SIGNAL RECOVERY FROM PARTIAL INFORMATION VIA ORTHOGONAL MATCHING PURSUIT

JOEL A. TROPP AND ANNA C. GILBERT

Abstract. This article demonstrates theoretically and empirically that a greedy algorithm called Orthogonal Matching Pursuit (OMP) can reliably recover a signal with m nonzero entries in dimension d given $O(m \ln d)$ random linear measurements of that signal. This is a massive improvement over previous results for OMP, which require $O(m^2)$ measurements. The new results for OMP are comparable with recent results for another algorithm called Basis Pursuit (BP). The OMP algorithm is much faster and much easier to implement, which makes it an attractive alternative to BP for signal recovery problems.

1. Introduction

Let $\boldsymbol{s}$ be a d-dimensional real signal with at most m nonzero components. This type of signal is called m-sparse. Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ be a sequence of measurement vectors in $\mathbb{R}^d$ that does not depend on the signal. We use these vectors to collect N linear measurements of the signal:

$$\langle \boldsymbol{s},\ \boldsymbol{x}_1 \rangle\,, \quad \langle \boldsymbol{s},\ \boldsymbol{x}_2 \rangle\,, \quad \dots, \quad \langle \boldsymbol{s},\ \boldsymbol{x}_N \rangle\,.$$

The problem of *signal recovery from partial information* asks two distinct questions:

(1) How many measurements are necessary to reconstruct the signal?
(2) Given these measurements, what algorithms can perform the reconstruction task?

As we will see, signal recovery is dual to sparse approximation, a problem of significant interest [MZ93, CDS99, RKD99, Mil02, Tem02].

To the first question, we can immediately respond that no fewer than m measurements will do. Even if the measurements were adapted to the signal, it would still take m pieces of information to determine all the components of an m-sparse signal. In the other direction, d nonadaptive measurements always suffice because we could simply list the d components of the signal. Although it is not obvious, sparse signals can be reconstructed with far less information.

The method for doing so has its origins during World War II. The US Army had a natural interest in screening soldiers for syphilis. But syphilis tests were expensive, and the Army realized that it was wasteful to perform individual assays to detect an occasional case. Their solution was to pool blood from groups of soldiers and test the pooled blood. If a batch checked positive, further tests could be performed. This method, called *group testing*, was subsequently studied in the computer science and statistics literatures. See [DH93] for a survey.

Very recently, a specific type of group testing has been proposed by the computational harmonic analysis community. The idea is that, by randomly combining the entries of a sparse signal, it is possible to generate a small set of summary statistics that allow us to identify the nonzero entries of the signal. The following theorem describes one example of this remarkable phenomenon.

Date: 11 April 2005.

2000 *Mathematics Subject Classification.* 41A46, 68Q25, 68W20, 90C27.

Key words and phrases. Algorithms, approximation, Basis Pursuit, group testing, Orthogonal Matching Pursuit, signal recovery, sparse approximation.

The authors are with the Department of Mathematics, The University of Michigan at Ann Arbor, 2074 East Hall, 530 Church St., Ann Arbor, MI 48109-1043. E-mail: {jtropp|annacg}@umich.edu. ACG has been supported by NSF DMS 0354600.

Theorem 1 (Rudelson–Vershynin [RV05, Thm. 2.1]). *Let $N \geq \mathrm{K}\, m \ln(d/m)$, and choose N vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ independently from the standard Gaussian distribution on $\mathbb{R}^d$. The following statement is true with probability exceeding $(1 - \mathrm{e}^{-\mathrm{k}\, N})$. It is possible to reconstruct every m-sparse signal $\boldsymbol{s}$ in $\mathbb{R}^d$ from the data $\{\langle \boldsymbol{s},\ \boldsymbol{x}_n \rangle\}$. The numbers* K *and* k *are (unspecified) universal constants.*

An important detail is that a particular choice of the Gaussian measurement vectors succeeds for *every* m-sparse signal with high probability. A similar theorem appears in a new paper of Candès and Tao [CT05]. These works extend earlier results of Candès–Romberg–Tao [CRT04], Donoho [Don04b], and Candès–Tao [CT04].

