

# Phase Transitions for Greedy Sparse Approximation Algorithms

Jeffrey D. Blanchard<sup>\*,a,1</sup>, Coralia Cartis<sup>b</sup>, Jared Tanner<sup>b,2</sup>, Andrew Thompson<sup>b</sup>

<sup>a</sup>*Department of Mathematics and Statistics, Grinnell College, Grinnell, Iowa 50112-1690, USA.*

<sup>b</sup>*School of Mathematics and the Maxwell Institute, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK.*

---

## Abstract

A major enterprise in compressed sensing and sparse approximation is the design and analysis of computationally tractable algorithms for recovering sparse, exact or approximate, solutions of underdetermined linear systems of equations. Many such algorithms have now been proven to have optimal-order uniform recovery guarantees using the ubiquitous Restricted Isometry Property (RIP) [11]. However, without specifying a matrix, or class of matrices, it is unclear when the RIP-based sufficient conditions on the algorithm are satisfied. Bounds on RIP constants can be inserted into the algorithms RIP based conditions, translating the conditions into requirements on the signal's sparsity level, length, and number of measurements. We illustrate this approach for Gaussian matrices on three of the state-of-the-art greedy algorithms: CoSaMP [29], Subspace Pursuit (SP) [13] and Iterative Hard Thresholding (IHT) [8]. Designed to allow a direct comparison of existing theory, our framework implies that, according to the best available analysis on these three algorithms, IHT requires the fewest number of compressed sensing measurements, has the best proven stability bounds, and has the lowest per iteration computational cost.

*Key words:* Compressed sensing, greedy algorithms, sparse solutions to underdetermined systems, restricted isometry property, phase transitions, Gaussian matrices.

---

## 1. Introduction

In compressed sensing [10, 11, 16], one works under the sparse approximation assumption, namely, that signals/vectors of interest can be well approximated by few components of a known basis. This assumption is often satisfied due to constraints imposed by the system which generates the signal. In this setting, it has been proven (originally in [11, 16] and by many others since) that the number of linear observations of the signal, required to guarantee recovery, need only be proportional to the sparsity of the signal's approximation. This is in stark contrast to the standard Shannon-Nyquist Sampling paradigm [36] where worst-case sampling requirements are imposed.

---

\*Corresponding author

*Email addresses:* [jeff@math.grinnell.edu](mailto:jeff@math.grinnell.edu) (Jeffrey D. Blanchard), [coralia.cartis@ed.ac.uk](mailto:coralia.cartis@ed.ac.uk) (Coralia Cartis), [jared.tanner@ed.ac.uk](mailto:jared.tanner@ed.ac.uk) (Jared Tanner), [a.thompson-8@sms.ed.ac.uk](mailto:a.thompson-8@sms.ed.ac.uk) (Andrew Thompson)

<sup>1</sup>JDB was supported by NSF DMS grant 0602219 while a VIGRE postdoctoral fellow at Department of Mathematics, University of Utah.

<sup>2</sup>JT acknowledges support from the Philip Leverhulme and the Alfred P. Sloan Fellowships.

Consider measuring a vector  $x_0 \in \mathbb{R}^N$  which either has exactly  $k < N$  nonzero entries or has  $k$  entries whose magnitudes are dominant. Let  $A$  be an  $n \times N$  matrix with  $n < N$  which we use to measure  $x_0$ . The observed measurements are often corrupted by additive noise, giving us the model  $y = Ax_0 + e$  for the  $n$  measurements where  $e$  denotes additive noise. From knowledge of  $y$  and  $A$  one seeks to recover the vector  $x_0$ , or a suitable approximation thereof, [9]. Let  $\chi^N(k) := \{x \in \mathbb{R}^N : \|x\|_0 \leq k\}$  denote the family of at most  $k$ -sparse vectors in  $\mathbb{R}^N$ , where  $\|\cdot\|_0$  counts the number of nonzero entries. From  $y$  and  $A$ , the optimal  $k$ -sparse signal is the solution of

$$\min_{x \in \chi^N(k)} \|Ax - y\|, \quad (1)$$

for a suitably chosen norm.

Solving (1) via a naive exhaustive search of all  $x \in \chi^N(k)$  is combinatorial in nature and NP-hard [28]. A major aspect of compressed sensing theory is the study of alternative methods for solving (1). Since the system  $y = Ax + e$  is underdetermined, any successful recovery of  $x$  will require some form of nonlinear reconstruction. Under certain conditions, various algorithms have been shown to successfully reduce (1) to a tractable problem, one with a computational cost which is a low degree polynomial of the problem dimensions, rather than the exponential cost associated with a direct combinatorial search for the solution of (1). While there are numerous reconstruction algorithms, they each generally fall into one of three categories: *greedy methods*, *regularizations*, or *combinatorial group testing*. For an in-depth discussion of compressed sensing recovery algorithms, see [29] and references therein.

The first uniform guarantees for exact reconstruction of every  $x \in \chi^N(k)$ , for a fixed  $A$ , came from  $\ell_1$ -regularization, where (1) is relaxed to solving the problem

$$\min_{x \in \mathbb{R}^N} \|x\|_1 \text{ subject to } \|Ax - y\|_2 < \gamma, \quad (2)$$

for some known noise level  $\gamma \sim \|e\|_2$  or by testing various values of  $\gamma$ .  $\ell_1$ -regularization has been extensively studied, see the pioneering works [11, 16]; also, see [15, 21, 5] for results analogous to those presented here. In this paper, we focus on three illustrative greedy algorithms, *Compressed Sensing Matching Pursuit* (CoSaMP) [29], *Subspace Pursuit* (SP) [13], and *Iterative Hard Thresholding* (IHT) [8], which boast similar uniform guarantees of successful recovery of sparse signals when the measurement matrix  $A$  satisfies the now ubiquitous *Restricted Isometry Property* (RIP) [11, 5]. The three algorithms are deeply connected and each have some advantage over the other. These algorithms are essentially support set recovery algorithms which use hard thresholding to iteratively update the approximate support set; their differences lie in the magnitude of the application of hard thresholding and the vectors to which the thresholding is applied, [18, 37]. The algorithms are restated in the Section 2. Other greedy methods with similar guarantees are available, see for example [12, 27]; several other greedy techniques have been developed ([24, 30, 14], etc.), but their theoretical analyses either do not currently subscribe to the above uniform framework, or as in the case of precursors to CoSaMP and SP, the algorithms OMP [14] and ROMP [30] are not known whether they achieve the optimal order.

As briefly mentioned earlier, the intriguing aspect of compressed sensing is its ability to recover  $k$ -sparse signals when the number of measurements required is proportional to the sparsity,  $n \sim k$ , as the problem size grows,  $n \rightarrow \infty$ . Each of the algorithms discussed here exhibit a phase transition property, where there exists a  $k_n^*$  such that for any  $\epsilon > 0$ , as  $k_n^*, n \rightarrow \infty$ , the algorithm successfully recovers *all*  $k$ -sparse vectors (exactly when no noise is present) provided  $k < (1 - \epsilon)k_n^*$  and does not recover all

$k$ -sparse vectors if  $k > (1 + \epsilon)k_n^*$ . For a description of phase transitions in the context of compressed sensing, see [22], while for empirically observed average-case phase transitions for greedy algorithms, see [18]. (Note that all recovery guarantees discussed here are for the recovery of *all*  $k$ -sparse vectors, and substantially different behavior is observed in average-case testing of the algorithms [18].) We consider the asymptotic setting where  $k$  and  $N$  grow proportionally with  $n$ , namely,  $(k, n, N) \rightarrow \infty$  with the ratios  $\frac{k}{n} \rightarrow \rho$ ,  $\frac{N}{n} \rightarrow \delta$  as  $n \rightarrow \infty$  for  $(\delta, \rho) \in (0, 1)^2$ ; also, we assume the matrix  $A$  is drawn i.i.d. from  $\mathcal{N}(0, n^{-1})$ , the normal distribution with mean 0 and variance  $n^{-1}$ . In this framework, we develop lower bounds on the phase transitions,  $k_n^*/n$ , for exact recovery of all  $k$ -sparse signals. These bounds provide curves in the unit square,  $(\delta, \rho) \in (0, 1)^2$ , below which there is *overwhelming probability* (probability approaching 1 exponentially in  $n$ ) on the draw of the Gaussian matrix  $A$ , that  $A$  will satisfy the sufficient RIP conditions and therefore solve (1). We utilize a more general, asymmetric version of the RIP, see Definition 1, to compute as precise a lower bound on the phase transitions as possible. This phase transition framework allows a direct comparison of the provable recovery regions of different algorithms in terms of the problem instance  $(\frac{N}{n}, \frac{k}{n})$ . We then compare the guaranteed recovery capabilities of these algorithms to the guarantees of  $\ell_1$ -regularization proven via a similar RIP analysis. For  $\ell_1$ -regularization, this phase transition framework has already been applied using the RIP [5], using the theory of convex polytopes [15] and geometric functional analysis [35].

The aforementioned lower bounds on the algorithmic sparse recovery phase transitions are presented in Theorems 9, 10, and 11. The curves are defined by functions  $\rho_S^{csp}(\delta)$  (CoSaMP; the black curve in Figure 1(a)),  $\rho_S^{sp}(\delta)$  (SP; the magenta curve in Figure 1(a)), and  $\rho_S^{iht}(\delta)$  (IHT; the red curve in Figure 1(a)). For comparison, the analogous lower bound on the phase transition for  $\rho_S^{\ell_1}(\delta)$  ( $\ell_1$ -regularization) derived using RIP is displayed as the blue curve in Figure 1(a). (For  $\ell_1$ -regularization substantially better bounds have been proven using other methods of analysis [15, 38].) From Figure 1, we are able to directly compare the provable recovery results of the three greedy algorithms as well as  $\ell_1$ -regularization. For a given problem instance  $(k, n, N)$  with the entries of  $A$  drawn i.i.d. from  $\mathcal{N}(0, n^{-1})$ , if  $\frac{k}{n} = \rho$  falls in the region below the curve  $\rho_S^{alg}(\delta)$  associated to a specific algorithm, then with probability approaching 1 exponentially in  $n$  the algorithm will exactly recover the  $k$ -sparse vector  $x \in \chi^N(k)$  no matter which  $x \in \chi^N(k)$  was measured by  $A$ . These lower bounds on the phase transition can also be interpreted as the minimum number of measurements known to guarantee recovery through the constant of proportionality:  $n > \left(\rho_S^{alg}\right)^{-1} k$ . Figure 1(b) portrays the inverse of the lower bounds on the phase transition. This gives a minimum known value for  $\left(\rho_S^{alg}\right)^{-1}$ . For example, from the blue curve, for a Gaussian matrix used in  $\ell_1$ -regularization, the minimum number of measurements proven (using RIP) to be sufficient to ensure recovery of all  $k$ -sparse vectors is  $n > 317k$ . By contrast, for greedy algorithms, the minimum number of measurements shown to be sufficient is significantly larger:  $n > 907k$  for IHT,  $n > 3124k$  for SP, and  $n > 4923k$  for CoSaMP.

More precisely, the main contributions of this article is the derivation of theorems and corollaries of the following form for each of the CoSaMP, SP, and IHT algorithms.

