where $x_k \in \mathbb{C}^N$ with $\mathrm{supp}(x_k) \subseteq \{M_{k-1}+1, \ldots, M_k\}$. Now let $\boldsymbol{s} = (s_1, \ldots, s_r)$ be local sparsities. For $x \in \mathbb{C}^N$, we write $L_{\boldsymbol{s},\boldsymbol{M}}(x)$ for the set

$$L_{\boldsymbol{s},\boldsymbol{M}}(x) = \bigcup_{k=1}^{r} L_{s_k}(x_k).$$

In other words, this is the index set consisting, in each level $\{M_{i-1}+1, \ldots, M_i\}$, of the largest absolute s_i entries of x in that level. With this in hand, we define the *hard thresholding in levels* operator $H_{\boldsymbol{s},\boldsymbol{M}} : \mathbb{C}^N \to \mathbb{C}^N$ by

$$H_{\boldsymbol{s},\boldsymbol{M}}(x) = (H_{\boldsymbol{s},\boldsymbol{M}}(x)_i)_{i=1}^N, \qquad H_{\boldsymbol{s},\boldsymbol{M}}(x)_i = \begin{cases} x_i & i \in L_{\boldsymbol{s},\boldsymbol{M}}(x) \\ 0 & \text{otherwise} \end{cases}, \qquad x = (x_i)_{i=1}^N \in \mathbb{C}^N.$$

With these, we can state the algorithms of interest.

The *IHT in Levels (IHTL)* algorithm is

Function $\hat{x} = \mathrm{IHTL}(A, y, \boldsymbol{s}, \boldsymbol{M})$
Inputs: $A \in \mathbb{C}^{m\times N}$, $y \in \mathbb{C}^m$, local sparsities $\boldsymbol{s}$, sparsity levels $\boldsymbol{M}$
Initialization: $x^{(0)} \in \mathbb{C}^N$ (e.g. $x^{(0)} = 0$)
Iterate: Until some stopping criterion is met at $n = \overline{n}$, set

$$x^{(n+1)} = H_{\boldsymbol{s},\boldsymbol{M}}(x^{(n)} + A^*(y - Ax^{(n)}))$$

Output: $\hat{x} = x^{(\overline{n})}$

and *CoSaMP in Levels (CoSaMPL)* is defined by

Function $\hat{x} = \mathrm{CoSaMPL}(A, y, \boldsymbol{s}, \boldsymbol{M})$
Inputs: $A \in \mathbb{C}^{m\times N}$, $y \in \mathbb{C}^m$, local sparsities $\boldsymbol{s}$, sparsity levels $\boldsymbol{M}$
Initialization: $x^{(0)} \in \mathbb{C}^N$ (e.g. $x^{(0)} = 0$)
Iterate: Until some stopping criterion is met at $n = \overline{n}$, set

$$\begin{aligned} U^{(n+1)} &= \mathrm{supp}(x^{(n)}) \cup L_{2\boldsymbol{s},\boldsymbol{M}}(A^*(y - Ax^{(n)})) \\ u^{(n+1)} &\in \operatorname*{argmin}_{z\in\mathbb{C}^N}\{\|y - Az\|_{\ell^2} \,:\, \mathrm{supp}(z) \subset U^{(n+1)}\} \\ x^{(n+1)} &= H_{\boldsymbol{s},\boldsymbol{M}}(u^{(n+1)}) \end{aligned}$$

Output: $\hat{x} = x^{(\overline{n})}$

We again emphasize here that these differ from the non-levels based versions only in the threshold operator and the index set $L_{2\boldsymbol{s},\boldsymbol{M}}$, and do not change the main iteration steps at all. Thus much of the analysis and intuition of these algorithms in the sparse case may still be applied, albeit with care.

OMP also admits a generalization to this new setting, motivated by the desirable feature of terminating after s iterations. The levels situation differs however, in that now one is given a budget of target sparsities $\boldsymbol{s}$. Whereas before we simply chose the index most correlated with the residual, there is a question of how to enforce the levels structure. A single index can be selected at each step, or r indices at each step – one in each level. Thus, the usual operation of greedy index selection becomes more subtle, as it is not immediately obvious how one should select the "best" indices.

To resemble the spirit of OMP most closely, the choice here is the former. Given a budget of local sparsities $s_1, \dots, s_r$, we seek to have s_i nonzero entries in the ith level. As with OMP, we select the index most correlated with the residual at each iteration. The only issue occurs when the column most correlated with the residual lies in a level with s_i nonzero entries already in the approximation – the budget in that level is expended. Thus, the idea is to simply stop selecting any indices that lie in a level with s_i entries already selected, and continue the algorithm. However, other extensions of OMP - using a different criterion to determine which indices to select – may well perform numerically as well as our proposed version.

Within the work done for this thesis, stability and robustness results for OMPL still prove elusive. While we will demonstrate its numerical effectiveness, it is unclear whether the proposed generalization is best. It is possible that this – or another reasonable formulation of OMPL - will admit similar recovery results as in the sparse case.

Function $\hat{x} = \text{OMPL}(A, y, \boldsymbol{s}, \boldsymbol{M})$
Inputs: $A \in \mathbb{C}^{m\times N}$, $y \in \mathbb{C}^m$, local sparsities $\boldsymbol{s}$, sparsity levels $\boldsymbol{M}$ (with $M_r = N$)
Initialization: Choose initial $x^{(0)}$, and set $S^{(0)} = \emptyset$, $\boldsymbol{s}^{(0)} = \mathbf{0}$ and $\mathcal{L} = \emptyset$.
Iterate: For each $k = 1, \ldots, s$, set

$$j_k \in \underset{\substack{j=1,\ldots,N \\ j\notin\{M_{p-1}+1,\ldots M_p\},\forall p\in\mathcal{L}}}{\operatorname{argmax}} |(A^*(y - Ax^{(k-1)}))_j|$$

and denote l as the level such that $j_k \in \{M_{l-1}+1, \ldots M_l\}$.

Update $\boldsymbol{s}^{(k)} = \boldsymbol{s}^{(k-1)} + \boldsymbol{e}_l$, where $\boldsymbol{e}_l$ is the l-th standard unit vector
Update $S^{(k)} = S^{(k-1)} \cup \{j_k\}$
If $\boldsymbol{s}_l^{(k)} = \boldsymbol{s}_l$, then update the set of saturated levels $\mathcal{L} = \mathcal{L} \cup \{l\}$

Set $x^{(k)} \in \operatorname{argmin}\left\{\|y - Az\|_{\ell^2} \,:\, z \in \mathbb{C}^N \text{ s.t. } \operatorname{supp}(z) \subseteq S^{(k)}\right\}$
Output: $\hat{x} = x^{(s)}$

Chapter 4

Analysis of levels based methods

Much as in the sparse case, the analysis of these algorithms is based on the RIPL. The following results are the main contributions of this work, and will be proven at the end of this chapter. These generalize Theorem 10 and Theorem 11, which will serve as a comparison point in our discussion.

Theorem 12. *Suppose that the $(6\boldsymbol{s}, \boldsymbol{M})$-th RICL constant of $A \in \mathbb{C}^{m\times N}$ satisfies $\delta_{6\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{3}}$, and let $w \in \mathbb{R}^N$, with $w > 0$, be a set of weights constant in each level, i.e. as in (2.19). Then, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$, the sequence $(x^{(n)})_{n\geq 0}$ defined by* $\mathrm{IHTL}(A, y, 2\boldsymbol{s}, \boldsymbol{M})$ *with $y = Ax + e$ and $x^{(0)} = 0$ satisfies, for any $n \geq 0$,*

$$\|x - x^{(n)}\|_{\ell^1_w} \leq C\frac{\sqrt{\zeta}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D\sqrt{\zeta}\|e\|_{\ell^2} + 2\sqrt{\zeta}\rho^n\|x\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \leq \frac{E}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + F\|e\|_{\ell^2} + \rho^n\|x\|_{\ell^2},$$

where $\rho = \sqrt{3}\delta_{6\boldsymbol{s},\boldsymbol{M}} < 1$ and $C, D, E, F > 0$ only depend on $\delta_{6\boldsymbol{s},\boldsymbol{M}}$, and ζ, ξ are as in (2.20).

An analogous result holds for CoSaMPL.

Theorem 13. *Suppose that the $(8\boldsymbol{s}, \boldsymbol{M})$-th RICL constant of $A \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{8\boldsymbol{s},\boldsymbol{M}} < \frac{\sqrt{\frac{11}{3}} - 1}{4} \approx 0.478,$$

and let $w \in \mathbb{R}^N$, with $w > 0$, be a set of weights constant in each level as in Theorem 12. Then, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$ the sequence $(x^{(n)})_{n\geq 0}$ constructed by $\mathrm{CoSaMPL}(A, y, 2\boldsymbol{s}, \boldsymbol{M})$ *with $y = Ax + e$ and $x^{(0)} = 0$, satisfies for any $n \geq 0$,*

$$\|x - x^{(n)}\|_{\ell^1_w} \leq C\frac{\sqrt{\zeta}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D\sqrt{\zeta}\|e\|_{\ell^2} + 2\sqrt{\zeta}\rho^n\|x\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \leq \frac{E}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + F\|e\|_{\ell^2} + \rho^n\|x\|_{\ell^2},$$

where ζ *and* ξ *are as in Theorem 12 and* $\rho = \sqrt{\frac{2\delta_{8s,M}^2(1+3\delta_{8s,M}^2)}{1-\delta_{8s}^2}}1$ *and* $C, D, E, F > 0$ *only depend on* $\delta_{8\boldsymbol{s},\boldsymbol{M}}$*.*

The proofs of these main results are at the end of this chapter. However, we first discuss the usefulness and implications of these results.

4.1 Discussion

Firstly, some general comments on the results: what do they mean? Foremost, these result guarantee stable and robust recovery for all sparse in levels vectors using either IHTL or CoSaMPL. These are the first results of this type for any iterative method. These result include sparse vectors, and all the other aforementioned applications, as a special case. Furthermore the weights appear as parameters in both results, but are not used in the algorithms themselves – for a given x these error bounds hold simultaneously for any choice of weights. This is in contrast to wQCBP, which explicitly requires weights to be chosen ahead of time for the algorithm.

There is also no noise assumption: we do not require any a priori knowledge of $\|e\|_{\ell^2}$ – which is very desirable, as many signal recovery problems do not have have any model for the signal noise. In fact, results of similar flavor for basis pursuit, e.g. Eq. (2.28), require explicit estimates of a noise level η. More precisely, given $\|e\|_{\ell^2} \leq \eta$ for wQCBP, Corollary 3 has an ℓ^2-norm error that scales with $(1 + (\zeta/\xi)^{1/4})\eta$ – serving to amplify noise.

It should be noted that the assumptions on the RICL constants for each result are no more stringent than the sparse case. For example for IHT and IHTL, we require $\delta_s, \delta_{\boldsymbol{s},\boldsymbol{M}} < 1/\sqrt{3}$. Similarly, the constant ρ in these more general results has the same dependence on δ from the sparse case. In fact, the results above simplify to nearly *exactly* the sparse case for $r = 1$ and constant weights.

There is a subtle detail that the constants E, F have differing scaling with δ in the levels setting – these constants are derived in Theorem 14, and arise from a different proof technique than was used in the sparse case. To be precise, Theorem 14 proves distance bounds for the ℓ^2 and ℓ^1_w norms independently, whereas the sparse case proves a similar result for all ℓ^p norms, $1 \leq p \leq 2$ simultaneously. Even with this technical detail, for a constant τ depending on δ these constant only differ roughly by

$$C = 1 + \sqrt{8}\tau \neq 1 + \sqrt{2}\tau = E, \qquad D = 2\tau \neq F = \tau.$$

4.1.1 Some useful and interesting cases

There are several useful special cases to consider. For the case of arbitrary number of levels r, but constant weights $w_1 = \ldots = w_r = 1$, we have that

$$\zeta = s, \qquad \xi = \min_{i,\ldots,r} s_i$$

and thus for sufficiently large n our error bounds reduce to

$$\|x - x^{(n)}\|_{\ell^1} \lesssim \frac{\sqrt{s}}{\sqrt{\min_i s_i}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1} + \sqrt{s}\|e\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \lesssim \frac{1}{\sqrt{\min_i s_i}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1} + \|e\|_{\ell^2}.$$

The issue in this case arises when $\min_i s_i$ is small in relation to the total sparsity s, as then the scaling term in front of $\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1}$ becomes large. In fact, to make this factor of moderate size, we would need $0 < \min_i s_i \approx \max_i s_i$. Accordingly, a good choice of weights is realized by making ζ/ξ order one, which results in the error bound in the ℓ^1_w-norm being optimal up to a constant.