All five of the papers [CRT04, Don04b, CT04, RV05, CT05] offer constructive demonstrations of the recovery phenomenon by proving that the original signal $\boldsymbol{s}$ is the unique solution to the linear program

$$\min_{\boldsymbol{f}} \ \|\boldsymbol{f}\|_1 \quad \text{subject to} \quad \langle \boldsymbol{f},\ \boldsymbol{x}_n \rangle = \langle \boldsymbol{s},\ \boldsymbol{x}_n \rangle \quad \text{for } n = 1, \dots, N. \tag{BP}$$

This optimization problem provides an answer to our second question about how to reconstruct the sparse signal. Note that this formulation requires knowledge of the measurement vectors.

When we talk about (BP), we often say that the linear program can be solved in polynomial time with standard scientific software, and we cite books on convex programming such as [BV04]. This line of talk is misleading because it takes a long time to solve the linear program, even for signals of moderate length. For example, when $d = 8192$ and $N = 1024$, one implementation requires about an hour to solve (BP) [Rom05]. Furthermore, when off-the-shelf optimization software is not available, the implementation of optimization algorithms demands serious effort. Therefore, one might wish to consider alternate methods for reconstructing sparse signals from partial information.

To that end, we adapted a sparse approximation algorithm called Orthogonal Matching Pursuit (OMP) [PRK93, DMA97] to handle the signal recovery problem. The major advantages of this algorithm are its ease of implementation and its speed. In constrast with (BP), when $d = 8192$ and $N = 1024$, OMP can recover a 56-sparse signal in under four seconds. On the other hand, conventional wisdom on OMP has been pessimistic about its performance outside the simplest régimes. This complaint dates to a 1996 paper of DeVore and Temlyakov [DT96]. Pursuing their reasoning leads to an example of a nonrandom ensemble of measurement vectors and a sparse signal that OMP cannot identify without d measurements [CDS99, Sec. 2.3.2]. Other negative results, such as Theorem 3.10 of [Tro04] and Theorem 5 of [Don04a], echo this concern.

But these negative results about OMP are very deceptive. Indeed, the empirical evidence suggests that OMP can recover an m-sparse signal when the number of measurements N is a constant multiple of m. The goal of this paper is to present a rigorous proof that OMP can perform this feat. In particular, we establish the following theorem.

Theorem 2 (OMP with Gaussian Measurements). *Fix a positive number p, and let $N \geq \mathrm{K}_p\, m \ln d$. Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$, and choose N measurement vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ independently from the standard Gaussian distribution on $\mathbb{R}^d$. Given the data $\{\langle \boldsymbol{s},\ \boldsymbol{x}_n \rangle\}$ and the measurement vectors, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $(1 - 2\, d^{-p})$. For this theoretical result, it suffices that $\mathrm{K}_p = 8\,(p + 1)$.*

In comparison, previous positive results, such as Theorem 3.6 from [Tro04], only demonstrate that Orthogonal Matching Pursuit can recover m-sparse signals when the number of measurements N is on the order of m^2. Theorem 2 improves massively on this earlier work.

Theorem 2 is weaker than Theorem 1 for several reasons. First, our result requires somewhat more measurements than the result for (BP). Second, the quantifiers are ordered differently. Whereas we prove that OMP can recover any sparse signal given random measurements independent from the signal, the result for (BP) shows that a single set of random measurement vectors can be used to recover all sparse signals. In Section 6, we argue that these formal distinctions may be irrelevant in practice. Indeed, we believe that the large advantages of Orthogonal Matching Pursuit make Theorem 2 extremely compelling.

2. Orthogonal Matching Pursuit for Signal Recovery

This section describes a greedy algorithm for signal recovery. This method is analogous with Orthogonal Matching Pursuit, an algorithm for sparse approximation. First, let us motivate the computational technique.

Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$, and let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ be a family of N measurement vectors. Form an $N \times d$ matrix $\boldsymbol{\Phi}$ whose *rows* are the measurement vectors, and observe that the N measurements of the signal can be collected in an N-dimensional data vector $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$. We refer to $\boldsymbol{\Phi}$ as the *measurement matrix* and denote its columns by $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_d$.

As we mentioned, it is natural to think of signal recovery as a problem dual to sparse approximation. Since $\boldsymbol{s}$ has only m nonzero components, the data vector $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$ is a linear combination of m columns from $\boldsymbol{\Phi}$. In the language of sparse approximation, we say that $\boldsymbol{v}$ has an m-term representation over the dictionary $\boldsymbol{\Phi}$. This perspective allows us to transport results on sparse approximation to the signal recovery problem.

In particular, sparse approximation algorithms can be used for signal recovery. To identify the ideal signal $\boldsymbol{s}$, we need to determine *which* columns of $\boldsymbol{\Phi}$ participate in the measurement vector $\boldsymbol{v}$. The idea behind the algorithm is to pick columns in a greedy fashion. At each iteration, we choose the column of $\boldsymbol{\Phi}$ that is most strongly correlated with the remaining part of $\boldsymbol{v}$. Then we subtract off its contribution to $\boldsymbol{v}$ and iterate on the residual. One hopes that, after m iterations, the algorithm will have identified the correct set of columns.

Algorithm 3 (OMP for Signal Recovery)**.**
Input:

- *An $N \times d$ measurement matrix $\boldsymbol{\Phi}$*
- *An N-dimensional data vector $\boldsymbol{v}$*
- *The sparsity level m of the ideal signal*

Output:

- *An estimate $\widehat{\boldsymbol{s}}$ in $\mathbb{R}^d$ for the ideal signal*
- *A set Λ_m containing m elements from $\{1, \dots, d\}$*
- *An N-dimensional approximation $\boldsymbol{a}_m$ of the data vector $\boldsymbol{v}$*
- *An N-dimensional residual $\boldsymbol{r}_m = \boldsymbol{v} - \boldsymbol{a}_m$*

Procedure:

(1) *Initialize the residual $\boldsymbol{r}_0 = \boldsymbol{v}$, the index set $\Lambda_0 = \emptyset$, and the iteration counter $t = 1$.*
(2) *Find the index λ_t that solves the easy optimization problem*

$$\lambda_t \;=\; \arg\max_{j=1,\dots,d} \; |\langle \boldsymbol{r}_{t-1},\; \boldsymbol{\varphi}_j \rangle| \,.$$

If the maximum occurs for multiple indices, break the tie deterministically.
(3) *Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$.*
(4) *Determine the orthogonal projector $\boldsymbol{P}_t$ onto* $\operatorname{span}\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_t\}$.
(5) *Calculate the new approximation and residual:*

$$\begin{aligned} \boldsymbol{a}_t &= \boldsymbol{P}_t\,\boldsymbol{v} \\ \boldsymbol{r}_t &= \boldsymbol{v} - \boldsymbol{a}_t. \end{aligned}$$

(6) *Increment t, and return to Step 2 if $t < m$.*
(7) *The estimate $\widehat{\boldsymbol{s}}$ for the ideal signal has nonzero indices at the components listed in Λ_m. The values of the estimate in these components appear in the series expansion*

$$\boldsymbol{a}_m \;=\; \sum\nolimits_{\lambda \in \Lambda_m} \widehat{s}_\lambda\, \boldsymbol{\varphi}_\lambda.$$

The running time of the OMP algorithm is dominated by Step 2, whose total cost is $O(m\,d\,N)$. No one has reported the cost of solving (BP) with a dense, unstructured measurement matrix.

Steps 4, 5, and 7 have been written to emphasize the conceptual structure of the algorithm. It is possible to implement them efficiently using standard techniques for least-squares problems. See [GVL96, Chap. 5] for extensive details. It is important to recognize that the residual $\boldsymbol{r}_t$ is always orthogonal to the columns indexed in Λ_t. Therefore, the set $\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_t\}$ is linearly independent and contains t distinct vectors.