**Theorem 1.** *Given a matrix  $A$  with entries drawn i.i.d. from  $\mathcal{N}(0, n^{-1})$ , for any  $x \in \chi^N(k)$ , let  $y = Ax + e$  for some (unknown) noise vector  $e$ . For any  $\epsilon \in (0, 1)$ , as  $(k, n, N) \rightarrow \infty$  with  $n/N \rightarrow \delta \in (0, 1)$  and  $k/n \rightarrow \rho \in (0, 1)$ , there exists  $\mu^{alg}(\delta, \rho)$  with  $\rho_S^{alg}(\delta)$  the unique solution to  $\mu^{alg}(\delta, \rho) = 1$ . If  $\rho < (1 - \epsilon)\rho_S^{alg}(\delta)$ , there is overwhelming probability on the draw of  $A$  that the output of the algorithm*

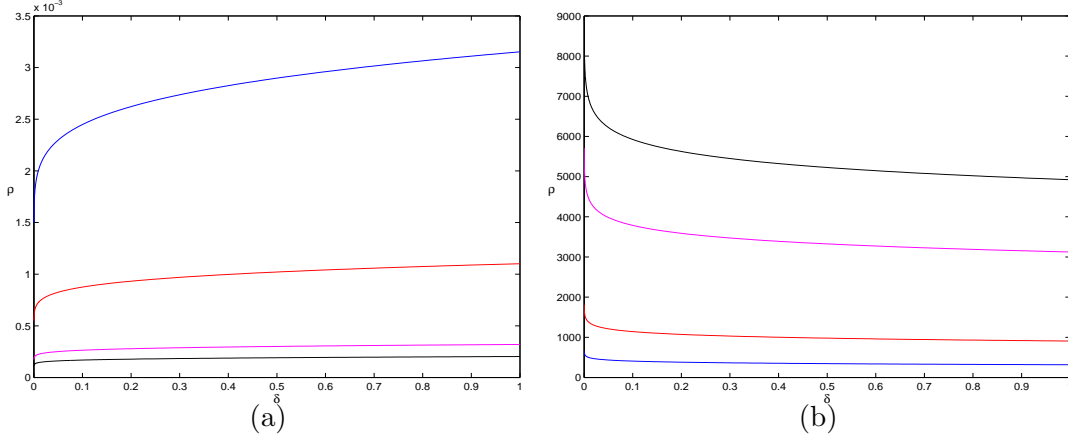


Figure 1: (a): The lower bounds on the Strong exact recovery phase transition for Gaussian matrices for the algorithms  $\ell_1$ -regularization ([5]  $\rho_S^{\ell_1}(\delta)$ , blue), IHT (Theorem 11,  $\rho_S^{\text{iht}}(\delta)$ , red), SP (Theorem 10,  $\rho_S^{\text{sp}}(\delta)$ , magenta), and CoSaMP (Theorem 9,  $\rho_S^{\text{csp}}(\delta)$ , black). (b): The inverse of the phase transition lower bounds in the left panel (a).

at the  $l^{\text{th}}$  iteration,  $\hat{x}$ , approximates  $x$  within the bound

$$\|x - \hat{x}\|_2 \leq \kappa^{\text{alg}}(\delta, (1 + \epsilon)\rho) \left[ \mu^{\text{alg}}(\delta, (1 + \epsilon)\rho) \right]^l \|x\|_2 + \frac{\xi^{\text{alg}}(\delta, (1 + \epsilon)\rho)}{1 - \mu^{\text{alg}}(\delta, (1 + \epsilon)\rho)} \|e\|_2, \quad (3)$$

for some  $\kappa^{\text{alg}}(\delta, \rho)$  and  $\xi^{\text{alg}}(\delta, \rho)$ .

The factors  $\mu^{\text{alg}}(\delta, \rho)$  and  $\frac{\xi^{\text{alg}}}{1 - \mu^{\text{alg}}}(\delta, \rho)$  for CoSaMP, SP, and IHT are displayed in Figure 2, while formulae for their calculation are deferred to Section 3. Even more than Theorem 1 can be said when the measurements are exact.

**Corollary 2.** *Given a matrix  $A$  with entries drawn i.i.d. from  $\mathcal{N}(0, n^{-1})$ , for any  $x \in \chi^N(k)$ , let  $y = Ax$ . For any  $\epsilon \in (0, 1)$ , with  $n/N \rightarrow \delta \in (0, 1)$  and  $k/n \rightarrow \rho < (1 - \epsilon)\rho_S^{\text{alg}}(\delta)$  as  $(k, n, N) \rightarrow \infty$ , there is overwhelming probability on the draw of  $A$  that the algorithm exactly recovers  $x$  from  $y$  and  $A$  in a finite number of iterations not to exceed*

$$\ell_{\text{max}}^{\text{alg}}(x) := \left\lceil \frac{\log \nu_{\text{min}}(x) - \log \kappa^{\text{alg}}(\delta, \rho)}{\log \mu^{\text{alg}}(\delta, \rho)} \right\rceil + 1 \quad (4)$$

where

$$\nu_{\text{min}}(x) := \frac{\min_{i \in T} |x_i|}{\|x\|_2} \quad (5)$$

with  $T := \{i : x_i \neq 0\}$  and  $\lceil m \rceil$ , the smallest integer greater than or equal to  $m$ .

Corollary 2 implies that  $\rho_S^{\text{alg}}(\delta)$  delineates a region where, when there exists an  $x \in \chi^N(k)$  such that  $y = Ax$ , the algorithm is guaranteed to recover  $x$  exactly. However, if no such  $x$  exists, as  $\rho$  approaches  $\rho_S^{\text{alg}}(\delta)$  the guarantees on the number of iteratives required and stability factors become unbounded. Further bounds on the convergence factor  $\mu^{\text{alg}}(\delta, \rho)$  and the stability factor  $\frac{\xi^{\text{alg}}}{1 - \mu^{\text{alg}}}(\delta, \rho)$  result in yet lower curves  $\rho_S^{\text{alg}}(\delta; \text{bound})$  for a specified *bound*; recall that  $\rho_S^{\text{alg}}(\delta)$  corresponds to the relationship  $\mu^{\text{alg}}(\delta, \rho) = 1$ .

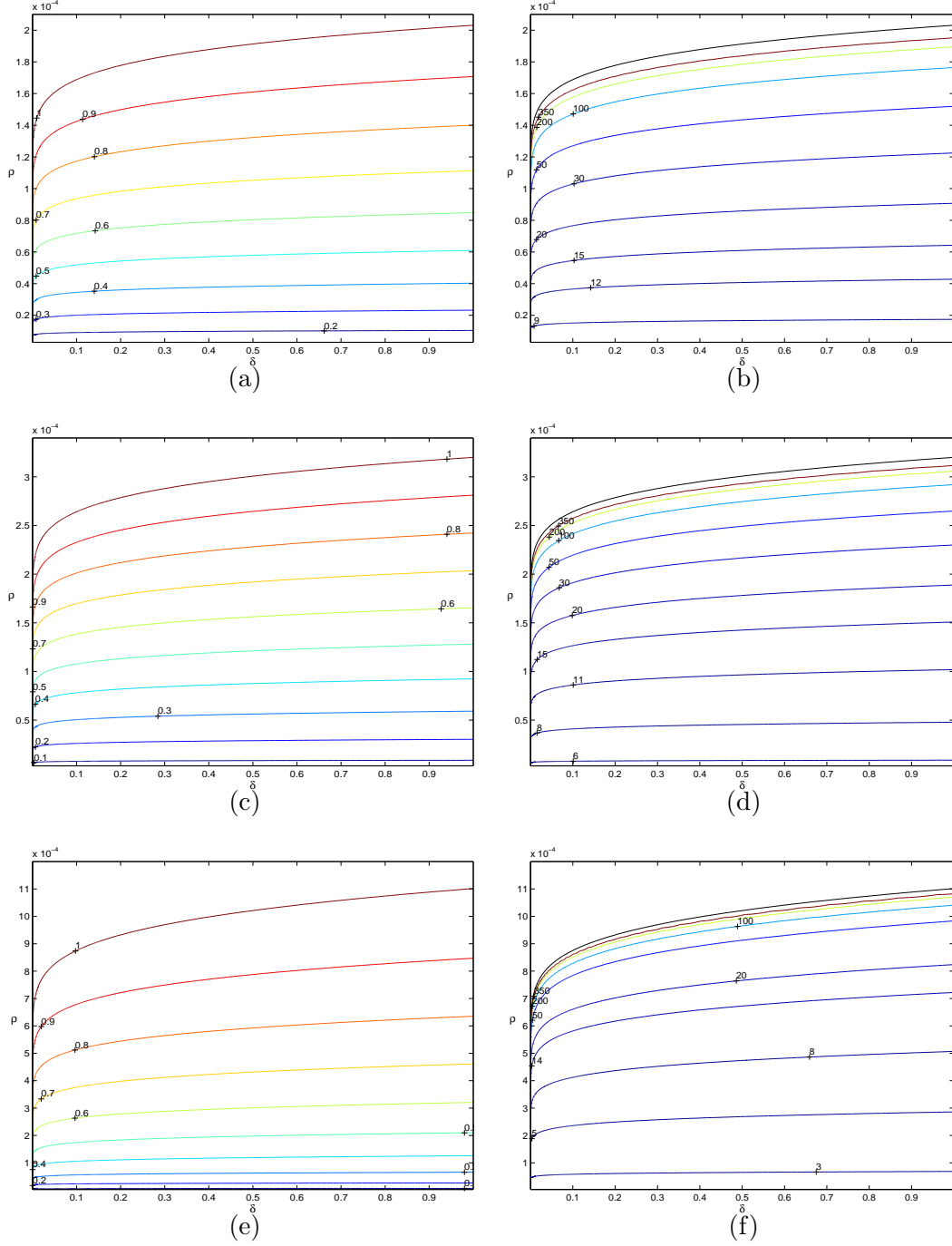


Figure 2: Contour plots of the convergence factor  $\mu^{alg}(\delta, \rho)$  and the stability factor  $\frac{\xi^{alg}}{1-\mu^{alg}}(\delta, \rho)$ , in the left and right panels respectively, for CoSaMP (a-b), SP (c-d), and IHT (e-f). The highest curve in the left panels corresponds to  $\mu^{alg}(\delta, \rho) = 1$  which implies  $\frac{\xi^{alg}}{1-\mu^{alg}}(\delta, \rho) = \infty$ .

In the next section, we recall the three algorithms and introduce necessary notation. Then we present the asymmetric RIP and formulate weaker restricted isometry conditions on a matrix  $A$  that ensure the respective algorithm will successfully recover all  $k$ -sparse signals. In order to make quantitative comparisons of these results, we must select a matrix ensemble for analysis. In Section 3, we present the lower bounds on the phase transition for each algorithm when the measurement matrix is a Gaussian matrix. Phase transitions are developed in the case of exact sparse signals while bounds on the multiplicative stability constants are also compared through associated level curves. Section 4 is a discussion of our interpretation of these results and shows how this phase transition framework is a unifying model for the comparison of compressed sensing algorithms.

## 2. Greedy Algorithms and the Asymmetric Restricted Isometry Property

Let us define some notation used in the algorithms discussed here. For an index set  $I \subset \{1, \dots, N\}$ , let  $x_I$  denote the restriction of a vector  $x \in \mathbb{R}^N$  to the set  $I$ , i.e.,  $(x_I)_i = x_i$  for  $i \in I$  and  $(x_I)_j = 0$  for  $j \notin I$ . Also, let  $A_I$  denote the submatrix of  $A$  obtained by selecting the columns  $A$  indexed by  $I$ .  $A_I^*$  is the conjugate transpose of  $A_I$  while  $A_I^\dagger = (A_I^* A_I)^{-1} A_I^*$  is the pseudoinverse of  $A_I$ . In each of the algorithms, thresholding is applied by selecting  $m$  entries of a vector with largest magnitude; we refer to this as hard thresholding of magnitude  $m$ .