Inspired by this idea, another choice of weights – that requires *a priori* knowledge of the local sparsities – is $w_i = \sqrt{s/s_i}$ for $i = 1, \dots, r$. With this choice, we have error bounds where the constant factors only depend on the number of levels r and the total sparsity s. More precisely,

$$\zeta = rs, \qquad \xi = s$$

and thus our bounds reduce to

$$\|x - x^{(n)}\|_{\ell^1_w} \lesssim \sqrt{r} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \sqrt{rs}\|e\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \lesssim \frac{1}{\sqrt{s}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \|e\|_{\ell^2}.$$

Notably we have a dependence scaling with the number of levels.

In both of these results, and in the recovery guarantees for wQCBP (Corollary 3), the factor ζ/ξ appears repeatedly. This ratio is seemingly ubiquitous, and thus a choice of weights to minimize this factor is ideal - such as the one above. In the results for wQCBP the scaling on $\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w}$ improves by a factor of $\sqrt{\zeta/\xi}$. Conversely, the ℓ^2-norm error bound is better for the IHTL and CoSaMPL decoders, by a factor of $(\zeta/\xi)^{1/4}$. However, the scaling on noise - identifying η with $\|e\|_{\ell^2}$ - is the same.

4.1.2 Scaling with ζ, ξ and its implications

Finally, it is worthwhile to point out how the assumptions for each theorem scale with ζ, ξ. For Corollary 3 the assumption on the RICL constant explicitly depends on the parameters ζ, ξ. Contrasting, the assumptions for IHTL and CoSaMPL are independent of these. For example, in the unweighted case the condition (2.27) for QCBP becomes

$$\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{2s/\min_i\{s_i\} + 1}}, \tag{4.1}$$

which depends on the ratio of the total sparsity s and the minimal local sparsity. Conversely, if the weights are chosen as in $w_i = \sqrt{s/s_i}$, the condition (2.27) becomes

$$\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{2r}+1}. \tag{4.2}$$

Observe from (2.29) that the number of measurements that guarantees an RIP generally scales like δ^{-2}. Combining this observation with condition (4.2) suggests that m should scale linearly in r for wQCBP to ensure stable and robust recovery, whereas for IHTL and CoSaMPL the corresponding condition on m would be independent of r.

This scaling is extremely important in applications. Recall the Fourier-Haar example as described at the end of chapter 2. The levels structure therein required r scale roughly like $\log(N)$. This means the measurement condition necessary to apply wQCBP in practice increases by a factor of $\log^2(N)$. As natural imaging problems can become large in N, this additional log factor is far from insignificant.

4.2 Proofs

4.2.1 Outline

As this series of proofs is quite lengthy, we begin by outlining the main steps. Both Theorem 12 and Theorem 13 are direct generalizations of standard sparse results, giving useful bounds involving the RICL constant. Using these, we prove a key result, Theorem 14, which gives conditions on any vector x' and RICL constant $\delta_{\boldsymbol{s},\boldsymbol{M}}$ to guarantee the true solution x and x' are sufficiently close. Using this result, the overall argument for both IHTL and CoSaMPL is similar. In either case we use Theorem 12 and Theorem 13, along with careful tracking of index sets, to show that $x^{(n)} = x'$ satisfies the assumptions of Theorem 14. From this, the final results follow immediately. This style of argument is extended from the sparse case contained in [27].

4.2.2 Preliminary Lemmas

The following two results are based on [27, Lemma 6.16], and [27, Lemma 6.20] respectively.

Lemma 4. *Let $u, v \in \mathbb{C}^N$ be $(\boldsymbol{s}', \boldsymbol{M})$-sparse and $(\boldsymbol{s}'', \boldsymbol{M})$-sparse respectively, and $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$ be arbitrary. Then for any matrix $A \in \mathbb{C}^{m\times N}$,*

$$(i) \qquad |\langle u, (I - A^*A)v\rangle| \leq \delta_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}} \|u\|_{\ell^2} \|v\|_{\ell^2}$$
$$(ii) \qquad \|P_\Delta (I - A^*A)v\|_{\ell^2} \leq \delta_{\boldsymbol{s}+\boldsymbol{s}'',\boldsymbol{M}} \|v\|_{\ell^2}$$

Proof. To show (i) we expand the inner product

$$|\langle u, (I - A^*A)v\rangle| = |\langle u, v\rangle - \langle Au, Av\rangle|$$

and define $\Xi = \mathrm{supp}(u) \cup \mathrm{supp}(v) \in D_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}}$. Then the above may be written as

$$
\begin{aligned}
|\langle u, v\rangle - \langle Au, Av\rangle| &= |\langle P_\Xi u, P_\Xi v\rangle - \langle (AP_\Xi)P_\Xi u, (AP_\Xi)P_\Xi v\rangle| \\
&= |\langle P_\Xi u, (P_\Xi - (P_\Xi A^* A P_\Xi)P_\Xi)v\rangle| \\
&\leq \|P_\Xi u\|_{\ell^2} \|P_\Xi - (P_\Xi A^* A P_\Xi)\|_{\ell^2} \|P_\Xi v\|_{\ell^2}. \qquad (4.3)
\end{aligned}
$$

As $P_\Xi v$ is $(\boldsymbol{s}'', \boldsymbol{M})$-sparse and thus $(\boldsymbol{s}' + \boldsymbol{s}'', \boldsymbol{M})$-sparse, we use that

$$
\delta_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}} = \sup_{\Xi \in D_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}}} \|P_\Xi - P_\Xi A^* A P_\Xi\|_{\ell^2},
$$

to obtain that the right-hand side of (4.3) may be written as

$$
\|P_\Xi u\|_{\ell^2} \|P_\Xi - (P_\Xi A^* A P_\Xi)\|_{\ell^2} \|P_\Xi v\|_{\ell^2} \leq \Xi_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}} \|P_\Xi u\|_{\ell^2} \|P_\Xi v\|_{\ell^2} = \Xi_{\boldsymbol{s}'+\boldsymbol{s}'',\boldsymbol{M}} \|u\|_{\ell^2} \|v\|_{\ell^2},
$$

which gives (i).

For (ii), we note that

$$
\|P_\Delta (I - A^*A)v\|_{\ell^2}^2 = |\langle P_\Delta (I - A^*A)v, (I - A^*A)v\rangle|,
$$

and apply (i) with $u = P_\Delta(I - A^*A)v$, giving

$$
\|P_\Delta (I - A^*A)v\|_{\ell^2}^2 \leq \delta_{\boldsymbol{s}+\boldsymbol{s}'',\boldsymbol{M}} \|P_\Delta (I - A^*A)v\|_{\ell^2} \|v\|_{\ell^2},
$$

and dividing through by $\|P_\Delta(I - A^*A)v\|_{\ell^2}$ gives the desired result. □

Lemma 5. *Let $e \in \mathbb{C}^m$, $A \in \mathbb{C}^{m \times N}$ with RICL $\delta_{\boldsymbol{s},\boldsymbol{M}}$ and $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$. Then*

$$
\|P_\Delta A^* e\|_{\ell^2} \leq \sqrt{1 + \delta_{\boldsymbol{s},\boldsymbol{M}}} \|e\|_{\ell^2}.
$$

Proof. We compute

$$
\|P_\Delta A^* e\|_{\ell^2}^2 = \langle A^* e, P_\Delta A^* e\rangle = \langle e, AP_\Delta A^* e\rangle \leq \|e\|_{\ell^2} \|AP_\Delta A^* e\|_{\ell^2}.
$$

But as $P_\Delta A^* e$ is $(\boldsymbol{s}, \boldsymbol{M})$-sparse we have

$$
\|e\|_{\ell^2} \|AP_\Delta A^* e\|_{\ell^2} \leq \|e\|_{\ell^2} \sqrt{1 + \delta_{\boldsymbol{s},\boldsymbol{M}}} \|P_\Delta A^* e\|_{\ell^2}
$$

and dividing through by $\|P_\Delta A^* e\|_{\ell^2}$ gives the desired result. □

With these is hand, we prove a key result. This theorem is directly extended from the sparse case in [27, Lemma 6.23].

Theorem 14. *Suppose $A \in \mathbb{C}^{m \times N}$ satisfies the RIPL of order $(\boldsymbol{s}, \boldsymbol{M})$ and has RICL $\delta_{\boldsymbol{s},\boldsymbol{M}} < 1$. Let $\kappa, \tau > 0, \lambda \geq 0$ and $e \in \mathbb{C}^m$ be given, and $w \in \mathbb{R}^N$, with $w > 0$, be a set of weights constant on each level, such that $w_i = w^{(k)}$, for $M_{k-1} < i \leq M_k$ and $1 \leq k \leq r$. Suppose we have $x, x' \in \mathbb{C}^N$ such that*

$$x' \in D_{\kappa \boldsymbol{s},\boldsymbol{M}}, \qquad \text{and} \qquad \|P_\Xi x - x'\|_{\ell^2} \leq \tau \|A P_{\Xi^c} x + e\|_{\ell^2} + \lambda,$$

where $\Xi = \Xi_1 \cup \ldots \cup \Xi_r$, and Ξ_i is the index set of the largest $2s_i$ entries of $P_{M_i}^{M_{i-1}} x$. Then, there exist constants $C_{\kappa,\tau}, D_{\kappa,\tau}, F_{\kappa,\tau} > 0$ depending only on κ and τ and $E_\kappa > 0$ depending only on κ such that

$$\|x - x'\|_{\ell^1_w} \leq C_{\kappa,\tau} \frac{\sqrt{\zeta}}{\sqrt{\xi}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D_{\kappa,\tau} \sqrt{\zeta} \|e\|_{\ell^2} + E_\kappa \sqrt{\zeta} \lambda,$$
$$\|x - x'\|_{\ell^2} \leq \frac{F_\tau}{\sqrt{\xi}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau \|e\|_{\ell^2} + \lambda,$$

where

$$\zeta = \sum_{i=1}^r (w^{(i)})^2 s_i, \qquad \xi = \min_{i=1,\ldots,r} (w^{(i)})^2 s_i.$$

Proof. Let us consider some fixed level i, and Ξ defined as above. We consider the case of the weighted 1-norm first. Projecting onto level i gives

$$\|P_{M_i}^{M_{i-1}}(x - x')\|_{\ell^1_w} \leq w^{(i)} \|P_{\Xi_i^c} x\|_{\ell^1} + w^{(i)} \|P_{\Xi_i} x - P_{M_i}^{M_{i-1}} x'\|_{\ell^1},$$

and where $\Xi_i^c = \{M_{i-1} + 1, \ldots, M_i\} \setminus \Xi_i$ is the relative complement of Ξ_i with respect to the level i. We bound the latter term by noting that $P_{\Xi_i} x - P_{M_i}^{M_{i-1}} x'$ is $(2+\kappa)s_i$-sparse, so that

$$w^{(i)} \|P_{\Xi_i} x - P_{M_i}^{M_{i-1}} x'\|_{\ell^1} \leq \sqrt{(2+\kappa) s_i (w^{(i)})^2} \|P_{\Xi_i} x - P_{M_i}^{M_{i-1}} x'\|_{\ell^2}.$$

Further defining $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$ to be the index set of a best $(\boldsymbol{s}, \boldsymbol{M})$-term approximation to x, we bound the former term by $w^{(i)} \|P_{\Xi_i^c} x\|_{\ell^1} \leq w^{(i)} \|P_{\Delta_i^c} x\|_{\ell^1}$. Summing over all levels $i = 1, \ldots, r$ and using the Cauchy-Schwarz inequality gives

$$\|x - x'\|_{\ell^1_w} \leq \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \sqrt{(2+\kappa)\zeta} \|P_\Xi x - x'\|_{\ell^2}.$$

By supposition we have then

$$\|x - x'\|_{\ell^1_w} \leq \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \sqrt{(2+\kappa)\zeta} (\tau \|A P_{\Xi^c} x + e\|_{\ell^2} + \lambda). \tag{4.4}$$

We now perform a particular decomposition of $\Delta^c = \{1, \ldots, N\} \setminus \Delta$, letting

$$\Delta_i^c = \Xi_{i,1} \cup \Xi_{i,2} \cup \ldots,$$

where $\Xi_{i,1}$ is index set of the s_i largest entries of $P_{\Delta_i^c}x$, $\Xi_{i,2}$ is the index set of the largest s_i entries of $P_{(\Delta_i \cup \Xi_{i,1})^c}x$, and so on. (Note that Δ_i^c, $(\Delta_i \cup \Xi_{i,1})^c$, etc. are relative complements with respect to the level i). This allows us to define the collection $\Xi^{(k)}$ for $k = 1, 2, \dots$

$$\Xi^{(k)} = \bigcup_{i=1}^{r} \Xi_{k,i}, \qquad \text{where by construction } \Xi^{(k)} \in D_{\boldsymbol{s},\boldsymbol{M}},$$

and furthermore $\Xi^c = \bigcup_{k\geq 2} \Xi^{(k)}$. Using this decomposition and the RIPL assumption we have

$$\begin{aligned}
\|AP_{\Xi^c}x + e\|_{\ell^2} &\leq \sum_{k\geq 2} \sqrt{1+\delta_{\boldsymbol{s},\boldsymbol{M}}}\|P_{\Xi^{(k)}}x\|_{\ell^2} + \|e\|_{\ell^2} \leq \sqrt{2}\sum_{k\geq 1}\sqrt{\sum_{i=1}^{r}\frac{1}{s_i}\|P_{\Xi_{i,k}}x\|_{\ell^1}^2} + \|e\|_{\ell^2} \\
&\leq \sqrt{2}\frac{1}{\sqrt{\xi}}\sum_{k\geq 1}\sqrt{\sum_{i=1}^{r}\|P_{\Xi_{i,k}}x\|_{\ell^1_w}^2} + \|e\|_{\ell^2} \leq \frac{\sqrt{2}}{\sqrt{\xi}}\|P_{\Delta^c}x\|_{\ell^1_w} + \|e\|_{\ell^2} \\
&= \frac{\sqrt{2}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \|e\|_{\ell^2}.
\end{aligned}$$

Combining this result with (4.4), we have

$$\begin{aligned}
\|x - x'\|_{\ell^1_w} &\leq \left[1 + \frac{\sqrt{(4+2\kappa)\zeta}\tau}{\sqrt{\xi}}\right]\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \sqrt{(2+\kappa)\zeta}\tau\|e\|_{\ell^2} + \sqrt{(2+\kappa)\zeta}\lambda \\
&= C_{\kappa,\tau}\frac{\sqrt{\zeta}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D_{\kappa,\tau}\sqrt{\zeta}\|e\|_{\ell^2} + E_\kappa\sqrt{\zeta}\lambda,
\end{aligned}$$

as was to be shown.

For the 2-norm case we again focus on particular level i. Using the definition of Ξ_i and Stechkin's inequality (see, e.g., [27, Proposition 6.23]), we see that

$$\begin{aligned}
\|P_{M_i}^{M_{i-1}}(x - x')\|_{\ell^2}^2 &= \|P_{\Xi_i^c}x\|_{\ell^2}^2 + \|P_{\Xi_i}x - P_{M_i}^{M_{i-1}}x'\|_{\ell^2}^2 \\
&\leq \frac{1}{(w^{(i)})^2 s_i}\|P_{\Xi_i^c}x\|_{\ell^1_w}^2 + \|P_{\Xi_i}x - P_{M_i}^{M_{i-1}}x'\|_{\ell^2}^2.
\end{aligned}$$

Summing over all levels $i = 1, \dots, r$ we have

$$\|x - x'\|_{\ell^2}^2 \leq \frac{1}{\xi}\|P_{\Xi^c}x\|_{\ell^1_w}^2 + \|P_\Xi x - x'\|_{\ell^2}^2 \leq \frac{1}{\xi}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w}^2 + (\tau\|AP_{\Xi^c}x + e\|_{\ell^2} + \lambda)^2,$$

where we have applied the definitions of ξ, Ξ, and the assumptions of the theorem. As a result we also have

$$\|x - x'\|_{\ell^2} \leq \frac{1}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau\|AP_{\Xi^c}x + e\|_{\ell^2} + \lambda.$$

As we have already bounded this second term, we have

$$\|x-x'\|_{\ell^2} \leq \frac{1+\sqrt{2}\tau}{\sqrt{\xi}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau\|e\|_{\ell^2} + \lambda = \frac{F_\tau}{\sqrt{\xi}} \sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau\|e\|_{\ell^2} + \lambda,$$

thus completing the proof. □

4.2.3 Proof of Theorem 12

The following theorem is based on [27, Theorem 6.18].

Theorem 15. *Suppose $x \in \mathbb{C}^N$ is $(\boldsymbol{s},\boldsymbol{M})$-sparse in levels, with the RIPL constant satisfying*

$$\delta_{3\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{3}}.$$

Then, for all $x \in \mathbb{C}^N$, $e \in \mathbb{C}^m$ and $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$, the sequence $(x^{(n)})_{n\geq 0}$ defined by $\mathrm{IHTL}(A, y, \boldsymbol{s}, \boldsymbol{M})$ *for $y = Ax + e$ satisfies*

$$\|x^{(n)} - P_\Delta x\|_{\ell^2} \leq \rho^n \|x^{(0)} - P_\Delta x\|_{\ell^2} + \tau\|AP_{\Delta^c}x + e\|_{\ell^2}.$$

where $\rho = \sqrt{3}\delta_{3\boldsymbol{s},\boldsymbol{M}} < 1$ and $\tau > 0$ only depends on ρ and $\delta_{3\boldsymbol{s},\boldsymbol{M}}$, with $\tau \leq 2.18/(1-\rho)$.

Proof. We firstly define $\Delta_i = \Delta \cap \{M_{i-1}+1,\dots,M_i\}$ and $\Delta_i^c = \{M_{i-1}+1,\dots,M_i\}\backslash\Delta_i$. Analogously to the proof of Theorem 14, it will prove to be convenient to decompose

$$\{1,\dots,N\} = \Delta \cup \Delta^c = \bigcup_{i=1}^{r} \Delta_i \cup \Delta_i^c.$$

Similarly we define Δ_i^{n+1} as the index set of the largest s_i entries of $x^{(n+1)}$ in the band $\{M_{i-1}+1,\dots M_i\}$. With this decomposition, we may use techniques near-identical to those in [27, Theorem 6.18], and thus we give a brief treatment where possible. By definition, for any Δ_i,

$$\|P_{\Delta_i}(x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2} \leq \|P_{\Delta_i^{n+1}}(x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2}.$$

Then we cancel any shared contribution on the set $\Delta_i \cap \Delta_i^{n+1}$,

$$\|P_{\Delta_i\backslash\Delta_i^{n+1}}(x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2} \leq \|P_{\Delta_i^{n+1}\backslash\Delta_i}(x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2}. \tag{4.5}$$

Here, making the observation that $P_{\Delta_i}x = 0$ on $\Delta_i^{n+1}\backslash\Delta_i$ and $P_{M_i}^{M_{i-1}}x^{(n+1)} = 0$ on $\Delta_i\backslash\Delta_i^{(n+1)}$, we write the right-hand side of Eq. (4.5) as

$$\|P_{\Delta_i^{n+1}\backslash\Delta_i}(x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2} = \|P_{\Delta_i^{n+1}\backslash\Delta_i}(x^{(n)} - P_{\Delta_i}x + A^*(y - Ax^{(n)}))\|_{\ell^2}.$$

and we bound the left-hand side of Eq. (4.5) from below as

$$\|P_{\Delta_i\setminus\Delta_i^{n+1}}(x^{(n)}+A^*(y-Ax^{(n)}))\|_{\ell^2}\geq\|P_{\Delta_i\setminus\Delta_i^{n+1}}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2} - \|P_{\Delta_i\setminus\Delta_i^{n+1}}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}.$$

Combining these both into Eq. (4.5) we find that

$$\|P_{\Delta_i\setminus\Delta_i^{n+1}}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2}\leq\sqrt{2}\|P_{\Delta_i^{n+1}\ominus\Delta_i}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}, \tag{4.6}$$

where $\Delta_i^{n+1}\ominus\Delta_i=(\Delta_i^{n+1}\setminus\Delta_i)\cup(\Delta_i\setminus\Delta_i^{n+1})$ is the symmetric difference. We now seek to bound the left-hand side further from below. To do so, we decompose

$$\begin{aligned}\|P_{M_i}^{M_{i-1}}&x^{(n+1)}-P_{\Delta_i}x\|_{\ell^2}^2\\ &=\|P_{\Delta_i^{n+1}}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2}^2+\|P_{(\Delta_i^{n+1})^c}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2}^2\\ &=\|P_{\Delta_i^{n+1}}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2+\|P_{(\Delta_i^{n+1})^c}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2}^2.\end{aligned}$$

Further observing that $P_{M_i}^{M_{i-1}}x^{(n+1)}=0$ on $(\Delta_i^{n+1})^c$, and $P_{\Delta_i}x=0$ on Δ_i^c, we can write

$$\|P_{M_i}^{M_{i-1}}x^{(n+1)}-P_{\Delta_i}x\|_{\ell^2}^2=\|P_{\Delta_i^{n+1}}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2 +\|P_{\Delta_i\setminus\Delta_i^{n+1}}(P_{\Delta_i}x-P_{M_i}^{M_{i-1}}x^{(n+1)})\|_{\ell^2}^2.$$

Combining this argument with the previous bound Eq. (4.6) we have in summary

$$\begin{aligned}\|P_{M_i}^{M_{i-1}}x^{(n+1)}-P_{\Delta_i}x\|_{\ell^2}^2&\leq\|P_{\Delta_i^{n+1}}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2\\ &\quad+2\|P_{\Delta_i^{n+1}\ominus\Delta_i}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2\\ &\leq 3\|P_{\Delta_i^{n+1}\cup\Delta_i}(x^{(n)}-P_{\Delta_i}x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2.\end{aligned}$$

Summing this over all levels $i=1,\dots,r$, we have that

$$\|x^{(n+1)}-P_\Delta x\|_{\ell^2}^2\leq 3\|P_{\Delta^{n+1}\cup\Delta}(x^{(n)}-P_\Delta x+A^*(y-Ax^{(n)}))\|_{\ell^2}^2.$$

By redefining $y=Ax+e=AP_\Delta x+e'$, with $e'=e+AP_{\Delta^c}x$, we may further bound this from above as

$$\begin{aligned}\|x^{(n+1)}-P_\Delta x\|_{\ell^2}&\leq\sqrt{3}\left[\|P_{\Delta^{n+1}\cup\Delta}(x^{(n)}-P_\Delta x+A^*A(P_\Delta x-x^{(n)})\|_{\ell^2}+\|P_{\Delta^{n+1}\cup\Delta}A^*e'\|_{\ell^2}\right]\\ &\leq\sqrt{3}\left[\|P_{\Delta^{n+1}\cup\Delta}(I-A^*A)(x^{(n)}-P_\Delta x)\|_{\ell^2}+\|P_{\Delta^{n+1}\cup\Delta}A^*e'\|_{\ell^2}\right].\end{aligned}$$

Here we note that $\operatorname{supp}(x^{(n)} - P_\Delta(x)) \subset \Delta \cup \Delta^n$, and $(\Delta \cup \Delta^n) \cup (\Delta^{n+1} \cup \Delta) \in D_{3\boldsymbol{s},\boldsymbol{M}}$. These observations allow us to apply part (ii) of Lemma 4 on the first term, and Lemma 5 on the second term, giving

$$\|x^{(n+1)} - P_\Delta x\|_{\ell^2} \leq \sqrt{3}\left[\delta_{3\boldsymbol{s},\boldsymbol{M}}\|x^{(n)} - P_\Delta x\|_{\ell^2} + \sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}\|e'\|_{\ell^2}\right].$$

Finally by examining this inequality, we set

$$\rho = \sqrt{3}\delta_{3\boldsymbol{s},\boldsymbol{M}}, \qquad (1-\rho)\tau = \sqrt{3}\sqrt{1+\delta_{2\boldsymbol{s},\boldsymbol{M}}}.$$