### 2.1. CoSaMP

The CoSaMP recovery algorithm is a support recovery algorithm which applies hard thresholding by selecting the  $k$  largest entries of a vector obtained by applying a pseudoinverse to the measurement  $y$ . In CoSaMP, the columns of  $A$  selected for the pseudoinverse are obtained by applying hard thresholding of magnitude  $2k$  to  $A^*$  applied to the residual from the previous iteration and adding these indices to the approximate support set from the previous iteration. This larger pseudoinverse matrix of size  $2k \times n$  imposes the most stringent aRIP condition of the three algorithms. However, CoSaMP uses one fewer pseudoinverse per iteration than SP as the residual vector is computed with a direct matrix-vector multiply of size  $n \times k$  rather than with an additional pseudoinverse. Furthermore, when computing the output vector  $\hat{x}$ , CoSaMP does not need to apply another pseudoinverse as does SP. See Algorithm 1.

### 2.2. Subspace Pursuit

The Subspace Pursuit algorithm is also a support recovery algorithm which applies hard thresholding of magnitude  $k$  to a vector obtained by applying a pseudoinverse to the measurements  $y$ . The submatrix chosen for the pseudoinverse has its columns selected by applying  $A^*$  to the residual vector from the previous iteration, hard thresholding of magnitude  $k$ , and adding the indices of the terms to the previous approximate support set. The aforementioned residual vector is also computed via a pseudoinverse, this time selecting the columns from  $A$  by again applying a hard threshold of magnitude  $k$ . The computation of the approximation to the target signal also requires the application of a pseudoinverse for a matrix of size  $n \times k$ . See Algorithm 2.

### 2.3. Iterative Hard Thresholding

Iterative Hard Thresholding (IHT) is also a support recovery algorithm. However, IHT applies hard thresholding to an approximation of the target signal, rather than to the residuals. This completely eliminates the use of a pseudoinverse, reducing the computational cost per iteration. In particular,

---

**Algorithm 1** CoSaMP [29]

---

**Input:**  $A, y, k$ **Output:** A  $k$ -sparse approximation  $\hat{x}$  of the target signal  $x$ 

---

**Initialization:**

- 1: Set  $T^0 = \emptyset$
- 2: Set  $y^0 = y$

**Iteration:** During iteration  $l$ , **do**

- 1:  $\tilde{T}^l = T^{l-1} \cup \{2k \text{ indices of largest magnitude entries of } A^*y^{l-1}\}$
  - 2:  $\tilde{x} = A_{\tilde{T}^l}^\dagger y$
  - 3:  $T^l = \{k \text{ indices of largest magnitude entries of } \tilde{x}\}$
  - 4:  $y^l = y - A_{T^l} \tilde{x}_{T^l}$
  - 5: **if**  $\|y^l\|_2 = 0$  **then**
  - 6:     **return**  $\hat{x}$  defined by  $\hat{x}_{\{1, \dots, N\} - T^l} = 0$  and  $\hat{x}_{T^l} = \tilde{x}_{T^l}$
  - 7: **else**
  - 8:     Perform iteration  $l + 1$
  - 9: **end if**
- 

---

**Algorithm 2** Subspace Pursuit [13]

---

**Input:**  $A, y, k$ **Output:** A  $k$ -sparse approximation  $\hat{x}$  of the target signal  $x$ 

---

**Initialization:**

- 1: Set  $T^0 = \{k \text{ indices of largest magnitude entries of } A^*y\}$
- 2: Set  $y_r^0 = y - A_{T^0} A_{T^0}^\dagger y$

**Iteration:** During iteration  $l$ , **do**

- 1:  $\tilde{T}^l = T^{l-1} \cup \{k \text{ indices of largest magnitude entries of } A^*y_r^{l-1}\}$
  - 2: Set  $\tilde{x} = A_{\tilde{T}^l}^\dagger y$
  - 3:  $T^l = \{k \text{ indices of largest magnitude entries of } \tilde{x}\}$
  - 4:  $y_r^l = y - A_{T^l} A_{T^l}^\dagger y$
  - 5: **if**  $\|y_r^l\|_2 = 0$  **then**
  - 6:     **return**  $\hat{x}$  defined by  $\hat{x}_{\{1, \dots, N\} - T^l} = 0$  and  $\hat{x}_{T^l} = A_{T^l}^\dagger y$
  - 7: **else**
  - 8:     Perform iteration  $l + 1$
  - 9: **end if**
- 

hard thresholding of magnitude  $k$  is applied to an updated approximation of the target signal,  $x$ , obtained by matrix-vector multiplies of size  $n \times N$  that represent a move by a fixed stepsize  $\omega$  along the steepest descent direction from the current iterate for the residual  $\|Ax - y\|_2^2$ . See Algorithm 3.

**Remark 1. (Stopping criteria for greedy methods)** *In the case of corrupted measurements, where  $y = Ax + e$  for some noise vector  $e$ , the stopping criteria listed in Algorithms 1-3 may never be achieved. Therefore, a suitable alternative stopping criteria must be employed. For our analysis on bounding the error of approximation in the noisy case, we bound the approximation error if the*

---

**Algorithm 3** Iterative Hard Thresholding [8]

---

**Input:**  $A, y, \omega \in (0, 1), k$ **Output:** A  $k$ -sparse approximation  $\hat{x}$  of the target signal  $x$ 

---

**Initialization:**

- 1: Set  $x^0 = 0$
- 2: Set  $T^0 = \emptyset$
- 3: Set  $y^0 = y$

**Iteration:** During iteration  $l$ , **do**

- 1:  $x^l = x_{T^{l-1}}^{l-1} + wA^*y^{l-1}$
  - 2:  $T^l = \{k \text{ indices of largest magnitude entries of } x^l\}$
  - 3:  $y^l = y - A_{T^l}x_{T^l}^l$
  - 4: **if**  $\|y^l\|_2 = 0$  **then**
  - 5:     **return**  $\hat{x}$  defined by  $\hat{x}_{\{1, \dots, N\} - T^l} = 0$  and  $\hat{x}_{T^l} = x_{T^l}^l$
  - 6: **else**
  - 7:     Perform iteration  $l + 1$
  - 8: **end if**
- 

algorithm terminates after  $l$  iterations. For example, we could change the algorithm to require a maximum number of iterations  $l$  as an input and then terminate the algorithm if our stopping criteria is not met in fewer iterations. In practice, the user would be better served to stop the algorithm when the residual is no longer improving. For a more thorough discussion of suitable stopping criteria for each algorithm in the noisy case, see the original announcement of the algorithms [8, 13, 29].

#### 2.4. The Asymmetric Restricted Isometry Property

In this section we relax the sufficient conditions originally placed on Algorithms 1-3 by employing a more general notion of a restricted isometry. As discussed in [5], the singular values of the  $n \times k$  submatrices of an arbitrary measurement matrix  $A$  do not, in general, deviate from unity symmetrically. The standard notion of the *restricted isometry property* (RIP) [11] has an inherent symmetry which is unnecessarily restrictive. Hence, seeking the best possible conditions for the measurement matrix under which Algorithms 1-3 will provably recovery every  $k$  sparse vector, we reformulate the sufficient conditions in terms of the *asymmetric restricted isometry property* (aRIP) [5]. (Foucart and Lai also proposed an aRIP motivated by imposing scale invariance on the RIP [26].)

**Definition 1.** For an  $n \times N$  matrix  $A$ , the asymmetric RIP constants  $L(k, n, N)$  and  $U(k, n, N)$  are defined as:

$$L(k, n, N) := \min_{c \geq 0} c \text{ subject to } (1 - c)\|x\|_2^2 \leq \|Ax\|_2^2, \forall x \in \chi^N(k); \quad (6)$$

$$U(k, n, N) := \min_{c \geq 0} c \text{ subject to } (1 + c)\|x\|_2^2 \geq \|Ax\|_2^2, \forall x \in \chi^N(k). \quad (7)$$

**Remark 2.** 1. The more common, symmetric definition of the RIP constants is recovered by defining  $R(k, n, N) = \max\{L(k, n, N), U(k, n, N)\}$ . In this case, a matrix  $A$  of size  $n \times N$  has the RIP constant  $R(k, n, N)$  if

$$R(k, n, N) := \min_{c \geq 0} c \text{ subject to } (1 - c)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + c)\|x\|_2^2, \forall x \in \chi^N(k).$$

2. Observe that  $\chi^N(k) \subset \chi^N(k+1)$  for any  $k$  and therefore the constants  $L(k, n, N)$ ,  $U(k, n, N)$ , and  $R(k, n, N)$  are nondecreasing in  $k$  [11].
3. For all expressions involving  $L(\cdot, n, N)$  it is understood, without explicit statement, that the first argument is limited to the range where  $L(\cdot, n, N) < 1$ . Beyond this range of sparsity, there exist vectors which are mapped to zero, and are unrecoverable.

Using the aRIP, we analyze the three algorithms in the case of a general measurement matrix  $A$  of size  $n \times N$ . For each algorithm, the application of Definition 1 results in a relaxation of the previously known RIP based conditions imposed on  $A$  to provably guarantee recovery of all  $x \in \chi^N(k)$ . We first present a stability result for each algorithm in terms of bounding the approximation error of the output after  $l$  iterations. The bounds show a multiplicative stability constant in terms of aRIP constants that amplifies the total energy of the noise. As a corollary, we obtain a sufficient condition on  $A$  in terms of the aRIP for exact recovery of all  $k$ -sparse vectors. The proofs of Theorems 5, 6, and 7 are available in an extended preprint [6]. These theorems and corollaries take the same form, differing for each algorithm only by the formulae for various factors. We state the general form of the theorems and corollaries, analogous to Theorem 1 and Corollary 2, and then state the formulae for each of the algorithms CoSaMP, SP, and IHT.

**Theorem 3.** *Given a matrix  $A$  of size  $n \times N$  with aRIP constants  $L(\cdot, n, N)$  and  $U(\cdot, n, N)$ , for any  $x \in \chi^N(k)$ , let  $y = Ax + e$ , for some (unknown) noise vector  $e$ . Then there exists  $\mu^{alg}(k, n, N)$  such that if  $\mu^{alg}(k, n, N) < 1$ , the output  $\hat{x}$  of algorithm “alg” at the  $l^{th}$  iteration approximates  $x$  within the bound*

$$\|x - \hat{x}\|_2 \leq \kappa^{alg}(k, n, N) \left[ \mu^{alg}(k, n, N) \right]^l \|x\|_2 + \frac{\xi^{alg}(k, n, N)}{1 - \mu^{alg}(k, n, N)} \|e\|_2, \quad (8)$$

for some  $\kappa^{alg}(k, n, N)$  and  $\xi^{alg}(k, n, N)$ .

**Corollary 4.** *Given a matrix  $A$  of size  $n \times N$  with aRIP constants  $L(\cdot, n, N)$  and  $U(\cdot, n, N)$ , for any  $x \in \chi^N(k)$ , let  $y = Ax$ . Then there exists  $\mu^{alg}(k, n, N)$  such that if  $\mu^{alg}(k, n, N) < 1$ , the algorithm “alg” exactly recovers  $x$  from  $y$  and  $A$  in a finite number of iterations not to exceed*

$$\ell_{max}^{alg}(x) := \left\lceil \frac{\log \nu_{min}(x) - \log \kappa^{alg}(k, n, N)}{\log \mu^{alg}(k, n, N)} \right\rceil + 1 \quad (9)$$

with  $\nu_{min}(x)$  defined as in (5).

We begin with Algorithm 1, the Compressive Sampling Matching Pursuit recovery algorithm of Needell and Tropp [29]. We relax the sufficient recovery condition in [29] via the aRIP.