Recalling that $e' = AP_{\Delta^c}x + e$, we have

$$\|x^{(n+1)} - P_\Delta x\|_{\ell^2} \leq \rho\|x^{(n)} - P_\Delta x\|_{\ell^2} + (1-\rho)\tau\|AP_{\Delta^c}x + e\|_{\ell^2},$$

which, by induction on n, gives

$$\|x^{(n)} - P_\Delta x\|_{\ell^2} \leq \rho^n\|x^{(0)} - P_\Delta x\|_{\ell^2} + \tau\|AP_{\Delta^c}x + e\|_{\ell^2}.$$

This was precisely the result to be shown, noting that

$$\rho < 1 \Leftrightarrow \delta_{3\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{3}} \qquad \text{and so} \qquad \tau = \frac{\sqrt{3}\sqrt{1+\delta_{3\boldsymbol{s},\boldsymbol{M}}}}{1-\rho} < \frac{\sqrt{3+\sqrt{3}}}{1-\rho} < \frac{2.18}{1-\rho}.$$

This concludes the proof. □

Proof. (Theorem 12) Using Theorem 15 with $(2\boldsymbol{s},\boldsymbol{M})$ instead of $(\boldsymbol{s},\boldsymbol{M})$, there exist constants $\rho \in (0,1)$, $\tau > 0$ depending on $\delta_{6\boldsymbol{s},\boldsymbol{M}}$ such that

$$\|x^{(n)} - P_\Xi x\|_{\ell^2} \leq \rho^n\|x^{(0)} - P_\Xi x\|_{\ell^2} + \tau\|AP_{\Xi^c}x + e\|_{\ell^2},$$

where $\Xi = \Xi^{(0)} \cup \ldots \cup \Xi_n$ and Ξ_i is the index set of the largest $2s_i$ entries of $P_{M_i}^{M_{i-1}}x$ (note that we applied Theorem 15 with $(2\boldsymbol{s},\boldsymbol{M})$ since $\Xi \in D_{2\boldsymbol{s},\boldsymbol{M}}$). Then, by letting $x' = x^{(n)}$ and $\lambda = \rho^n\|P_\Xi x\|_{\ell^2}$ (recall that $x^{(0)} = 0$) we may apply Theorem 14 with $\kappa = 2$ to assert

$$\|x - x^{(n)}\|_{\ell^1_w} \leq C\frac{\sqrt{\zeta}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D\sqrt{\zeta}\|e\|_{\ell^2} + 2\sqrt{\zeta}\rho^n\|x\|_{\ell^2},$$
$$\|x - x^{(n)}\|_{\ell^2} \leq \frac{E}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau\|e\|_{\ell^2} + \rho^n\|x\|_{\ell^2}$$

where $C, D, E > 0$ depend on τ, ρ, κ and thus only on $\delta_{6\boldsymbol{s}}, \boldsymbol{M}$. □

4.2.4 Proof of Theorem 13

The following is based on [27, Theorem 6.27].

Theorem 16. *Suppose the* $(4\boldsymbol{s}, \boldsymbol{M})$*-th RICL constant of the matrix* $A \in \mathbb{C}^{m\times N}$ *satisfies*

$$\delta_{4\boldsymbol{s},\boldsymbol{M}} < \frac{\sqrt{\sqrt{\frac{11}{3}} - 1}}{2}$$

Then for $x \in \mathbb{C}^N$*,* $e \in \mathbb{C}^m$ *and index set* $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$*, the sequence* $(x^{(n)})_{n\geq 0}$ *defined by* $\mathrm{CoSaMPL}(A, y, \boldsymbol{s}, \boldsymbol{M})$ *with* $y = Ax + e$ *satisfies*

$$\|x^{(n)} - P_\Delta x\|_{\ell^2} \leq \rho^n \|x^{(0)} - P_\Delta x\|_{\ell^2} + \tau \|AP_{\Delta^c} x + e\|_{\ell^2},$$

where $\rho \in (0, 1)$ *and* $\tau > 0$ *are constants only depending on* $\delta_{4\boldsymbol{s},\boldsymbol{M}}$*.*

Proof. As before, with correct treatment of our index sets, many of the algebraic manipulations follow near-identically from [27, Theorem 6.27]. Then,

$$\|P_{U^{(n+1)}}(P_\Delta x - x^{(n+1)})\|_{\ell^2} \leq \|u^{(n+1)} - x^{(n+1)}\|_{\ell^2} + \|u^{(n+1)} - P_{U^{(n+1)}\cap\Delta} x\|_{\ell^2}. \tag{4.7}$$

Further as $x^{(n+1)} = H_{\boldsymbol{s},\boldsymbol{M}}(u^{(n+1)})$ we bound $\|u^{(n+1)} - x^{(n+1)}\|_{\ell^2} \leq \|u^{(n+1)} - P_{U^{(n+1)}\cap\Delta} x\|_{\ell^2}$. This result, combined with the fact that $P_{(U^{(n+1)})^c} x^{(n+1)} = P_{(U^{(n+1)})^c} u^{(n+1)} = 0$, asserts

$$\begin{aligned} \|P_\Delta x - x^{(n+1)}\|^2_{\ell^2} &= \|P_{(U^{(n+1)})^c}(P_\Delta x - x^{(n+1)})\|^2_{\ell^2} + \|P_{U^{(n+1)}}(P_\Delta x - x^{(n+1)})\|^2_{\ell^2} \\ &\leq \|P_{(U^{(n+1)})^c}(P_\Delta x - u^{(n+1)})\|^2_{\ell^2} + 4\|P_{U^{(n+1)}}(P_\Delta x - u^{(n+1)})\|^2_{\ell^2}. \end{aligned} \tag{4.8}$$

We will us this bound later, but we now examine the latter term more closely.

We first make the observation that $P_{U^{(n+1)}} A^*(y - Au^{(n+1)}) = 0$, as $u^{(n+1)}$ satisfies the normal equations when restricted to its support. Thus we may write $P_{U^{(n+1)}} A^* A(P_\Delta x - u^{(n+1)}) = -P_{U^{(n+1)}} A^* e'$, where $e' = AP_{\Delta^c} x + e$. We use this to write

$$\|P_{U^{(n+1)}}(P_\Delta x - u^{(n+1)})\|_{\ell^2} \leq \|(I - A^*A)(P_\Delta x - u^{(n+1)})\|_{\ell^2} + \|P_{U^{(n+1)}} A^* e'\|_{\ell^2}.$$

Now as $\Delta \in D_{\boldsymbol{s},\boldsymbol{M}}$ and $U^{(n+1)} \in D_{3\boldsymbol{s},\boldsymbol{M}}$, we have their union is in $D_{4\boldsymbol{s},\boldsymbol{M}}$. Thus using Lemma 4 (ii) gives $\|(I - A^*A)(P_\Delta x - u^{(n+1)})\|_{\ell^2} \leq \delta_{4\boldsymbol{s},\boldsymbol{m}} \|P_\Delta x - u^{(n+1)}\|_{\ell^2}$, and so

$$\|P_{U^{(n+1)}}(P_\Delta x - u^{(n+1)})\|_{\ell^2} \leq \delta_{4\boldsymbol{s},\boldsymbol{m}} \|P_\Delta x - u^{(n+1)}\|_{\ell^2} + \|P_{U^{(n+1)}} A^* e'\|_{\ell^2}. \tag{4.9}$$

From here denoting $\delta_{4\boldsymbol{s},\boldsymbol{M}} = \delta$, we wish to derive the inequality

$$\|P_{U^{(n+1)}}(P_\Delta x - u^{(n+1)})\|_{\ell^2} \leq \frac{\delta}{\sqrt{1-\delta^2}} \|P_{(U^{(n+1)})^c}(P_\Delta x - u^{(n+1)})\|_{\ell^2} + \frac{1}{1-\delta} \|P_{U^{(n+1)}} A^* e'\|_{\ell^2}. \tag{4.10}$$

We here split into cases. If $\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 \leq \frac{1}{1-\delta}\|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2$, then the desired inequality is immediate. As

$$\begin{aligned}\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 &\leq \frac{1}{1-\delta}\|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2 \\ &\leq \frac{\delta}{\sqrt{1-\delta}}\|P^{\perp}_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 + \frac{1}{1-\delta}\|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2.\end{aligned}$$

Otherwise, if $\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 > \frac{1}{1-\delta}\|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2$, we rearrange (4.9) to write

$$\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 - \|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2 \leq \delta\|P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)}\|_2.$$

Squaring both sides, which in this case are both nonnegative, and decomposing the right hand side gives

$$\begin{aligned}(\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 - \|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2)^2 &\leq \delta^2\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2^2 \\ &\quad + \delta^2\|P^{\perp}_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2^2.\end{aligned}$$

For temporary convenience, we rewrite the above as

$$(a-c)^2 \leq \delta^2 a^2 + \delta^2 b^2,$$

where $a = \|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2$, $b = \|P^{\perp}_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2$ and $c = \|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2$. Using this we rearrange to find

$$(a-c)^2 - \delta^2 a^2 \leq \delta^2 b^2$$

or after some simplification,

$$(1-\delta^2)\left(a - \frac{c}{(1+\delta)}\right)\left(a - \frac{c}{(1-\delta)}\right) \leq \delta^2 b^2.$$

Here we use our assumption that $\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 > \frac{1}{1-\delta}\|P_{U^{n+1}}(A^*\boldsymbol{e}')\|_2$, or in this new notation $a - \frac{c}{(1-\delta)} > 0$. Thus we may bound

$$a - \frac{c}{(1+\delta)} \geq a - \frac{c}{(1-\delta)} > 0,$$

which allows us to assert that

$$\left(a - \frac{c}{(1-\delta)}\right)^2 \leq \frac{\delta^2}{(1-\delta^2)} b^2.$$

Now taking a square-root and rearranging gives

$$a \leq \frac{\delta}{\sqrt{1-\delta^2}} b + \frac{c}{1-\delta},$$

and by substituting back in the values of a, b, c we find the desired inequality:

$$\|P_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{(n+1)})\|_2 \leq \frac{\delta}{\sqrt{1-\delta^2}} \|P^{\perp}_{U^{n+1}}(P_\Delta \boldsymbol{x} - \boldsymbol{u}^{n+1})\|_2 + \frac{1}{1-\delta}\|P_{U^{n+1}}(A^* \boldsymbol{e}')\|_2. \quad (4.11)$$

Recalling that CoSaMPL defines $S^{(n)} = \mathrm{supp}(x^{(n)})$, and that $S^{(n)} \subset U^{(n+1)}$. Further we define $T^{(n+1)} = L_{2\boldsymbol{s},\boldsymbol{M}}(A^*(y - Ax^{(n)}))$. As this is the index set of the largest $(2\boldsymbol{s}, \boldsymbol{M})$ entries of $A^*(y - Ax^{(n)})$, and $\Delta \cup S^{(n)} \in D_{2\boldsymbol{s},\boldsymbol{M}}$ we have

$$\|P_{\Delta \cup S^{(n)}} A^*(y - Ax^{(n)})\|_{\ell^2} \leq \|P_{T^{(n+1)}} A^*(y - Ax^{(n)})\|_{\ell^2}.$$

In turn, eliminating the shared contribution on $(\Delta \cup S^{(n)}) \cap T^{(n+1)}$ we find

$$\|P_{(\Delta \cup S^{(n)}) \setminus T^{(n+1)}} A^*(y - Ax^{(n)})\|_{\ell^2} \leq \|P_{T^{(n+1)} \setminus (\Delta \cup S^{(n)})} A^*(y - Ax^{(n)})\|_{\ell^2}.$$