**Theorem 5 (CoSaMP).** *Theorem 3 and Corollary 4 are satisfied by CoSaMP, Algorithm 1, with  $\kappa^{csp}(k, n, N) := 1$  and  $\mu^{csp}(k, n, N)$  and  $\xi^{csp}(k, n, N)$  defined as*

$$\mu^{csp}(k, n, N) := \frac{1}{2} \left( 2 + \frac{L(4k, n, N) + U(4k, n, N)}{1 - L(3k, n, N)} \right) \left( \frac{L(2k, n, N) + U(2k, n, N) + L(4k, n, N) + U(4k, n, N)}{1 - L(2k, n, N)} \right) \quad (10)$$

and

$$\xi^{csp}(k, n, N) := 2 \left\{ \left( 2 + \frac{L(4k, n, N) + U(4k, n, N)}{1 - L(3k, n, N)} \right) \left( \frac{\sqrt{1 + U(2k, n, N)}}{1 - L(2k, n, N)} \right) + \frac{1}{\sqrt{1 - L(3k, n, N)}} \right\}. \quad (11)$$

Next, we apply the aRIP to Algorithm 2, Dai and Milenkovic's Subspace Pursuit [13]. Again, the aRIP provides a sufficient condition that admits a wider range of measurement matrices than admitted by the symmetric RIP condition derived in [13].

**Theorem 6 (SP).** *Theorem 3 and Corollary 4 are satisfied by Subspace Pursuit, Algorithm 2, with  $\kappa^{sp}(k, n, N)$ ,  $\mu^{sp}(k, n, N)$ , and  $\xi^{sp}(k, n, N)$  defined as*

$$\kappa^{sp}(k, n, N) := 1 + \frac{U(2k, n, N)}{1 - L(k, n, N)}, \quad (12)$$

$$\mu^{sp}(k, n, N) := \frac{2U(3k, n, N)}{1 - L(k, n, N)} \left( 1 + \frac{2U(3k, n, N)}{1 - L(2k, n, N)} \right) \left( 1 + \frac{U(2k, n, N)}{1 - L(k, n, N)} \right) \quad (13)$$

and

$$\begin{aligned} \xi^{sp}(k, n, N) := & \frac{\sqrt{1 + U(k, n, N)}}{1 - L(k, n, N)} \left[ 1 - \mu^{sp}(k, n, N) + 2\kappa^{sp}(k, n, N) \left( 1 + \frac{2U(3k, n, N)}{1 - L(2k, n, N)} \right) \right] \\ & + \frac{2\kappa^{sp}(k, n, N)}{\sqrt{1 - L(2k, n, N)}}. \end{aligned} \quad (14)$$

Finally, we apply the aRIP analysis to Algorithm 3, Iterative Hard Thresholding for Compressed Sensing introduced by Blumensath and Davies [8]. Theorem 7 employs the aRIP to provide a weaker sufficient condition than derived in [8].

**Theorem 7 (IHT).** *Theorem 3 and Corollary 4 are satisfied by Iterative Hard Thresholding, Algorithm 3, with  $\kappa^{iht}(k, n, N) := 1$  and  $\mu^{iht}(k, n, N)$  and  $\xi^{iht}(k, n, N)$  defined as*

$$\mu^{iht}(k, n, N) := 2\sqrt{2} \max \{ \omega [1 + U(3k, n, N)] - 1, 1 - \omega [1 - L(3k, n, N)] \}. \quad (15)$$

and

$$\xi^{iht}(k, n, N) := 2\omega \sqrt{1 + U(2k, n, N)}. \quad (16)$$

**Remark 3.** *Each of Theorems 5, 6 and 7 are derived following the same recipe as in [29], [13] and [8], respectively, using the aRIP rather than the RIP and taking care to maintain the least restrictive bounds at each step (for details, see the extended preprint [6]). For IHT, the aRIP is simply a scaling of the matrix so that its RIP bounds are minimal. This is possible for IHT as the factors in  $\mu^{iht}(k, n, N)$  involve  $L(\alpha k, n, N)$  and  $U(\alpha k, n, N)$  for only one value of  $\alpha$ , here  $\alpha = 3$ . No such scaling interpretation is possible for CoSaMP and SP.*

### 3. Phase Transitions for Greedy Algorithms with Gaussian Matrices

The quantities  $\mu^{alg}(k, n, N)$  and  $\xi^{alg}(k, n, N)$  in Theorems 5, 6, and 7 dictate the current theoretical convergence bounds for CoSaMP, SP, and IHT. (These are uniform bounds over *all*  $k$ -sparse vectors  $x$ .) Although some comparisons can be made between the forms of  $\mu^{alg}$  and  $\xi^{alg}$  for different algorithms, it is not possible to quantitatively state for what range of  $k$  the algorithm will satisfy bounds on  $\mu^{alg}(k, n, N)$  and  $\xi^{alg}(k, n, N)$  for a specific value of  $n$  and  $N$ . To establish quantitative interpretations of the conditions in Theorems 5, 6 and 7, it is necessary to have quantitative bounds on the behavior of the aRIP constants  $L(k, n, N)$  and  $U(k, n, N)$  for the matrix  $A$  in question, [4, 5]. Currently, there

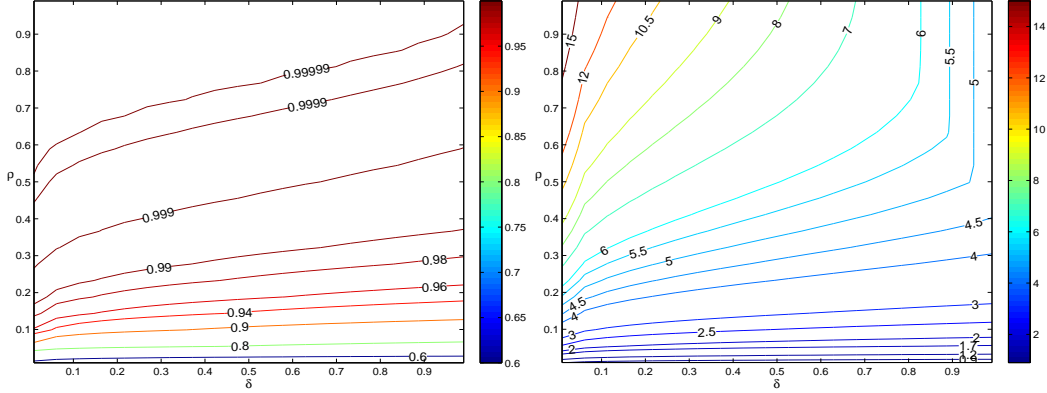


Figure 3: Bounds,  $\mathcal{L}(\delta, \rho)$  and  $\mathcal{U}(\delta, \rho)$  (left and right respectively), above which it is exponentially unlikely that the RIP constants  $L(k, n, N)$  and  $U(k, n, N)$  exceed, with entries in  $A$  drawn i.i.d.  $\mathcal{N}(0, n^{-1})$  and in the limit as  $\frac{k}{n} \rightarrow \rho$  and  $\frac{n}{N} \rightarrow \delta$  as  $n \rightarrow \infty$ , see Theorem 8.

is no known explicit family of matrices  $A$  for which it has been proven that  $U(k, n, N)$  and  $L(k, n, N)$  remain bounded above and away from one, respectively, as  $n$  grows, for  $k$  and  $N$  proportional to  $n$ . However, it is known that for some random matrix ensembles, with overwhelming probability on the draw of  $A$ ,  $\frac{1}{1-L(k, n, N)}$  and  $U(k, n, N)$  do remain bounded as  $n$  grows, for  $k$  and  $N$  proportional to  $n$ . The ensemble with the best known bounds on the growth rates of  $L(k, n, N)$  and  $U(k, n, N)$  in this setting are the Gaussian matrices. In this section, we consider large problem sizes as  $(k, n, N) \rightarrow \infty$ , with  $\frac{n}{N} \rightarrow \delta$  and  $\frac{k}{n} \rightarrow \rho$  for  $\delta, \rho \in (0, 1)$ . We study the implications of the sufficient conditions from Section 2 for matrices with Gaussian i.i.d. entries, namely, entries drawn i.i.d. from the normal distribution with mean 0 and variance  $n^{-1}$ ,  $\mathcal{N}(0, n^{-1})$ .

Gaussian matrices are well studied and much is known about the behavior of their eigenvalues. The first bounds on the RIP constants for Gaussian matrices were presented by Candés and Tao [11]. Recently, the first three authors [5] derived upper bounds on the aRIP constants  $L(k, n, N)$  and  $U(k, n, N)$ , for matrices of size  $n \times N$  with Gaussian i.i.d. entries. Level curves of the bounds  $\mathcal{L}(\delta, \rho)$  and  $\mathcal{U}(\delta, \rho)$  presented in Theorem 8 are shown in Figure 3.

**Theorem 8 (Blanchard, Cartis, and Tanner [5]).** *Let  $A$  be a matrix of size  $n \times N$  whose entries are drawn i.i.d. from  $\mathcal{N}(0, n^{-1})$  and let  $n \rightarrow \infty$  with  $\frac{k}{n} \rightarrow \rho$  and  $\frac{n}{N} \rightarrow \delta$ . Let  $H(p) := p \log(1/p) + (1-p) \log(1/(1-p))$  denote the usual Shannon Entropy with base  $e$  logarithms, and let*

$$\psi_{\min}(\lambda, \rho) := H(\rho) + \frac{1}{2} [(1-\rho) \log \lambda + 1 - \rho + \rho \log \rho - \lambda], \quad (17)$$

$$\psi_{\max}(\lambda, \rho) := \frac{1}{2} [(1+\rho) \log \lambda + 1 + \rho - \rho \log \rho - \lambda]. \quad (18)$$

Define  $\lambda_{\min}(\delta, \rho)$  and  $\lambda_{\max}(\delta, \rho)$  as the solution to (19) and (20), respectively:

$$\delta \psi_{\min}(\lambda_{\min}(\delta, \rho), \rho) + H(\rho \delta) = 0 \quad \text{for} \quad \lambda_{\min}(\delta, \rho) \leq 1 - \rho \quad (19)$$

$$\delta \psi_{\max}(\lambda_{\max}(\delta, \rho), \rho) + H(\rho \delta) = 0 \quad \text{for} \quad \lambda_{\max}(\delta, \rho) \geq 1 + \rho. \quad (20)$$

Define  $\mathcal{L}(\delta, \rho)$  and  $\mathcal{U}(\delta, \rho)$  as

$$\mathcal{L}(\delta, \rho) := 1 - \lambda_{\min}(\delta, \rho) \quad \text{and} \quad \mathcal{U}(\delta, \rho) := \min_{\nu \in [\rho, 1]} \lambda_{\max}(\delta, \nu) - 1. \quad (21)$$

For any  $\epsilon > 0$ , as  $n \rightarrow \infty$ ,

$$\text{Prob}(L(k, n, N) < \mathcal{L}(\delta, \rho) + \epsilon) \rightarrow 1 \quad \text{and} \quad \text{Prob}(U(k, n, N) < \mathcal{U}(\delta, \rho) + \epsilon) \rightarrow 1.$$

With Theorem 8, we are able to formulate quantitative statements about the sufficient aRIP conditions from Section 2 where  $A$  has Gaussian  $\mathcal{N}(0, n^{-1})$  entries. A naive replacement of each  $L(\cdot, n, N)$  and  $U(\cdot, n, N)$  in Theorems 5-7 with the asymptotic aRIP bounds in Theorem 8 is valid in these cases. The properties necessary for this replacement are detailed in Lemma 12, stated in the Appendix. For each algorithm (CoSaMP, SP and IHT) the recovery conditions can be stated in the same format as Theorem 1 and Corollary 2, with only the expressions for  $\kappa(\delta, \rho)$ ,  $\mu(\delta, \rho)$  and  $\xi(\delta, \rho)$  differing. These recovery factors are stated in Theorems 9-11.

**Theorem 9.** *Theorem 1 and Corollary 2 are satisfied for CoSaMP, Algorithm 1, with  $\kappa^{csp}(\delta, \rho) := 1$  and  $\mu^{csp}(\delta, \rho)$  and  $\xi^{csp}(\delta, \rho)$  defined as*

$$\mu^{csp}(\delta, \rho) := \frac{1}{2} \left( 2 + \frac{\mathcal{L}(\delta, 4\rho) + \mathcal{U}(\delta, 4\rho)}{1 - \mathcal{L}(\delta, 3\rho)} \right) \left( \frac{\mathcal{L}(\delta, 2\rho) + \mathcal{U}(\delta, 2\rho) + \mathcal{L}(\delta, 4\rho) + \mathcal{U}(\delta, 4\rho)}{1 - \mathcal{L}(\delta, 2\rho)} \right). \quad (22)$$

and

$$\xi^{csp}(\delta, \rho) := 2 \left\{ \left( 2 + \frac{\mathcal{L}(\delta, 4\rho) + \mathcal{U}(\delta, 4\rho)}{1 - \mathcal{L}(\delta, 3\rho)} \right) \left( \frac{\sqrt{1 + \mathcal{U}(\delta, 2\rho)}}{1 - \mathcal{L}(\delta, 2\rho)} \right) + \frac{1}{\sqrt{1 - \mathcal{L}(\delta, 3\rho)}} \right\}. \quad (23)$$

The phase transition lower bound  $\rho_S^{csp}(\delta)$  is defined as the solution to  $\mu^{csp}(\delta, \rho) = 1$ .  $\rho_S^{csp}(\delta)$  is displayed as the black curve in Figure 1(a).  $\mu^{csp}(\delta, \rho)$  and  $\xi^{csp}(\delta, \rho)/(1 - \mu^{csp}(\delta, \rho))$  are displayed in Figure 2 panels (a) and (b) respectively.

**Theorem 10.** *Theorem 1 and Corollary 2 are satisfied for Subspace Pursuit, Algorithm 2, with  $\kappa^{sp}(\delta, \rho)$ ,  $\mu^{sp}(\delta, \rho)$ , and  $\xi^{sp}(\delta, \rho)$  defined as*

$$\kappa^{sp}(\delta, \rho) := 1 + \frac{\mathcal{U}(\delta, 2\rho)}{1 - \mathcal{L}(\delta, \rho)}, \quad (24)$$

$$\mu^{sp}(\delta, \rho) := \frac{2\mathcal{U}(\delta, 3\rho)}{1 - \mathcal{L}(\delta, \rho)} \left( 1 + \frac{2\mathcal{U}(\delta, 3\rho)}{1 - \mathcal{L}(\delta, 2\rho)} \right) \left( 1 + \frac{\mathcal{U}(\delta, 2\rho)}{1 - \mathcal{L}(\delta, \rho)} \right), \quad (25)$$

and

$$\begin{aligned} \xi^{sp}(\delta, \rho) := & \frac{\sqrt{1 + \mathcal{U}(\delta, \rho)}}{1 - \mathcal{L}(\delta, \rho)} \left[ 1 - \mu^{sp}(\delta, \rho) + 2\kappa^{sp}(\delta, \rho) \left( 1 + \frac{2\mathcal{U}(\delta, 3\rho)}{1 - \mathcal{L}(\delta, 2\rho)} \right) \right] \\ & + \frac{2\kappa^{sp}(\delta, \rho)}{\sqrt{1 - \mathcal{L}(\delta, 2\rho)}}. \end{aligned} \quad (26)$$

The phase transition lower bound  $\rho_S^{sp}(\delta)$  is defined as the solution to  $\mu^{sp}(\delta, \rho) = 1$ .  $\rho_S^{sp}(\delta)$  is displayed as the magenta curve in Figure 1(a).  $\mu^{sp}(\delta, \rho)$  and  $\xi^{sp}(\delta, \rho)/(1 - \mu^{sp}(\delta, \rho))$  are displayed in Figure 2 panels (c) and (d) respectively.

**Theorem 11.** *Theorem 1 and Corollary 2 are satisfied for Iterative Hard Thresholding, Algorithm 3, with  $\omega := 2/(2 + \mathcal{U}(\delta, 3\rho) - \mathcal{L}(\delta, 3\rho))$ ,  $\kappa^{iht}(\delta, \rho) := 1$ , and  $\mu^{iht}(\delta, \rho)$  and  $\xi^{iht}(\delta, \rho)$  defined as*

$$\mu^{iht}(\delta, \rho) := 2\sqrt{2} \left( \frac{\mathcal{L}(\delta, 3\rho) + \mathcal{U}(\delta, 3\rho)}{2 + \mathcal{U}(\delta, 3\rho) - \mathcal{L}(\delta, 3\rho)} \right) \quad (27)$$

and

$$\xi^{iht}(\delta, \rho) := \frac{4\sqrt{1 + \mathcal{U}(\delta, 2\rho)}}{2 + \mathcal{U}(\delta, 3\rho) - \mathcal{L}(\delta, 3\rho)}. \quad (28)$$

The phase transition lower bound  $\rho_S^{iht}(\delta)$  is defined as the solution to  $\mu^{iht}(\delta, \rho) = 1$ .  $\rho_S^{iht}(\delta)$  is displayed as the red curve in Figure 1(a).  $\mu^{iht}(\delta, \rho)$  and  $\xi^{iht}(\delta, \rho)/(1 - \mu^{iht}(\delta, \rho))$  are displayed in Figure 2 panels (e) and (f) respectively.

An analysis similar to that presented here for the greedy algorithms CoSaMP, SP, and IHT was previously carried out in [5] for the  $\ell_1$ -regularization problem (2). For comparison, the associated  $\rho_S^{\ell_1}(\delta)$  implied by aRIP for the Gaussian ensemble is displayed in Figure 1, for details see [5]. The form of the results for  $\ell_1$  differs from those of Theorem 1 and Corollary 2 in that no algorithm is usually specified for how (2) is solved. For this reason, no results are stated for the convergence rate or number of iterations. However, (2) can be reformulated as a convex quadratic or second-order cone programming problem — and its noiseless variant as a linear program — which have polynomial complexity when solved using interior point methods [33]. Moreover, convergence and complexity of other alternative algorithms for solving (2) such as gradient projection have long been studied by the optimization community for more general problems [3, 31, 34], and recently, more specifically for (2) [25, 32] and many more.

#### 4. Discussion and Conclusions

*Summary.* We have presented a framework in which recoverability results for sparse approximation algorithms derived using the ubiquitous RIP can be easily compared. This phase transition framework, [15, 20, 5], translates the generic RIP-based conditions of Theorem 3 into specific sparsity levels  $k$  and problem sizes  $n$  and  $N$  for which the algorithm is guaranteed to satisfy the sufficient RIP conditions with high probability on the draw of the measurement matrix; see Theorem 1. Deriving (bounds on) the phase transitions requires bounds on the behaviour of the measurement matrix' RIP constants [4]. To achieve the most favorable quantitative bounds on the phase transitions, we used the less restrictive aRIP constants; moreover, we employed the best known bounds on aRIP constants, those provided for Gaussian matrices [5], see Theorem 8. This framework was illustrated on three exemplar greedy algorithms: CoSaMP [29], SP [13], and IHT [8]. The lower bounds on the phase transitions in Theorems 9-11 allow for a direct comparison of the current theoretical results/guarantees for these algorithms.

*Computational Cost of CoSaMP, SP and IHT.* The major computational cost per iteration in these algorithms is the application of one or more pseudoinverses. SP uses two pseudoinverses of dimensions  $2k \times n$  and  $k \times n$  per iteration and another to compute the output vector  $\hat{x}$ ; see Algorithm 2. CoSaMP uses only one pseudoinverse per iteration but of dimensions  $3k \times n$ ; see Algorithm 1. Consequently, CoSaMP and SP require leading order  $10nk^2$  and  $18nk^2$  floating point operations per iteration, respectively, if the pseudoinverse is solved using an exact  $QR$  factorization. IHT avoids computing a pseudoinverse altogether in internal iterations, but is aided by one pseudoinverse of dimensions  $k \times n$

on the final support set. Thus IHT has a substantially lower computational cost per iteration than CoSaMP and SP. Note that pseudoinverses may be computed approximately by an iterative method such as conjugate gradients [29]. As such, the exact application of a pseudoinverse could be entirely avoided, improving the implementation costs of these algorithms, especially of CoSaMP and SP.

Globally, all three algorithms converge linearly; in fact, they converge in a finite number of iterations provided there exists a  $k$ -sparse solution to  $Ax = y$  and a sufficient aRIP condition is satisfied, see Corollary 2. For each algorithm, the upper bound on the required number of iterations grows unbounded as the function  $\mu^{alg}(k, n, N) \rightarrow 1$ . Hence, according to the bounds presented here, to ensure rapid convergence, it is advantageous to have a matrix that satisfies a more strict condition, such as  $\mu^{alg}(k, n, N) < \frac{1}{2}$ . Similarly, the factor controlling stability to additive noise, namely the vector  $e$  in Theorem 1, blows up as the function  $\mu^{alg}(k, n, N) \rightarrow 1$ . Again, according to the bounds presented here, in order to guarantee stability with small amplification of the additive noise, it is necessary to restrict the range of  $\frac{\xi^{alg}}{1-\mu^{alg}}(k, n, N)$ . A phase transition function analogous to the functions  $\rho_S^{alg}(\delta)$  can be easily read from Figure 2 in these settings as well, resulting in curves lower than those presented in Figure 1(a). For example, requiring a multiplicative noise penalty of no more than 8 significantly reduces the region of the phase space below the phase transition curve for all three algorithms; see Figure 2(b,d,f). This is the standard trade-off of compressed sensing, where one must determine the appropriate balance between computational efficiency, stability, and minimizing the number of measurements.

*Comparison of Phase Transitions and Constants of Proportionality.* From Figure 1(a), we see that for uniform guarantees over all  $k$ -sparse  $x$ , the best known lower bounds on the phase transitions for the three greedy algorithms satisfy the ordering  $\rho_S^{csp}(\delta) < \rho_S^{sp}(\delta) < \rho_S^{iht}(\delta)$  for Gaussian measurement matrices. Moreover, the level curves bounding the stability and convergence factors follow the same ordering for every bound, see Figure 2. Therefore, we now know that, at least for Gaussian matrices, according to existing theory, IHT has the largest region where recovery for all signals can be guaranteed; the regions with similar guarantees for SP and CoSaMP are considerably smaller. Moreover, IHT has a lower bound on its computational cost and better known stability bounds for additive noise.

The phase transition bounds  $\rho_S^{alg}(\delta)$  also allow a precise comparison of the recoverability results derived for these greedy algorithms with those proven for  $\ell_1$ -regularization using the aRIP, see Figure 1. Although [29, 13, 8] have provided guarantees of successful sparse recovery analogous to those for  $\ell_1$ -regularization, the greedy algorithms place a more restrictive aRIP condition on the suitable matrices to be used in the algorithm. However, some of the algorithms for solving the  $\ell_1$ -regularization problem, such as interior point methods, are, in general, computationally more expensive than the greedy methods discussed in this paper, and hence attention needs to be paid to the method of choice for solving the  $\ell_1$ -regularization problem [2, 25].