Now as $P_\Delta x - x^{(n)} = 0$ on $T^{(n+1)} \setminus (\Delta \cup S^{(n)})$ we may write the right-hand side of the above as

$$\|P_{T^{(n+1)} \setminus (\Delta \cup S^{(n)})} A^*(y - Ax^{(n)})\|_{\ell^2} = \|P_{T^{(n+1)} \setminus (\Delta \cup S^{(n)})}(P_\Delta x - x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2},$$

whereas for the left-hand side we apply a reverse triangle inequality

$$\begin{aligned}
&\|P_{(\Delta \cup S^{(n)}) \setminus T^{(n+1)}} A^*(y - Ax^{(n)})\|_{\ell^2} \\
&\quad \geq \|P_{(\Delta \cup S^{(n)}) \setminus T^{(n+1)}}(P_\Delta x - x^{(n)})\|_{\ell^2} - \|P_{(\Delta \cup S^{(n)}) \setminus T^{(n+1)}}(P_\Delta x - x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2} \\
&\quad = \|P_{(T^{(n+1)})^c}(P_\Delta x - x^{(n)})\|_{\ell^2} - \|P_{(\Delta \cup S^{(n)}) \setminus T^{(n+1)}}(P_\Delta x - x^{(n)} + A^*(y - Ax^{(n)}))\|_{\ell^2}.
\end{aligned}$$

Combining these two observations and rearranging gives

$$\begin{aligned}
&\|P_{(T^{(n+1)})^c}(P_\Delta x - x^{(n)})\|_{\ell^2} \\
&\qquad \leq \sqrt{2}\|P_{T^{(n+1)} \ominus (\Delta \cup S^{(n)})}(P_\Delta x - x^{(n)} + A^*(y - Ax^{(n)})\|_{\ell^2} \\
&\qquad \leq \sqrt{2}\|P_{T^{(n+1)} \ominus (\Delta \cup S^{(n)})}(I - A^*A)(x^{(n)} - P_\Delta x)\|_{\ell^2} + \sqrt{2}\|P_{T^{(n+1)} \ominus (\Delta \cup S^{(n)})} A^* e'\|_{\ell^2},
\end{aligned}$$

where $\ominus$ denotes the symmetric difference, and $y = AP_\Delta x + e'$ is as before. Now, as $T^{(n+1)} \subset U^{(n+1)}$ and $S^{(n)} \subset U^{(n+1)}$ by the definition of CoSaMPL, we may bound the left-hand side of the above equation from below by

$$\begin{aligned}\|P_{(T^{(n+1)})^c}(P_\Delta x - x^{(n)})\|_{\ell^2} &\geq \|P_{(U^{(n+1)})^c}(P_\Delta x - x^{(n)})\|_{\ell^2} = \|P_{(U^{(n+1)})^c}P_\Delta x\|_{\ell^2} \\ &= \|P_{(U^{(n+1)})^c}(P_\Delta x - u^{(n+1)})\|_{\ell^2}.\end{aligned}$$

With this lower bound in hand, we note that $\Delta, S^{(n)} \in D_{\boldsymbol{s},\boldsymbol{M}}$ and $T^{(n+1)} \in D_{2\boldsymbol{s},\boldsymbol{M}}$ so that we may apply Lemma 5 (ii) with $T^{(n+1)} \ominus (\Delta \cup S^{(n)}) \subset T^{(n+1)} \cup (\Delta \cup S^{(n)}) \in D_{4\boldsymbol{s},\boldsymbol{M}}$ on the term $\|P_{T^{(n+1)}\ominus(\Delta\cup S^{(n)})}(I - A^*A)(x^n - P_\Delta x)\|_{\ell^2}$. Combining this series of observations gives

$$\|P_{(U^{(n+1)})^c}(P_\Delta x - u^{(n+1)})\|_{\ell^2} \leq \sqrt{2}\delta_{4\boldsymbol{s},\boldsymbol{M}}\|x^n - P_\Delta x\|_{\ell^2} + \sqrt{2}\|P_{T^{(n+1)}\ominus(\Delta\cup S^{(n)})}(A^*e')\|_{\ell^2}. \tag{4.12}$$

To conclude our argument, it remains to combine the three distinct results of equations (4.8), (4.11) and (4.12). Again, this is near identical to the sparse case in [27, Theorem 6.27], and contains purely algebraic manipulations. This leads to the inequality

$$\begin{aligned}\|P_\Delta x - x^{(n+1)}\|_{\ell^2} \leq &\sqrt{\frac{2\delta^2(1+3\delta^2)}{1-\delta^2}}\|x^{(n)} - P_\Delta x\|_{\ell^2} \\ &+ \sqrt{\frac{2(1+3\delta^2)}{1-\delta^2}}\|P_{T^{(n+1)}\ominus(\Delta\cup S^{(n)})}A^*e'\|_{\ell^2} + \frac{2}{1-\delta}\|P_{U^{(n+1)}}A^*e'\|_{\ell^2}.\end{aligned}$$

Now using Lemma 5 on the sets $T^{(n+1)} \ominus (\Delta \cup S^{(n)}) \in D_{4\boldsymbol{s},\boldsymbol{M}}$ and $U^{(n+1)} \in D_{3\boldsymbol{s},\boldsymbol{M}} \subset D_{4\boldsymbol{s},\boldsymbol{M}}$ we find

$$\|P_\Delta x - x^{(n+1)}\|_{\ell^2} \leq \sqrt{\frac{2\delta^2(1+3\delta^2)}{1-\delta^2}}\|x^{(n)} - P_\Delta x\|_{\ell^2} + \left(\sqrt{\frac{2(1+\delta)(1+3\delta^2)}{1-\delta^2}} + \frac{2\sqrt{1+\delta}}{1-\delta}\right)\|e'\|_{\ell^2},$$

which is exactly

$$\|P_\Delta x - x^{(n+1)}\|_{\ell^2} \leq \rho\|x^{(n)} - P_\Delta x\|_{\ell^2} + \tau\|AP_{\Delta^c}x + e\|_{\ell^2},$$

for suitable $\rho, \tau > 0$ depending only on δ. Then by a simple induction we have

$$\|x^{(n)} - P_\Delta x\|_{\ell^2} \leq \rho^n\|x^{(0)} - P_\Delta x\|_{\ell^2} + \tau\|AP_{\Delta^c}x + e\|_{\ell^2}.$$

Which is precisely the result to be shown. Notably, the constant $\rho < 1$ only if

$$\sqrt{\frac{2\delta^2(1+3\delta^2)}{1-\delta^2}} < 1 \Leftrightarrow 6\delta^4 + 3\delta^2 - 1 < 0,$$

which by solving this quadratic in δ^2 for its largest root gives us that we require $\delta^2 < \frac{\sqrt{\frac{11}{3}}-1}{4}$ as was assumed. □

Proof. (Theorem 13) Under the hypotheses of the theorem, let us denote $\Xi = L_{2\boldsymbol{s},\boldsymbol{M}}(x)$ to be the index set corresponding to the largest $(2\boldsymbol{s},\boldsymbol{M})$ entries of x. First, we may apply Theorem 16 to assert there exist $\rho \in (0,1)$ and $\tau > 0$ depending only on $\delta_{8\boldsymbol{s},\boldsymbol{M}}$ such that, for any $n \geq 0$,

$$\|x^{(n)} - P_{\Xi}x\|_{\ell^2} \leq \rho^n \|P_{\Xi}x\|_{\ell^2} + \tau\|AP_{\Xi^c}x + e\|_{\ell^2}.$$

Then, we may apply Theorem 14 with $x' = x^{(n)}$ and $\lambda = \rho^n\|P_{\Xi}x\|_{\ell^2} \leq \rho^n\|x\|_{\ell^2}$ to give us that

$$\begin{aligned}
\|x - x^{(n)}\|_{\ell^1_w} &\leq C\frac{\sqrt{\zeta}}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + D\sqrt{\zeta}\|e\|_{\ell^2} + 2\sqrt{\zeta}\rho^n\|x\|_{\ell^2}, \\
\|x - x^{(n)}\|_{\ell^2} &\leq \frac{E}{\sqrt{\xi}}\sigma_{\boldsymbol{s},\boldsymbol{M}}(x)_{\ell^1_w} + \tau\|e\|_{\ell^2} + \rho^n\|x\|_{\ell^2}
\end{aligned}$$

where $C, D, E > 0$ depend only on $\tau, \rho, \delta_{8\boldsymbol{s},\boldsymbol{M}}$ and thus only on $\delta_{8\boldsymbol{s},\boldsymbol{M}}$. This is exactly the result that was to be shown. □

Chapter 5

Numerical examination of the methods

With the theory of IHTL and CoSaMPL developed, it is important to also examine the practical performance of these iterative approaches. There are several types of comparisons that are interesting to analyse. Foremost of interest is the comparison of the levels-based algorithms against the sparse counterparts. Then, of course, is the comparison of the levels methods against each other. Also of interest is the performance of these iterative methods against that of wQCBP – or in the sparse case, QCBP.

5.1 Experimental setup

A standard methodology is to attempt to recover randomly generated vectors over many trials, with a varying number of samples m. Over these trials and m values, we will plot the empirical recovery probability. The goal of such experiments is to see numerically which algorithms require the fewest number of samples. This, as stated in previous chapters, was one of the guiding goals of this work. While we have proved under the RIP that recovery is guaranteed, it is not a necessary condition and we may succeed in recovery even for somewhat smaller m. Thus, we are interested in three regions of m values: the region of zero recovery probability, the region of probability one, and the thin region between them, termed the *phase transition*.

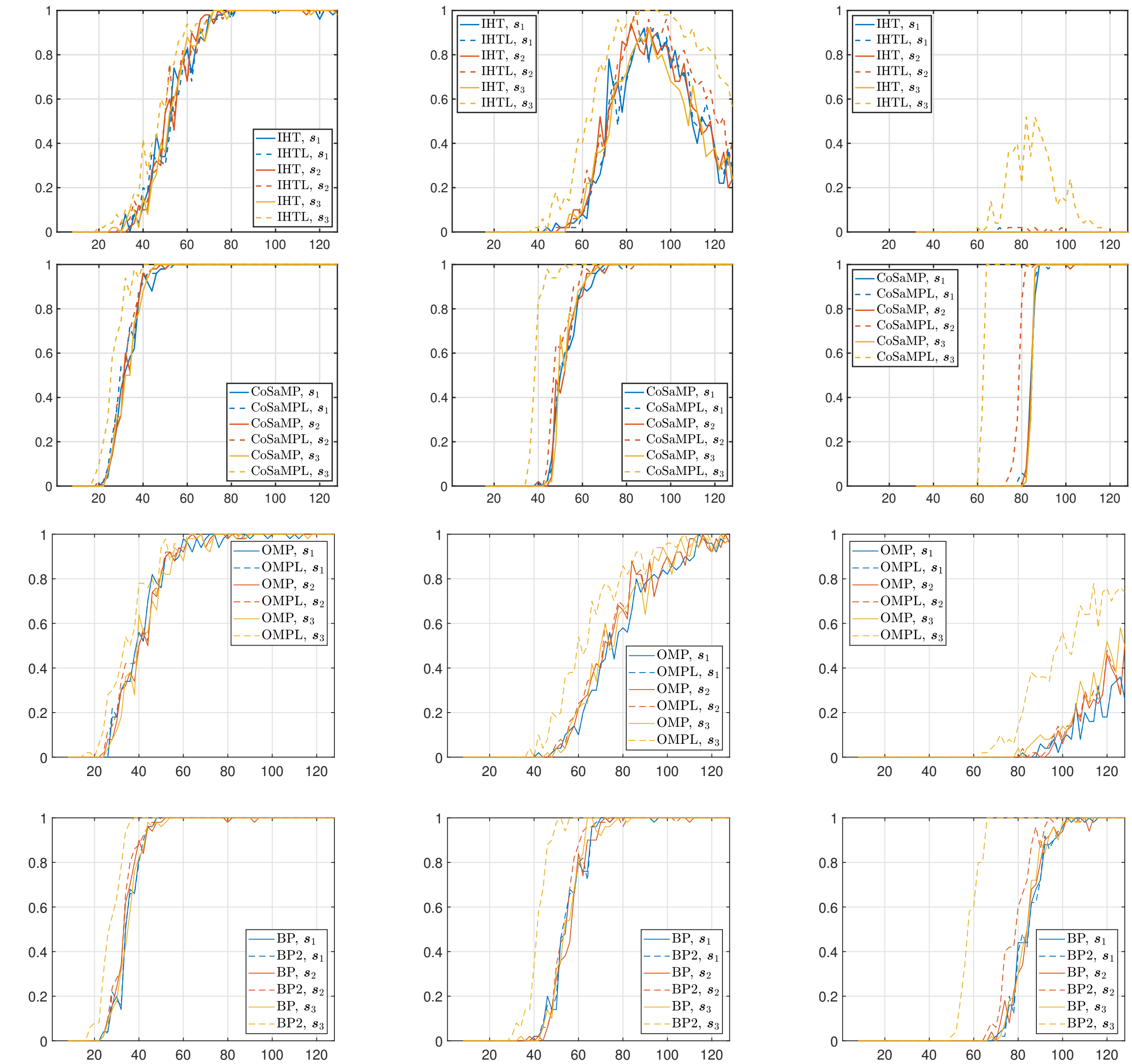
To be precise, all numerical experiments share the following setup. For each fixed total sparsity s and number of measurements m, we generate an $(\boldsymbol{s}, \boldsymbol{M})$-sparse in levels vector x of length $N = 128$ with random support and unit normal random entries. The local sparsity pattern $\boldsymbol{s}$ depends on the experiment, as outlined below. Using a measurement matrix A that is a Gaussian random matrix (independent, normally distributed entries with mean zero and variance $1/\sqrt{m}$), linear samples $y = Ax$ are generated. Then we compute an approximation $\hat{x}$ to the vector x and record the relative error $\|x - \hat{x}\|_{\ell^2}/\|x\|_{\ell^2}$. Over 50 trials, we compute the success probability with the success criterion that the relative error be less than 10^{-2}. For IHTL and CoSaMPL, we have the additional stopping criterion that the algorithms terminate either when $\|x^{(n+1)} - x^{(n)}\|_{\ell^2}$ is less than a tolerance 10^{-4}, or if the algorithms exceeds 1000 iterations. For OMPL, we simply run s iterations.

We also add a comparison to weighted quadratically constrained basis pursuit. We do this with and without weights, to show the levels structure also boosts performance in the optimization setting. For practical purposes, we use weights $w_i = \sqrt{s/(s_i + 10^{-5})}$. This regularization is to avoid issues when $s_i = 0$, which is an interesting case to consider – and has the intended behavior of having large weights where the local sparsity is small. This convex optimization problem is solved using the SPGL1 package in Matlab [39], with default tolerances and default recommended maximum iterations of 10000. This should serve as a baseline comparison of accuracy, with the important note that this decoder is quite slow in comparison to the iterative methods.

5.2 The experiments

We firstly examine each method for various fixed total sparsities, Fig. 5.1. These experiments are designed to compare the levels based algorithms to the sparse versions. In the first numerical test, we recover a vector with underlying sparsity $\boldsymbol{s}_1 = (s/2, s/2)$, $\boldsymbol{s}_2 = (3s/4, s/4)$ and $\boldsymbol{s}_3 = (s, 0)$ for fixed levels $\boldsymbol{M} = (N/2, N)$. Here, intuition says that moving to the levels case should be extremely beneficial, especially for $\boldsymbol{s}_2$ – as we have most of our nonzero entries in the first level. That is, the local structure is very pronounced. However, for $\boldsymbol{s}_1$, we expect marginal or no benefits, as being $(s/2, s/2)$-sparse in levels $(1, N/2, N)$ is equivalent to simply being s-sparse. And, this behavior is exactly what we observe, in that the transition curve for $\boldsymbol{s}_1$ is the latest, that is, it takes the most measurements to successfully recovery $\boldsymbol{s}_1$ sparse vectors. For example With CoSaMPL and $s = 32$, this takes $m = 80$ measurements to find and empirical probability of $\mathbb{P} = 1$. Comparatively, recovery for s_3 – where we know that the first level contains all the nonzero entries – requires only 64 measurements to reach $\mathbb{P} = 1$, whereas $\boldsymbol{s}_2$ requires approximately $m = 75$. The takeaway of this experiment is thus that the levels based algorithms perform better in the presence of more distinct levels structure, when the nonzero entries are more concentrated in specific levels. Another important observation - and one that will reoccur throughout all the following experiments – will be that IHTL performs generally worse than CoSaMPL, and is much more likely to not converge. This is perhaps unsurprising considering the simplicity of IHT's iteration. For example, with $s = 32$, the empirical recovery probability for $\boldsymbol{s}_1, \boldsymbol{s}_2$ is always zero, and even for s_3-sparse vectors IHTL never exceeds $\mathbb{P} = 0.5$. In fact, in cases where IHT does not converge, iterates may 'blow up' in ℓ^2-norm. The iteration of CoSaMP avoids this behavior however, due to the least squares performed at each iteration.

Finally, we compare to basis pursuit. We overall have similar recovery behavior to CoSaMPL, with all three tests having the same appearance qualitatively. Quantitatively, this is most obvious for $\boldsymbol{s}_3$, where both CoSaMPL and wQCBP reach $\mathbb{P} = 1$ at approximately $m = 40, m = 50, m = 60$. To contrast these approaches however, the phase transition region for CoSaMPL is much sharper. For example in $s = 32$ and $\boldsymbol{s}_3$, CoSaMPL transitions between probability zero and one approximately on the interval $[60, 65]$ whereas wQCBP between $[40, 60]$. This means that BP has nonzero probability of succeeding at smaller m, at the tradeoff of reaching $\mathbb{P} = 1$ later.



$s = 8$ $s = 16$ $s = 32$

Figure 5.1: Horizontal phase transition line showing success probability versus m for various fixed total sparsities s. Two level sparsity with $\boldsymbol{M} = (N/2, N)$. The local sparsities are $\boldsymbol{s}_1 = (s/2, s/2)$, $\boldsymbol{s}_2 = (3s/4, s/4)$ and $\boldsymbol{s}_3 = (s, 0)$. The top row considers IHT and IHTL, the second row considers CoSaMP and CoSaMPL, the third OMPL, and the final BP.

With the same local sparsity patterns as above, we also test the speed of each method. The experimental setup is the same as the above, plotting median time of each test in seconds. Here, each row corresponds to a different total sparsity $s = 8, 16, 32$, and each column to a sparsity pattern. This allows the superposition of all the methods, to easily contrast between each approach on the same test set. There are several observations to be made. Foremost, running OMP and IHTL is negligible for every sparsity and every number of samples, with runtime $t < 0.01$s. This is particularly interesting for OMPL, which has reasonable recovery, as we will illustrate below. In comparison, CoSaMPL and BP have much longer runtime, on the order of median 0.07 seconds for CoSaMPL in the worst case $(\boldsymbol{s}_1, s = 32)$, and 0.05 for BP $(\boldsymbol{s}_3, s = 32)$. While this may sound extremely fast, this experiment is a "toy problem" in size, with $N = 128$ and 50 tests only. This increase, over many tests, with problem of reasonable size, can easily become difficult to run.

For more particular observations, it is interesting to point out that when the probability of recovery $\mathbb{P} = 1$ or $\mathbb{P} = 0$, CoSaMPL's runtime is also negilible, < 0.01s, whereas basis pursuit always needs some time greater than 0.01s, even with m on the order of N. This is discouraging as we will see below CoSaMPL and BP ave similar phase transitions. Counter to this is the regime where $s = 32$ for s_1 and s_2, where CoSaMPL has longer runtime than BP during the transition.

Overall, the takeaway is the tradeoff between recovery probability and runtime. IHTL takes this to the extreme end of runtime, and BP to the extreme end of recovery. Overall, OMPL seems to be the most "happy medium", balancing good phase transition behavior with fast runtime, and a guaranteed s iterations.

Next, we perform a similar experiment, now moving to the four levels case. The vector we recover is $(\boldsymbol{s}, \boldsymbol{M})$-sparse with $\boldsymbol{M} = (N/4, N/2, 3N/4, N)$ and either $\boldsymbol{s}_1 = (3s/8, s/8, 3s/8, s/8)$ or $\boldsymbol{s}_2 = (s/2, 0, s/2, 0)$. We then run IHTL, CoSaMPL, OMPL, and wQCBP with $1, 2$ or 4 levels each. Note that for the "2-levels" case, we have a working sparsity pattern of $(s/2, s/2)$ in levels $(N/2, N)$. Here, intuition – and the results – say that moving to more levels is using more accurate information of the local sparsity pattern, and thus should result in better recovery. This effect is extremely pronounced in the $\boldsymbol{s}_2$ case, where moving to 4 levels in the recovery algorithms result in a significant performance increase. Overall, the most gain is to be had in moving from 2 to 4 levels.

For both these experiments, it is useful to compare the observed results to the predictions of the theory. As our target vectors are exactly $(\boldsymbol{s}, \boldsymbol{M})$-sparse, and we are using constant weights for our iterative approaches – we expect exact recovery given sufficient iterations, provided we have the RIPL (with appropriate constant for each algorithm) of order $(2\boldsymbol{s}, \boldsymbol{M})$. For CoSaMPL in 4 levels, we observe this extremely clearly, with the transition to successful recover at measurements totalling $2s$. Of course, for A to have any chance of possessing the appropriate RIP, we must take at least $2s$ measurements, and we see for each algorithm that in the 4-levels case this is when recovery probability becomes nonzero. As before, it is

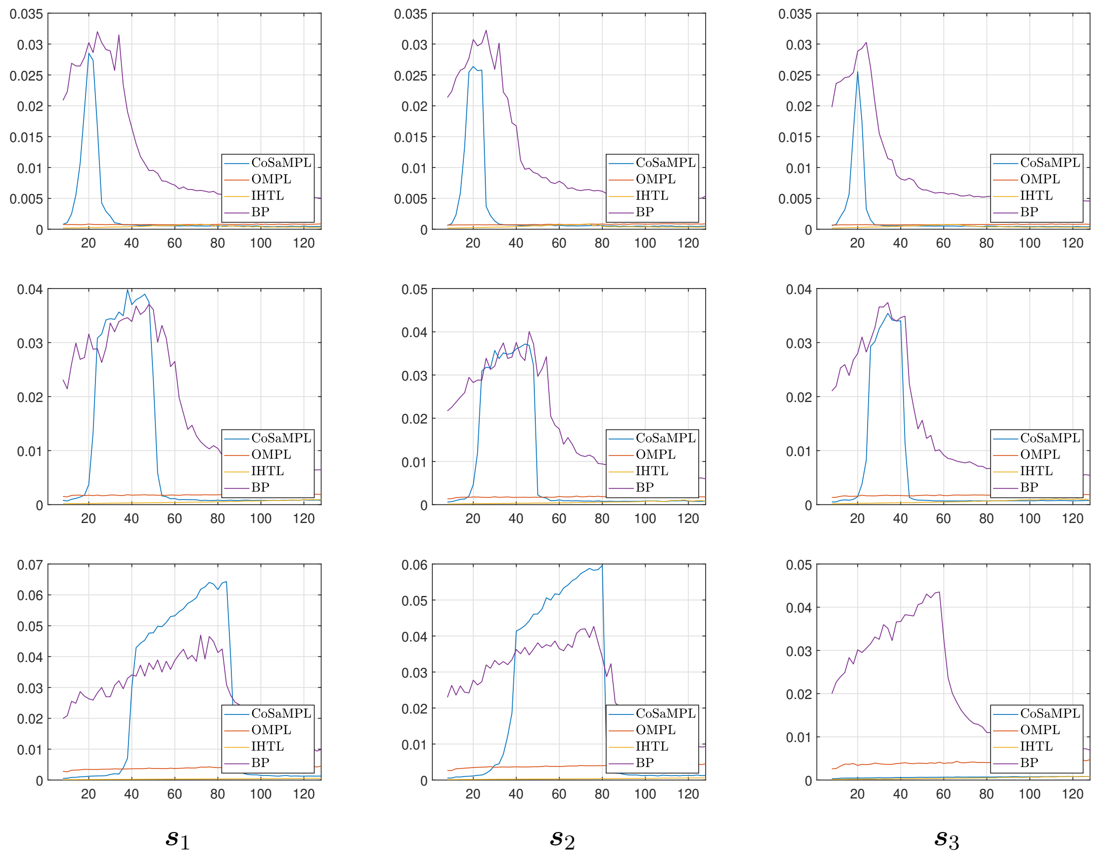


Figure 5.2: Time per iteration (s) against m for various fixed total sparsities s. Two level sparsity with $\boldsymbol{M} = (N/2, N)$. The local sparsities are $\boldsymbol{s}_1 = (s/2, s/2)$, $\boldsymbol{s}_2 = (3s/4, s/4)$ and $\boldsymbol{s}_3 = (s, 0)$, corresponding to the columns. Each row is a fixed total sparsity, with the top row $s = 8$, the next $s = 16$, and the last row $s = 32$.

interesting to observe that wQCBP performs reasonably similar to CoSaMPL, with a wider phase transition.