The lower bounds on the phase transitions presented here can also be read as lower bounds on the constant of proportionality in the oversampling rate, namely, taking  $n \geq (\rho_S^{alg}(\delta))^{-1}k$  measurements rather than the oracle rate of  $k$  measurements is sufficient if algorithm “alg” is used to recover the  $k$ -sparse signal. From Figure 1(b), it is clear that according to the conditions presented here, the convergence of greedy algorithms can only be guaranteed with substantially more measurements than for  $\ell_1$ -regularization. The lowest possible number of measurements (when  $n = N$  so  $\delta = 1$ ) for the algorithms are as follows:  $n \geq 907k$  for IHT,  $n \geq 3124k$  for SP, and  $n \geq 4923k$  for CoSaMP. On the other hand, an aRIP analysis of  $\ell_1$ -regularization yields that linear programming requires  $n \geq 317k$ . In fact, using a geometric, convex polytopes approach, Donoho has shown that for  $\ell_1$ -regularization,  $n \geq 5.9k$  is a sufficient number of measurements [5, 15, 17] when the target signal,  $x$ , is exactly

$k$ -sparse, and the multiple 5.9 increases smoothly as noise is added [38].

*Future Improvements and Conclusions.* The above bounds on greedy algorithms’ phase transitions could be improved by further refining the algorithms’ theory, for instance, deriving less strict aRIP conditions on the measurement matrix that still ensure convergence of the algorithm; numerous such advances have occurred for  $\ell_1$ -regularization and can be expected to also take place for greedy algorithms. The phase transition framework presented here can be directly applied to such advances, and the resulting lower bounds on the phase transitions can serve the crucial role of an unbiased measure of improvement; for a further discussion of the importance of the latter see [7].

Alternatively, increasing the lower bounds on the phase transitions could be expected to occur from improving the upper bounds we employed on the aRIP constants of the Gaussian measurement matrices, see Theorem 8. However, extensive empirical calculations of lower estimates of aRIP constants show the latter to be within a factor of 1.83 of our proven upper bounds [5], and are in fact much sharper for the range of  $\rho \ll 1$  relevant here. To test the effect improved aRIP bounds have on the phase transitions, the upper bounds on the aRIP constants used in  $\mu^{alg}(\delta, \rho)$  could be replaced with empirically observed lower bounds on the aRIP constants, yielding upper bounds as to how much the phase transitions could be raised by improving the aRIP bounds alone. This was done in [5] for  $\ell_1$ -regularization, showing that improving the aRIP constants alone cannot increase its phase transition by more than a multiple of 2.5. Similar limited improvement can also be expected for the greedy algorithms discussed here, but accurate testing of this is difficult at present since the small values of their  $\rho_S^{alg}(\delta)$  require testing of aRIP constants for matrices whose size is too large for existing algorithms; for instance, testing CoSaMP for sparsity even as small as  $k = 5$  requires matrices with approximately  $10^9$  entries. During the revision of this manuscript, improved bounds on the aRIP constants for the Gaussian ensemble were derived [1], tightening the bound to be within 1.57 of lower estimates. However, for the relevant range of  $\rho$  here,  $\rho \approx 10^{-3}$ , both bounds were already very sharp [1], and the resulting increase of the phase transitions shown here was under 0.5%.

## A. Proofs of Main Results

We present a framework by which RIP-based convergence results of the form presented in Theorem 3 can be translated into results of the form of Theorem 1; that is removing explicit dependencies on RIP constants in favor of their bounds for specified classes of matrices.

The proofs of Theorems 5, 6, and 7 follow their original derivations in [29], [13], and [8] with greater care taken to obtain the tightest bounds possible using aRIP; their derivation is available in an extended preprint [6].

Theorems 9, 10, and 11 follow from Theorems 5, 6, and 7 and the form of  $\mu^{alg}$  and  $\xi^{alg}$  as functions of  $\mathcal{L}$  and  $\mathcal{U}$ ; this latter point is summarized in Lemma 12 which is stated and proven in Section A.1. The resulting Theorems 9, 10, and 11 can then be interpreted in the phase transition framework advocated by Donoho et al. [15, 17, 19, 20, 23], as we have explained in Section 4.

### A.1. Technical Lemmas

Theorems 9, 10, and 11 follow from Theorems 5, 6, and 7 and the form of  $\mu^{alg}$  and  $\xi^{alg}$  as functions of  $\mathcal{L}$  and  $\mathcal{U}$ . We formalize the relevant functional dependencies in the next three lemmas.

**Lemma 12.** For some  $\tau < 1$ , define the set  $\mathcal{Z} := (0, \tau)^p \times (0, \infty)^q$  and let  $F : \mathcal{Z} \rightarrow \mathbb{R}$  be continuously differentiable on  $\mathcal{Z}$ . Let  $A$  be a Gaussian matrix of size  $n \times N$  with aRIP constants  $L(\cdot, n, N), U(\cdot, n, N)$  and let  $\mathcal{L}(\delta, \cdot), \mathcal{U}(\delta, \cdot)$  be defined as in Theorem 8. Define  $\mathbf{1}$  to be the vector of all ones, and

$$z(k, n, N) := [L(k, n, N), \dots, L(pk, n, N), U(k, n, N), \dots, U(qk, n, N)] \quad (29)$$

$$z(\delta, \rho) := [\mathcal{L}(\delta, \rho), \dots, \mathcal{L}(\delta, p\rho), \mathcal{U}(\delta, \rho), \dots, \mathcal{U}(\delta, q\rho)]. \quad (30)$$

(i) Suppose, for all  $t \in \mathcal{Z}$ ,  $(\nabla F[t])_i \geq 0$  for all  $i = 1, \dots, p + q$  and for any  $v \in \mathcal{Z}$  we have  $\nabla F[t] \cdot v > 0$ . Then for any  $c\epsilon > 0$ , as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta, \frac{k}{n} \rightarrow \rho$ , there is overwhelming probability on the draw of the matrix  $A$  that

$$\text{Prob}(F[z(k, n, N)] < F[z(\delta, \rho) + 1c\epsilon]) \rightarrow 1 \quad \text{as } n \rightarrow \infty. \quad (31)$$

(ii) Suppose, for all  $t \in \mathcal{Z}$ ,  $(\nabla F[t])_i \geq 0$  for all  $i = 1, \dots, p + q$  and there exists  $j \in \{1, \dots, p\}$  such that  $(\nabla F[t])_j > 0$ . Then there exists  $c \in (0, 1)$  depending only on  $F, \delta,$  and  $\rho$  such that for any  $\epsilon \in (0, 1)$

$$F[z(\delta, \rho) + 1c\epsilon] < F[z(\delta, (1 + \epsilon)\rho)], \quad (32)$$

and so there is overwhelming probability on the draw of  $A$  that

$$\text{Prob}(F[z(k, n, N)] < F[z(\delta, (1 + \epsilon)\rho])) \rightarrow 1 \quad \text{as } n \rightarrow \infty. \quad (33)$$

Also,  $F(z(\delta, \rho))$  is strictly increasing in  $\rho$ .

PROOF. To prove (i), suppose  $u, v \in \mathcal{Z}$  with  $v_i > u_i$  for all  $i = 1, \dots, p + q$ . From Taylor's Theorem,  $F[v] = F[u + (v - u)] = F[u] + \nabla F[t] \cdot [v - u]$  with  $t = u + \lambda[v - u]$  for some  $\lambda \in (0, 1)$ . Then

$$F[v] > F[u] \quad (34)$$

since, by assumption,  $\nabla F[t] \cdot [v - u] > 0$ .

From Theorem 8, for any  $c\epsilon > 0$  and any  $i = 1, \dots, p + q$ , as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta, \frac{k}{n} \rightarrow \rho$ ,

$$\text{Prob}(z(k, n, N)_i < z(\delta, \rho)_i + c\epsilon) \rightarrow 1,$$

with convergence to 1 exponential in  $n$ . Therefore, letting  $v_i := z(\delta, \rho)_i + c\epsilon$  and  $u_i := z(k, n, N)_i$ , for all  $i = 1, \dots, p + q$ , we conclude from (34) that

$$\text{Prob}(F[z(k, n, N)] < F[z(\delta, \rho) + 1c\epsilon]) \rightarrow 1,$$

again with convergence to 1 exponential in  $n$ .

To establish (ii), we take the Taylor expansion of  $F$  centered at  $z(\delta, \rho)$ , namely

$$F[z(\delta, \rho) + 1c\epsilon] = F[z(\delta, \rho)] + \nabla F[t_1] \cdot 1c\epsilon \quad \text{for } t_1 \in (z(\delta, \rho), z(\delta, \rho) + 1c\epsilon) \quad (35)$$

$$F[z(\delta, (1 + \epsilon)\rho)] = F[z(\delta, \rho)] + \left( \nabla F[z(\delta, \rho)] \cdot \frac{\partial}{\partial \rho} z(\delta, \rho) \right) \Big|_{\rho=t_2} \epsilon \rho \quad \text{for } t_2 \in (\rho, (1 + \epsilon)\rho). \quad (36)$$

Select

$$t_1^* = \operatorname{argmax} \{ \nabla F[t_1] : t_1 \in [z(\delta, \rho), z(\delta, \rho) + 1] \}$$

$$t_2^* = \operatorname{argmin} \left\{ \left( \nabla F[z(\delta, \rho)] \cdot \frac{\partial}{\partial \rho} z(\delta, \rho) \right) \Big|_{\rho=t_2} : t_2 \in [\rho, (1 + \epsilon)\rho] \right\}$$

so that

$$F[z(\delta, \rho) + 1c\epsilon] \leq F[z(\delta, \rho)] + \nabla F[t_1^*] \cdot 1c\epsilon \quad (37)$$

$$F[z(\delta, (1 + \epsilon)\rho)] \geq F[z(\delta, \rho)] + \left( \nabla F[z(\delta, \rho)] \cdot \frac{\partial}{\partial \rho} z(\delta, \rho) \right) \Big|_{\rho=t_2^*} \epsilon\rho. \quad (38)$$

Since  $\mathcal{L}(\delta, \rho)$  is strictly increasing in  $\rho$  [5], then  $\left( \frac{\partial}{\partial \rho} z(\delta, \rho) \Big|_{\rho=t_2^*} \right)_j > 0$  for all  $j = 1, \dots, p$ . Since  $\mathcal{U}(\delta, \rho)$  is nondecreasing in  $\rho$  [5], then  $\left( \frac{\partial}{\partial \rho} z(\delta, \rho) \Big|_{\rho=t_2^*} \right)_i \geq 0$  for all  $i = p + 1, \dots, p + q$ . Hence, by the hypotheses of (ii),

$$\begin{aligned} \left( \nabla F[z(\delta, \rho)] \cdot \frac{\partial}{\partial \rho} z(\delta, \rho) \right) \Big|_{\rho=t_2^*} &> 0 \\ \nabla F[t_1^*] \cdot 1 &> 0. \end{aligned}$$

Therefore, for any  $c$  satisfying

$$0 < c < \min \left\{ 1, \rho \frac{\left( \nabla F[z(\delta, \rho)] \cdot \frac{\partial}{\partial \rho} z(\delta, \rho) \right) \Big|_{\rho=t_2^*}}{\nabla F[t_1^*] \cdot 1} \right\},$$

(37) and (38) imply (32). Since the hypotheses of (ii) imply those of (i), (31) also holds, and so (33) follows.  $F(z(\delta, \rho))$  strictly increasing follows from the hypotheses of (ii) and  $\mathcal{L}(\delta, \rho)$  and  $\mathcal{U}(\delta, \rho)$  strictly increasing and nondecreasing in  $\rho$ , respectively [5]. ■

Let the superscript *alg* denote the algorithm identifier so that  $\mu^{alg}(k, n, N)$  is defined by one of (10), (13), (15), while  $\mu^{alg}(\delta, \rho)$  is defined by one of (22), (25), (27). Next, a simple property is summarized in Lemma 13, that further reveals some necessary ingredients of our analysis.

**Lemma 13.** *Assume that  $\mu^{alg}(\delta, \rho)$  is strictly increasing in  $\rho$  and let  $\rho_S^{alg}(\delta)$  solve  $\mu^{alg}(\delta, \rho) = 1$ . For any  $\epsilon \in (0, 1)$ , if  $\rho < (1 - \epsilon)\rho_S^{alg}(\delta)$ , then  $\mu^{alg}(\delta, (1 + \epsilon)\rho) < 1$ .*

PROOF. Let  $\rho_\epsilon^{alg}(\delta)$  be the solution to  $\mu^{alg}(\delta, (1 + \epsilon)\rho) = 1$ . Since by definition,  $\rho_S^{alg}(\delta)$  denotes a solution to  $\mu^{alg}(\delta, \rho) = 1$ , and this solution is unique as  $\mu^{alg}(\delta, \rho)$  is strictly increasing, we must have  $(1 + \epsilon)\rho_\epsilon^{alg}(\delta) = \rho_S^{alg}(\delta)$ . Since  $(1 - \epsilon) < (1 + \epsilon)^{-1}$  for all  $\epsilon \in (0, 1)$ , we have  $(1 - \epsilon)\rho_S^{alg}(\delta) < \rho_\epsilon^{alg}(\delta)$ . If  $\rho < (1 - \epsilon)\rho_S^{alg}(\delta)$ , then since  $\mu^{alg}(\delta, \rho)$  is strictly increasing in  $\rho$ ,

$$\mu^{alg}(\delta, (1 + \epsilon)\rho) < \mu^{alg}(\delta, (1 + \epsilon)(1 - \epsilon)\rho_S^{alg}(\delta)) < \mu^{alg}(\delta, (1 + \epsilon)\rho_\epsilon^{alg}(\delta)) = 1.$$

■

Note that Lemma 12 ii) with  $F := \mu^{alg}$  will be employed to show the first assumption in Lemma 13; this is but one of several good uses of Lemma 12 that we will make.

Corollaries 2 and 4 are easily derived from Lemma 14. Note that this lemma demonstrates only that the support set has been recovered. The proof of Lemma 14 is a minor generalization of a proof from [13, Theorem 7].

**Lemma 14.** *Suppose, after  $l$  iterations, algorithm  $alg$  returns the  $k$ -sparse approximation  $\hat{x}^l$  to a  $k$ -sparse target signal  $x$ . Suppose there exist constants  $\mu$  and  $\kappa$  independent of  $l$  and  $x$  such that*

$$\|x - \hat{x}^l\|_2 \leq \kappa \mu^l \|x\|_2. \quad (39)$$

*If  $\mu < 1$ , then the support set of  $\hat{x}^l$  coincides with the support set of  $x$  after at most  $\ell_{max}^{alg}(x)$  iterations, where*

$$\ell_{max}^{alg}(x) := \left\lceil \frac{\log \nu_{min}(x) - \log \kappa}{\log \mu} \right\rceil + 1, \quad (40)$$

*where  $\nu_{min}(x)$  is defined in (5).*

PROOF. Let  $T$  be the support set of  $x$  and  $T^l$  be the support set of  $\hat{x}^l$ ; as  $x, \hat{x}^l \in \chi^N(k)$ ,  $|T|, |T^l| \leq k$ . From the definition (40) of  $\ell_{max}^{alg}(x)$  and (5),  $\kappa \mu^{\ell_{max}^{alg}(x)} \|x\|_2 < \min_{i \in T} |x_i|$ . From (39), we then have

$$\|x - \hat{x}^{\ell_{max}^{alg}(x)}\|_2 \leq \kappa \mu^{\ell_{max}^{alg}(x)} \|x\|_2 < \min_{i \in T} |x_i|$$

which clearly implies that  $T \subset T^{\ell_{max}^{alg}(x)}$ . Since  $|T| = |T^{\ell_{max}^{alg}(x)}|$ , the sets must be equal. ■

To ensure exact recovery of the target signal, namely, to complete the proof of Corollaries 2 and 4, we actually need something stronger than recovering the support set as implied by Lemma 14. For CoSaMP and SP, since the algorithms employ a pseudoinverse at an appropriate step, the output is then the exact sparse signal. For IHT, no pseudoinverse has been applied; thus, to recover the signal exactly, one simply determines  $T$  from the output vector and then  $x = A_T^\dagger y$ . These comments and Lemma 14 now establish Corollaries 2 and 4 for each algorithm.

In each of the following subsections, we apply the above lemmas to derive Theorems 9, 10, and 11.

## A.2. Proofs for CoSaMP, Theorem 9

Let  $x, y, A$  and  $e$  satisfy the hypothesis of Theorem 9 and select  $\epsilon > 0$ . Fix  $\tau < 1$  and let

$$\begin{aligned} z(k, n, N) &= [L(2k, n, N), L(3k, n, N), L(4k, n, N), U(2k, n, N), U(4k, n, N)] \\ \text{and } z(\delta, \rho) &= [\mathcal{L}(\delta, 2\rho), \mathcal{L}(\delta, 3\rho), \mathcal{L}(\delta, 4\rho), \mathcal{U}(\delta, 2\rho), \mathcal{U}(\delta, 4\rho)]. \end{aligned}$$

Define  $\mathcal{Z} = (0, \tau)^3 \times (0, \infty)^2$  and define the functions  $F^{csp}, G^{csp} : \mathcal{Z} \rightarrow \mathbb{R}$ :

$$F^{csp}[z] := F^{csp}[z_1, \dots, z_5] = 2 \left( 2 + \frac{z_3 + z_5}{1 - z_2} \right) \left( \frac{z_1 + z_4 + z_3 + z_5}{1 - z_1} \right). \quad (41)$$

$$G^{csp}[z] := G^{csp}[z_1, \dots, z_5] = 2 \left\{ \left( 2 + \frac{z_3 + z_5}{1 - z_2} \right) \left( \frac{\sqrt{1 + z_4}}{1 - z_1} \right) + \frac{1}{\sqrt{1 - z_2}} \right\}. \quad (42)$$

Clearly,  $(\nabla F^{csp}[t])_i \geq 0$  for all  $i = 1, \dots, 5$  and

$$(\nabla F^{csp}[t])_1 = \frac{1}{2} \left( 2 + \frac{t_3 + t_5}{1 - t_2} \right) \left( \frac{1 + t_4 + t_3 + t_5}{(1 - t_1)^2} \right) > 0.$$

Hence the hypotheses of Lemma 12 (ii) are satisfied for  $F^{csp}$ . By (10), (22) and (41),  $F^{csp}[z(k, n, N)] = \mu^{csp}(k, n, N)$  and  $F^{csp}[z(\delta, \rho)] = \mu^{csp}(\delta, \rho)$ . Thus, by Lemma 12, as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta, \frac{k}{n} \rightarrow \rho$ ,

$$\text{Prob}(\mu^{csp}(k, n, N) < \mu^{csp}(\delta, (1 + \epsilon)\rho)) \rightarrow 1. \quad (43)$$

Also,  $\mu^{csp}(\delta, \rho)$  is strictly increasing in  $\rho$  and so Lemma 13 applies.

Similarly,  $G^{csp}$  satisfies the hypotheses of Lemma 12 (ii). Likewise, by (11), (23) and (42),  $G^{csp}[z(k, n, N)] = \xi^{csp}(k, n, N)$  and  $G^{csp}[z(\delta, \rho)] = \xi^{csp}(\delta, \rho)$ . Again, by Lemma 12, as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta$ ,  $\frac{k}{n} \rightarrow \rho$ ,

$$\text{Prob}(\xi^{csp}(k, n, N) < \xi^{csp}(\delta, (1 + \epsilon)\rho)) \rightarrow 1. \quad (44)$$

Therefore, for any  $x \in \chi^N(k)$  and any noise vector  $e$ , as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta$ ,  $\frac{k}{n} \rightarrow \rho$ , there is overwhelming probability on the draw of a matrix  $A$  with Gaussian i.i.d. entries that

$$[\mu^{csp}(k, n, N)]^l \|x\|_2 + \frac{\xi^{csp}(k, n, N)}{1 - \mu^{csp}(k, n, N)} \|e\|_2 \leq [\mu^{csp}(\delta, (1 + \epsilon)\rho)]^l \|x\|_2 + \frac{\xi^{csp}(\delta, (1 + \epsilon)\rho)}{1 - \mu^{csp}(\delta, (1 + \epsilon)\rho)} \|e\|_2. \quad (45)$$

Combining (45) with Theorem 5 completes the argument. ■

### A.3. Proofs for Subspace Pursuit, Theorem 10

Let  $x, y, A$ , and  $e$  satisfy the hypothesis of Theorem 10 and select  $\epsilon > 0$ . Fix  $\tau < 1$  and let

$$z(k, n, N) = [L(k, n, N), L(2k, n, N), U(k, n, N), U(2k, n, N), U(3k, n, N)]$$

and  $z(\delta, \rho) = [\mathcal{L}(\delta, \rho), \mathcal{L}(\delta, 2\rho), \mathcal{U}(\delta, \rho), \mathcal{U}(\delta, 2\rho), \mathcal{U}(\delta, 3\rho)]$ .

Define  $\mathcal{Z} = (0, \tau)^2 \times (0, \infty)^3$  and define the following functions mapping  $\mathcal{Z} \rightarrow \mathbb{R}$ :

$$F^{sp}[z] := F^{sp}[z_1, \dots, z_5] = 2 \frac{z_5}{1 - z_1} \left(1 + \frac{2z_5}{1 - z_2}\right) \left(1 + \frac{z_4}{1 - z_1}\right), \quad (46)$$

$$K[z] := K[z_1, \dots, z_5] = 1 + \frac{z_4}{1 - z_1}, \quad (47)$$

$$G^{sp}[z] := G^{sp}[z_1, \dots, z_5] = 2 \frac{\sqrt{1 + z_3}}{1 - z_1} \left(1 + \frac{2z_5}{1 - z_2}\right) + \frac{2}{\sqrt{1 - z_2}}, \quad (48)$$

$$H[z] := H[z_1, \dots, z_5] = \frac{\sqrt{1 + z_3}}{1 - z_1}. \quad (49)$$

For each of these functions, the gradient is clearly nonnegative componentwise on  $\mathcal{Z}$ , with the first entry of each gradient strictly positive which is sufficient to verify the hypotheses of Lemma 12 (ii). Moreover, from (12)–(14) and (24)–(26), we have

$$\begin{aligned} \kappa^{sp}(k, n, N) \mu^{sp}(k, n, N) &= K[z(k, n, N)] F^{sp}[z(k, n, N)], \\ \kappa^{sp}(\delta, \rho) \mu^{sp}(\delta, \rho) &= K[z(\delta, \rho)] F^{sp}[z(\delta, \rho)], \\ \frac{\xi^{sp}(k, n, N)}{1 - \mu^{sp}(k, n, N)} &= K[z(k, n, N)] \frac{G^{sp}[z(k, n, N)]}{1 - F^{sp}[z(k, n, N)]} + H[z(k, n, N)], \\ \frac{\xi^{sp}(\delta, \rho)}{1 - \mu^{sp}(\delta, \rho)} &= K[z(\delta, \rho)] \frac{G^{sp}[z(\delta, \rho)]}{1 - F^{sp}[z(\delta, \rho)]} + H[z(\delta, \rho)]. \end{aligned}$$