5.2.1 Phase transition plots

Next we proceed with *phase transition plots*. These are 2D arrays, in which each coordinate corresponds to a different value of $(s/N, m/N)$. For each value of s, m, we perform 50 tests of recovery, using the same procedure as before. Each pixel value records the empirical success probability, with yellow corresponding to 1, and blue 0. Correspondingly, more yellow regions are desirable, with a perfect oracle recovering up to the line $s = m$. These are very standard tools in the compressed sensing literature, as they reveal many of the benefits – or failings – of an approach in a visually obvious way.

We first perform an experiment similar to the above, comparing the performance of each algorithm against each other and their sparse counterparts. To observe the general

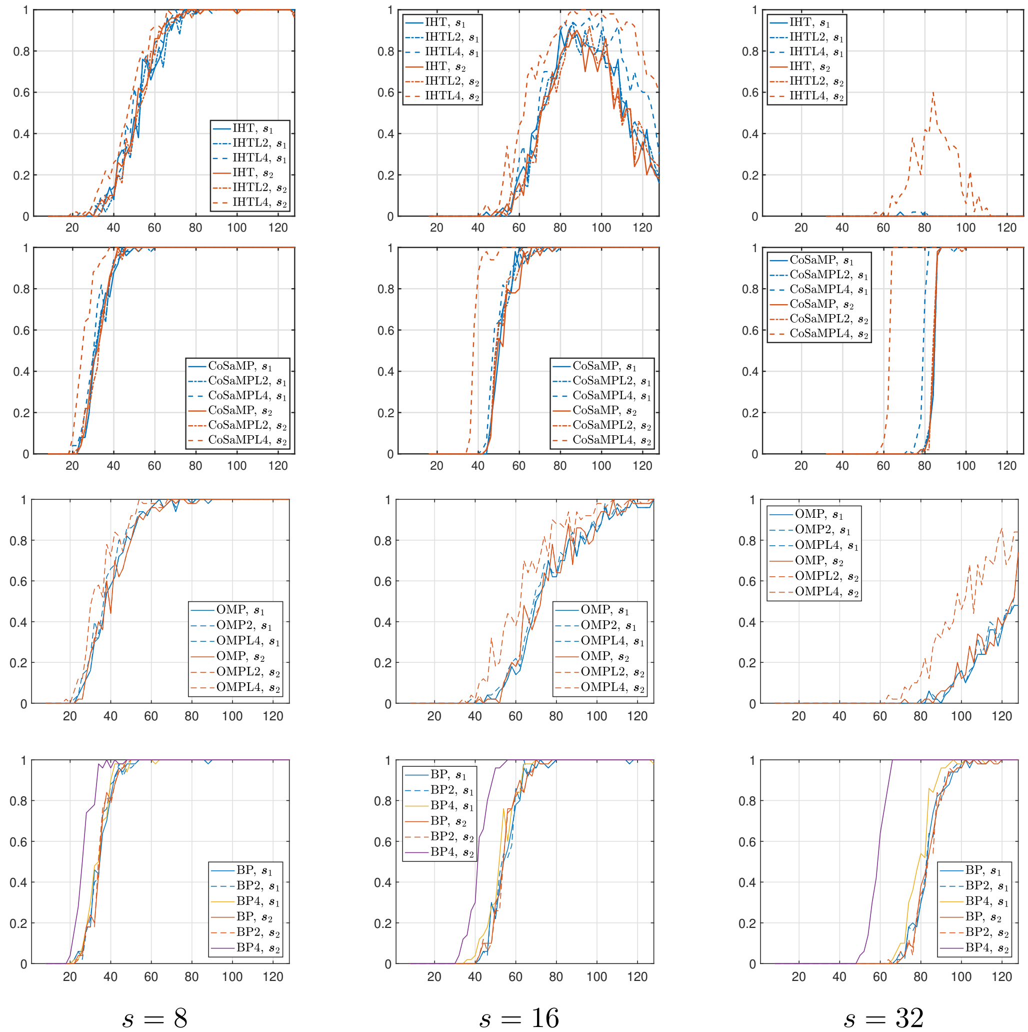


Figure 5.3: Horizontal phase transition line showing success probability versus m for various fixed total sparsities s. Four level sparsity with $\boldsymbol{M} = (N/4, N/2, 3N/4, N)$. The local sparsities are $\boldsymbol{s}_1 = (3s/8, s/8, 3s/8, s/8)$ and $\boldsymbol{s}_2 = (s/2, 0, s/2, 0)$. In the levels case we consider two-level algorithms based on $\boldsymbol{M} = (N/2, N)$ and $\boldsymbol{s} = (s/2, s/2)$ and four-level algorithms based on $\boldsymbol{M} = (N/4, N/2, 3N/4, N)$ and $\boldsymbol{s} = \boldsymbol{s}_1$ or $\boldsymbol{s} = \boldsymbol{s}_2$. Row one contains IHT, IHTL, row two CoSaMP and CoSaMPL, the third OMP and OMPL, and the final wQCBP.

behavior, we return to the simple two level case, with the true solution being $\boldsymbol{s} = (s/2, s/2)$ sparse in two levels $\boldsymbol{M} = (s/2, N)$. This is meant to model a natural problem of function approximation arising in compressive imaging [3]. This is exactly the problem of recovering a natural image using Fourier-Haar wavelet problem described at the end of Chapter 2, where one level is fully saturated to capture the coarse-scale wavelet coefficients, and exponentially fewer samples are required to recover fine scale details of an image. This is Fig. 5.4.

Here, we see marginal improvement for CoSaMPL and IHTL. However, OMPL does not improve over OMP in our experiments, contrasting with IHTL and CoSaMPL. This provokes the natural question of whether the formulation of OMPL here is the best possible - and if another variant would improve further. This calls back to the concerns we originally had when developing OMPL. Similarly, moving to QCBP with weights does not improve over its unweighted counterpart. This is not unexpected however, as the measurement condition for wQCBP here explicitly scales with the number of levels r. To be concrete, this falls into the special case of Eq. (2.28), requiring $\delta_{2\boldsymbol{s},\boldsymbol{M}} < \frac{1}{\sqrt{2r}+1}$. For suitable log factors L in s and N, a Guassian matrix requiresn$m \gtrsim \delta^{-2} \cdot sL$. Using our knowledge of the RICL this scales as $m \gtrsim 2r \cdot sL$. That is, we require a number of measurements scaling directly with the number of levels.

Finally we conclude with a four level experiment, Fig. 5.5. Here, the underlying vector is $\boldsymbol{s} = (s/2, 0, s/2, 0)$ sparse in four levels $\boldsymbol{M} = (N/4, N/2, 3N/4, N)$. Notably this sparsity pattern is only sensible up to $m = N/2$, so the axes are adjusted accordingly. Here, we see uniform improvement for each levels based algorithm. Again, we see CoSaMP appears to have the best phase transition curve, and IHT the worst.

It is also interesting to observe the phase transition behaivor of CoSaMP and wQCBP resembling each other rather closely. This is rather interesting considering the different in computational time, and sheds some positive light on CoSaMPL specifically.

As expected from the theory *a priori* knowledge allows for significant benefits, as then the recovery algorithm can respect the underlying sparsity structure. In the most extreme case, where some levels contain only zero entries, this – as the problem is essential reduced to one of smaller, in this case half, dimension.

There are several general takeaways from these experiments. Foremost is that using levels information in our recovery algorithms generally increases performance when the underlying solution exhibits sparsity in levels. This is especially pronounced for CoSaMP and IHT, whereas OMPL exhibits cases where it is outperformed by the sparse variant. However, in all our experiments, this was only present in a small subset.

Similarly to the sparse case, CoSaMP (and its variants) have overall better recovery in the sense of phase transitions, with OMP close behind. IHTL – as expected – performs generally the worst in terms of accuracy, trading off for computational time. However, the computational efficiency of OMPL combined with its reasonable accuracy and phase transitions, makes make it the most more desirable in larger scale computations. While CoSaMPL exhibits the best recovery, its runtime is lackluster and scales poorly along the phase transition region. Thus, attempting to reduce the number of measurements m leads to increased runtime for CoSaMPL - not a preferable design. Neither OMPL nor IHTL have this, with runtime generally constant in small examples.

To compare to our optimization baseline, wQCBP, OMPL and IHTL generally trade faster runtime for worse recovery for small m, whereas CoSaMPL has comparable runtime and better recovery than wQCBP. Thus depending on the computational budget, there is an iterative method outperforming wQCBP. In the low budget, large scale example, OMPL is the obvious choice, and in the large budget, small scale case, CoSaMPL is better suited. So, the summary of the numerics is that iterative methods – either in speed or in accuracy – can be used to outperform optimization approaches. Even in the cases where the performance is similar (i.e. moderately sized problems) – they are a useful tool to keep in mind, with guaranteed cost per iteration.

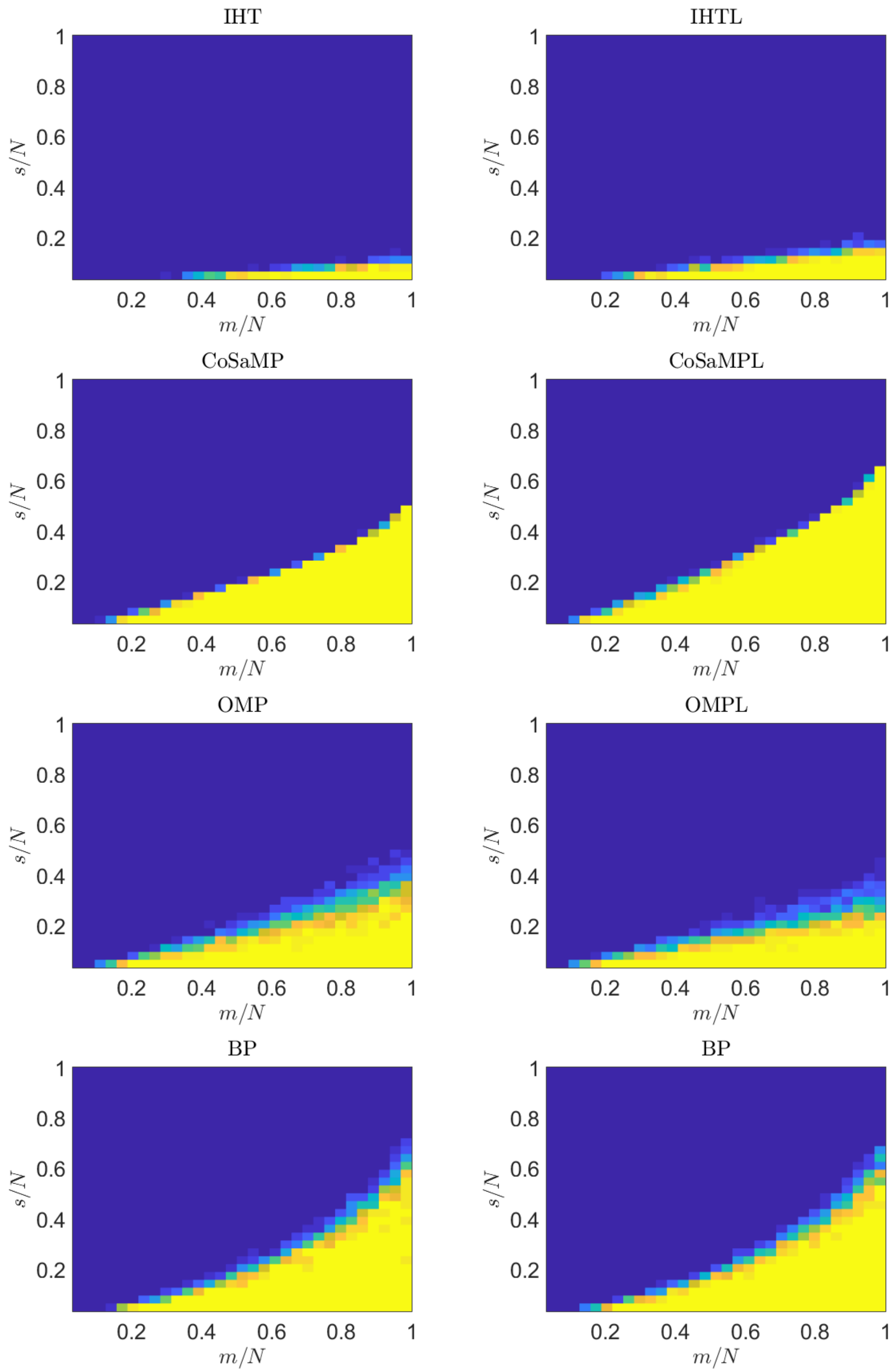


Figure 5.4: Phase transition plots comparing the standard sparse decoders of IHT, CoSaMP, and OMP against the levels-based generalizations for $N = 256$. Here, the underlying vector is $\boldsymbol{s} = (s/2, s/2)$ sparse in two levels $\boldsymbol{M} = (s/2, N)$. The final row is QCBP and wQCBP as a comparison.

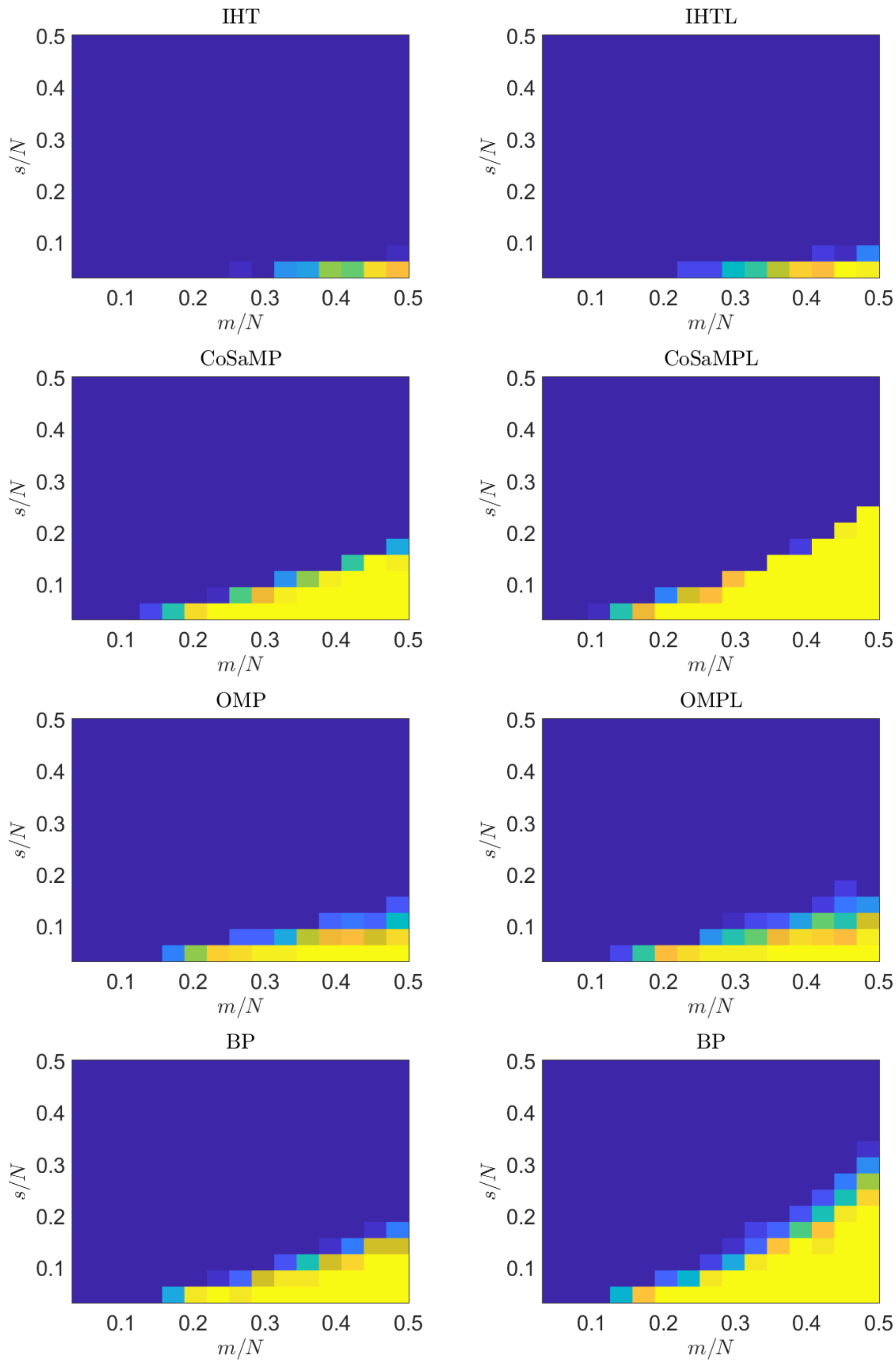


Figure 5.5: Phase transition plots comparing the standard sparse decoders of IHT, CoSaMP, and OMP against the levels-based generalizations for $N = 256$. Here, the underlying vector is $\boldsymbol{s} = (s/2, 0, s/2, 0)$ sparse in four levels $\boldsymbol{M} = (N/4, N/2, 3N/4, N)$. Row one contains IHT, IHTL, row two CoSaMP and CoSaMPL, the third OMP and OMPL, and the The final row is QCBP and wQCBP.

Chapter 6

Conclusions and Future Work

To conclude this thesis, we summarize the results within, and point in several interesting directions of future work.

The contributions within this work have foremost been the development and analysis of IHTL, CoSaMPL, and OMPL. These are extended from standard algorithms in the sparse case. We have extended stability and robustness for IHTL and CoSaMPL to sparse in levels vectors, under no more stringent assumptions than the sparse situation. Furthermore, these results have a beneficial scaling with the important parameters ζ, ξ. In the Fourier-Haar example, this was particularly noticeable as the measurement cost was reduced by $\log^2(N)$ compared to basis pursuit, a significant gain.

Numerically, we also observe the benefits of iterative approaches. Firstly, these methods are simple to implement and understand, and require no secondary algorithm to solve – unlike basis pursuit. Furthermore we have a guaranteed computational cost per iteration, and in the case of OMPL terminate after a fixed number of iterations. Numerically, this runtime improvement is observed rather strongly, with IHTL and OMPL having negligible runtime in comparison to BP. In combination with the reduced sampling complexity garnered from the sparse in levels structure, leads to extremely efficient recovery methods.

Furthermore, the phase transition behavior of the iterative methods is promising. CoSaMPL performs as well as, or better than QCBP, whereas OMPL and IHTL perform only marginally worse – and all methods have reasonable recovery for moderate number of samples. In fact, the levels based variations improve almost unilaterally over their sparse counterparts – with only OMPL having cases where it is outperformed by OMP.

In mentioning OMPL, it is worth emphasizing that the formulation here has not yet been shown to have formal guarantees of stability and robustness. In fact, this is only one of many potential versions of OMP in levels - and is a very interesting direction of future consideration. As mentioned throughout, OMP's guaranteed cost both per iteration and fixed number of iterations means a recovery result would be extremely desirable.

Putting this theory and numerical evidence together, we have developed an important tool for approximating sparse in levels signals. This is extremely useful in the case of compressive imaging. These tools – combined with design of the sensing matrix A – are possible approaches to construct to guaranteed polynomial time recovery of images from linear samples. This may be specifically useful in MRI imaging, where Fourier-Haar measurements with wavelet sparsity have seen success.

Bibliography

[1] B. Adcock. Infinite-dimensional compressed sensing and function interpolation. *Found. Comput. Math*, 18(3):661–701, 2018.

[2] B. Adcock, A. Bao, J. D. Jakeman, and A. Narayan. Compressed sensing with sparse corruptions: Fault-tolerant sparse collocation approximations. *SIAM/ASA J. Uncertain. Quantif.*, 6(4):1424–1453, 2018.

[3] B. Adcock, S. Brugiapaglia, and M. King-Roskamp. Do log factors matter? on optimal wavelet approximation and the foundations of compressed sensing. *arXiv:1905.10028*, 2019.

[4] B. Adcock, S. Brugiapaglia, and M. King-Roskamp. Iterative and greedy algorithms for the sparsity in levels model in compressed sensing. In *Wavelets and Sparsity XVIII*, volume 11138, page 1113809. International Society for Optics and Photonics, 2019.

[5] B. Adcock, S. Brugiapaglia, and C. Webster. Compressed sensing approaches for polynomial approximation of high-dimensional functions. In *Compressed Sensing and its Applications*, pages 93–124. Springer, 2017.

[6] B. Adcock and A. Hansen. *Compressive Imaging: Structure, Sampling, Learning.* 2020.

[7] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. *Forum Math. Sigma*, 5, 2017.

[8] B. Adcock, A. C. Hansen, and B. Roman. The quest for optimal sampling: computationally efficient, structure-exploiting measurements for compressed sensing. In *Compressed Sensing and Its Applications.* Birkhäuser, 2015.

[9] V. Antun, B. Adcock, and A. C. Hansen. Uniform recovery in infinite-dimensional compressed sensing and applications to structured binary sampling. *arXiv:1905.00126*, 2019.

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

[11] A. Bastounis and A. C. Hansen. On the absence of uniform recovery in many real-world applications of compressed sensing and the restricted isometry property and nullspace property in levels. *SIAM J. Imaging Sci.*, 10(1):335–371, 2017.

[12] T. Blumensath. Accelerated iterative hard thresholding. *Signal Process.*, 92:752–756, 2012.

[13] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.*, 14:629–654, 2008.

[14] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27:265?274, 2009.

[15] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE J. Sel. Top. Signal Process.*, 4(2), 2010.

[16] T. Cai and A. Zhang. Sharp RIP bound for sparse signal and low-rank matrix recovery. *Appl. Comput. Harmon. Anal.*, 35(1):74–93, 2013.

[17] E.J. Candes. The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris*, 346(9-10):589–592, 2008.

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

[19] I. Chun and B. Adcock. Optimal sparse recovery for multi-sensor measurements. In *IEEE Inf. Theory Workshop (ITW) 2016*, 2016.

[20] I.-Y. Chun and B. Adcock. Compressed sensing and parallel acquisition. *IEEE Trans. Inform. Theory*, 63(8):4860–4882, 2017.

[21] G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Engineering*, 33(7):2183 – 2191, 1994.

[22] S. Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *Found. Comput. Math.*, 16:1367–1396, 2016.

[23] A. Doostan and H. Owhadi. A non-adapted sparse approximation of pdes with stochastic inputs. *Journal of Computational Physics*, 230(8):3015–3034, 2011.

[24] D. Dorsch and H. Rauhut. Refined analysis of sparse mimo radar. *J. Fourier Anal. Appl.*, pages 1–45, 2016.

[25] M. F. Duarte and Y. C. Eldar. Structured compressed sensing: from theory to applications. *IEEE Trans. Signal Process.*, 59(9):4053–4085, 2011.

[26] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky. A wide-angle view at iterated shrinkage algorithms. In *Wavelets XII*, volume 6701, page 670102. International Society for Optics and Photonics, 2007.

[27] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhauser, 2013.

[28] C. Hegde, P. Indyk, and L. Schmidt. Approximation algorithms for model-based compressive sensing. *IEEE Trans. Inform. Theory*, 61(9), 2015.

[29] C. Li and B. Adcock. Compressed sensing with local structure: uniform recovery guarantees for the sparsity in levels class. *Appl. Comput. Harmon. Anal.*, 46:453—477, 2019.

[30] X. Li. Compressed sensing and matrix completion with a constant proportion of corruptions. *Constr. Approx.*, 37:73–99, 2013.

[31] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[32] D. Needell and J. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2008.

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

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

[35] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *in Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 1–3, 1993.

[36] B. Roman, A. C. Hansen, and B. Adcock. On asymptotic structure in compressed sensing. *arXiv:1406.4178*, 2014.

[37] Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all. *Appl. Comput. Harm. Anal.*, 45(1):170–205, 2018.

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

[39] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

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

[41] X. Yang and G. Karniadakis. Reweighted $\ell 1$ minimization method for stochastic elliptic differential equations. *Journal of Computational Physics*, 248:87–108, 2013.