Invoking Lemma 12 for each of the functions in (46)–(49) yields that with high probability on the draw of  $A$  from a Gaussian distribution,

$$\kappa^{sp}(k, n, N) [\mu^{sp}(k, n, N)]^l \|x\|_2 < \kappa^{sp}(\delta, (1 + \epsilon)\rho) [\mu^{sp}(\delta, (1 + \epsilon)\rho)]^l \|x\|_2, \quad (50)$$

$$\frac{\xi^{sp}(k, n, N)}{1 - \mu^{sp}(k, n, N)} \|e\|_2 < \frac{\xi^{sp}(\delta, (1 + \epsilon)\rho)}{1 - \mu^{sp}(\delta, (1 + \epsilon)\rho)} \|e\|_2. \quad (51)$$

Combining (50) and (51) with Theorem 6 completes the argument, recalling that Lemma 12 applied to  $F^{sp} = \mu^{sp}$  also implies that  $\mu^{sp}(\delta, \rho)$  is strictly increasing in  $\rho$  and so Lemma 13 holds. ■

A.4. Proofs for Iterative Hard Thresholding, Theorem 11

Let  $x, y, A$  and  $e$  satisfy the hypothesis of Theorem 11 and select  $\epsilon > 0$ . Fix  $\tau < 1$  and let

$$z(k, n, N) = [L(3k, n, N), U(2k, n, N), U(3k, n, N)]$$

$$\text{and } z(\delta, \rho) = [\mathcal{L}(\delta, 3\rho), \mathcal{U}(\delta, 2\rho), \mathcal{U}(\delta, 3\rho)].$$

Define  $\mathcal{Z} = (0, \tau) \times (0, \infty)^2$ . For an arbitrary weight  $\omega \in (0, 1)$ , define the functions  $F_\omega^{iht}, G_\omega^{iht} : \mathcal{Z} \rightarrow \mathbb{R}$ :

$$F_\omega^{iht}[z] := F_\omega^{iht}[z_1, z_2, z_3] = 2\sqrt{2} \max\{\omega[1 + z_3] - 1, 1 - \omega[1 - z_1]\}, \quad (52)$$

$$G_\omega^{iht}[z] := G_\omega^{iht}[z_1, z_2, z_3] = \frac{\omega}{\sqrt{2}} \left( \frac{\sqrt{1 + z_2}}{1 - \max\{\omega[1 + z_3] - 1, 1 - \omega[1 - z_1]\}} \right). \quad (53)$$

[Note that  $F_\omega^{iht}[z(k, n, N)] = \mu^{iht}(k, n, N)$  and  $G_\omega^{iht}[z(k, n, N)] = \xi^{iht}(k, n, N)/(1 - \mu^{iht}(k, n, N))$  due to (15) and (16).] Clearly the functions are nondecreasing so that, with any  $t \in \mathcal{Z}$ ,  $(\nabla F_\omega^{iht}[t])_i \geq 0$  and  $(\nabla G_\omega^{iht}[t])_i \geq 0$  for  $i = 1, 2, 3$ ; note that  $F_\omega^{iht}[t]$  and  $G_\omega^{iht}[t]$  have points of nondifferentiability, but that the left and right derivatives at those points remain nonnegative. Also, and for any  $v \in \mathcal{Z}$ , since  $t_i, v_i > 0$  for each  $i$ ,  $\nabla F_\omega^{iht}[t] \cdot v > 0$  and  $\nabla G_\omega^{iht}[t] \cdot v > 0$  as both functions clearly increase when each component of the argument increases. Hence,  $F_\omega^{iht}$  and  $G_\omega^{iht}$  satisfy the hypotheses of Lemma 12 (i). Therefore, for any  $\omega \in (0, 1)$ , as  $(k, n, N) \rightarrow \infty$  with  $\frac{n}{N} \rightarrow \delta$ ,  $\frac{k}{n} \rightarrow \rho$ ,

$$\text{Prob} \left( F_\omega^{iht}[z(k, n, N)] < F_\omega^{iht}[z(\delta, \rho) + 1c\epsilon] \right) \rightarrow 1, \quad (54)$$

$$\text{Prob} \left( G_\omega^{iht}[z(k, n, N)] < G_\omega^{iht}[z(\delta, \rho) + 1c\epsilon] \right) \rightarrow 1. \quad (55)$$

Now fix  $\omega^* := \frac{2}{2 + \mathcal{U}(\delta, 3\rho) - \mathcal{L}(\delta, 3\rho)}$  and define

$$\tilde{F}_{\omega^*}^{iht}[z] := \tilde{F}_{\omega^*}^{iht}[z_1, z_2, z_3] = 2\sqrt{2} \left( \frac{z_1 + z_3}{2 + z_3 - z_1} \right), \quad (56)$$

$$\tilde{G}_{\omega^*}^{iht}[z] := \tilde{G}_{\omega^*}^{iht}[z_1, z_2, z_3] = \frac{4\sqrt{1 + z_2}}{2 - (2\sqrt{2} - 1)z_3 - (2\sqrt{2} + 1)z_1}. \quad (57)$$

Then for any  $t \in \mathcal{Z}$ ,  $(\nabla \tilde{F}_{\omega^*}^{iht}[t])_i > 0$  for  $i = 1, 3$  and  $(\nabla \tilde{F}_{\omega^*}^{iht}[t])_2 = 0$ . Likewise,  $(\nabla \tilde{G}_{\omega^*}^{iht}[t])_i > 0$  for  $i = 1, 2, 3$ . Thus  $\tilde{F}_{\omega^*}^{iht}$  and  $\tilde{G}_{\omega^*}^{iht}$  satisfy the hypotheses of Lemma 12 (ii) and, therefore,

$$\tilde{F}_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon] < \tilde{F}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)], \quad (58)$$

$$\tilde{G}_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon] < \tilde{G}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)]. \quad (59)$$

Finally, observe that

$$F_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon] = \tilde{F}_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon], \quad (60)$$

$$G_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon] = \tilde{G}_{\omega^*}^{iht}[z(\delta, \rho) + 1c\epsilon]. \quad (61)$$

In (54) and (55), the weight was arbitrary; thus both statements certainly hold for the particular weight  $\omega^*$ . Therefore, combining (54), (58), (60) and combining (55), (59), (61) imply that with overwhelming probability on the draw of  $A$ ,

$$F_{\omega^*}^{iht}[z(k, n, N)] < \tilde{F}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)], \quad (62)$$

$$G_{\omega^*}^{iht}[z(k, n, N)] < \tilde{G}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)]. \quad (63)$$

Therefore, with the weight  $\omega^*$ , there is overwhelming probability on the draw of  $A$  from a Gaussian distribution that

$$\mu^{iht}(k, n, N) = F_{\omega^*}^{iht}[z(k, n, N)] < \tilde{F}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)] = \mu^{iht}(\delta, (1 + \epsilon)\rho), \quad (64)$$

$$\frac{\xi^{iht}(k, n, N)}{1 - \mu^{iht}(k, n, N)} = G_{\omega^*}^{iht}[z(k, n, N)] < \tilde{G}_{\omega^*}^{iht}[z(\delta, (1 + \epsilon)\rho)] = \frac{\xi^{iht}(\delta, (1 + \epsilon)\rho)}{1 - \mu^{iht}(\delta, (1 + \epsilon)\rho)}, \quad (65)$$

where we also employed (15), (16) with  $\omega = \omega^*$ , and (27), (28). The result follows by invoking Theorem 7 and applying (64) and (65); recall also that Lemma 13 holds since  $\mu^{iht}(\delta, \rho) = \tilde{F}_{\omega^*}^{iht}(z(\delta, \rho))$  is implied to be strictly increasing in  $\rho$  by Lemma 12 (ii). ■

## References

- [1] B. Bah and J. Tanner. Improved bounds on restricted isometry constants for gaussian matrices. submitted, 2010.
- [2] Ewout van den Berg and Michael P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.
- [3] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999. Second edition.
- [4] Jeffrey D. Blanchard, Coralia Cartis, and Jared Tanner. Decay properties for restricted isometry constants. *IEEE Signal Proc. Letters*, 16(7):572–575, 2009.
- [5] Jeffrey D. Blanchard, Coralia Cartis, and Jared Tanner. Compressed Sensing: How sharp is the restricted isometry property? *SIAM Review*, in press.
- [6] Jeffrey D. Blanchard, Coralia Cartis, Jared Tanner, and Andrew Thompson. Phase transitions for greedy sparse approximation algorithms. arXiv, 2009.
- [7] Jeffrey D. Blanchard and Andrew Thompson. On support sizes of restricted isometry constants. *Appl. Comput. Harmon. Anal.*, in press.
- [8] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, 27(3):265–274, 2009.
- [9] A. M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.
- [10] Emmanuel J. Candès. Compressive sampling. In *International Congress of Mathematicians. Vol. III*, pages 1433–1452. Eur. Math. Soc., Zürich, 2006.
- [11] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.
- [12] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Instance optimal decoding by thresholding in compressed sensing. Technical Report, 2008.
- [13] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inform. Theory*, 55(5):2230–2249, 2009.

- [14] M. A. Davenport and M. B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. submitted, 2009.
- [15] David L. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. Technical Report, Department of Statistics, Stanford University, 2004.
- [16] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [17] David L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.*, 35(4):617–652, 2006.
- [18] David L. Donoho and Arian Malehi. Optimally tuned iterative thresholding algorithms for compressed sensing. *IEEE Selected Topics in Signal Processing*, 4(2):330–341, 2010.
- [19] David L. Donoho and Victoria Stodden. Breakdown point of model selection when the number of variables exceeds the number of observations. In *Proceedings of the International Joint Conference on Neural Networks*, 2006.
- [20] David L. Donoho and Jared Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA*, 102(27):9446–9451, 2005.
- [21] David L. Donoho and Jared Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J. AMS*, 22(1):1–53, 2009.
- [22] David L. Donoho and Jared Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, in press.
- [23] David L. Donoho and Yaakov Tsaig. Fast solution of  $l_1$  minimization problems when the solution may be sparse. *IEEE Trans. Inform. Theory*, 54(11):4789–4812, 2008.
- [24] David L. Donoho, Yaakov Tsaig, Iddo Drori, and Jean-Luc Stark. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, submitted.
- [25] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Selected Topics in Signal Processing*, 1(4):586–597, 2007.
- [26] S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via  $l_q$ -minimization for  $0 < q \leq 1$ . *Appl. Comput. Harmon. Anal.*, 26(3):395–407, 2009.
- [27] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 337–344, New York, NY, USA, 2009. ACM.
- [28] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.
- [29] Deanna Needell and Joel Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comp. Harm. Anal.*, 26(3):301–321, 2009.

- [30] Deanna Needell and Roman Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Comp. Math.*, 9(3):317–334, 2009.
- [31] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
- [32] Yurii Nesterov. Gradient methods for minimizing composite objective functions. CORE Discussion Paper 2007/76, Center for Operations Research and Econometrics, Université Catholique de Louvain, Belgium, 2007.
- [33] Yurii Nesterov and Arkadi Nemirovski. *Interior Point Polynomial Methods in Convex Programming*. SIAM, Philadelphia, PA, 1994.
- [34] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Verlag, 2006. Second edition.
- [35] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.
- [36] C. E. Shannon. Communication in the presence of noise. *Proc. Inst. of Radio Engineers*, 37(1):10–21, 1949.
- [37] Joel A. Tropp and Steven J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, in press.
- [38] W. Xu and B. Hassibi. Compressed sensing over the grassmann manifold: A unified analytical framework. Forty-Sixth Annual Allerton Conference, 2008.