A prototype of this algorithm first appeared in the statistics community [FS81], but it developed a life of its own in the signal processing [MZ93, PRK93, DMA97] and approximation theory [DeV98, Tem02] literatures. Our adaptation for the signal recovery problem seems to be new.

3. Random Measurement Ensembles

This article demonstrates that Orthogonal Matching Pursuit can recover sparse signals given a set of random linear measurements. The two obvious distributions for the $N \times d$ measurement matrix $\boldsymbol{\Phi}$ are (1) Gaussian and (2) Bernoulli, normalized for mathematical convenience:

(1) Independently select each entry of $\boldsymbol{\Phi}$ from the $\textsc{normal}(0, N^{-1})$ distribution. For reference, the density function of this distribution is

$$p(x) \quad \stackrel{\text{def}}{=} \quad \frac{1}{\sqrt{2\pi\,N}}\,\mathrm{e}^{-x^2\,N/2}.$$

(2) Independently select each entry of $\boldsymbol{\Phi}$ to be $\pm 1/\sqrt{N}$ with equal probability.

Indeed, either one of these distributions can be used to collect measurements. More generally, the measurement ensemble can be chosen from any distribution that meets a few basic requirements. We will abstract these properties, even though we are primarily interested in the foregoing examples.

3.1. Admissible Measurement Matrices.

An *admissible measurement matrix* for m-sparse signals in $\mathbb{R}^d$ is an $N \times d$ random matrix $\boldsymbol{\Phi}$ with four properties.

(M0) Independence: The columns of $\boldsymbol{\Phi}$ are stochastically independent.

(M1) Normalization: $\mathbb{E}\,\|\boldsymbol{\varphi}_j\|_2^2 = 1$ for $j = 1, \dots, d$.

(M2) Joint correlation: Let $\{\boldsymbol{u}_t\}$ be a sequence of m vectors whose ℓ_2 norms do not exceed one. Let $\boldsymbol{\varphi}$ be a column of $\boldsymbol{\Phi}$ that is independent from this sequence. Then

$$\text{Prob}\,\{\max_t\,|\langle \boldsymbol{\varphi},\ \boldsymbol{u}_t\rangle| \ \le\ \varepsilon\} \ \ge\ 1 \ -\ 2\,m\exp\{-\mathrm{C}_2\,N\,\varepsilon^2\}$$

for a positive constant C_2.

(M3) Smallest singular value: Given an $N \times m$ submatrix $\boldsymbol{Z}$ from $\boldsymbol{\Phi}$, the m-th singular value $\sigma_{\min}(\boldsymbol{Z})$ satisfies

$$\text{Prob}\left\{\sigma_{\min}(\boldsymbol{Z}) \ \ge\ 1 - \sqrt{m/N} - \varepsilon\right\} \ \ge\ 1 \ -\ \exp\{-\mathrm{C}_3\,N\,\varepsilon^2\}$$

for a positive constant C_3.

Some remarks may help delineate the range of this definition. First, note that the columns of $\boldsymbol{\Phi}$ need not have the same distribution. Condition (M0) only requires independence of columns; the entries within each column may be correlated. The unit normalization in (M1) is chosen to simplify our proofs, but it should be obvious that the signal recovery problem does not depend on the scale of the measurement matrix. The property (M2) depends on the decay of the random variables $\|\boldsymbol{\varphi}_j\|_2$, and a version of (M2) holds whenever these variables have subgaussian tails. Many types of large random matrices satisfy something like (M3) on account of the Marčenko–Pastur Law for singular values [MP67], which holds under weak regularity assumptions.

In the next two subsections, we explain why the Gaussian and Bernoulli ensembles both yield admissible measurement matrices. Afterward, we compare admissible measurement matrices with other types of measurement ensembles that have appeared in the literature.

3.2. **Gaussian Measurements.** If the entries of $\mathbf{\Phi}$ are chosen from independent $\text{NORMAL}(0, N^{-1})$ distributions, then $\mathbf{\Phi}$ is an admissible measurement matrix with $\mathrm{C}_2 = \mathrm{C}_3 = \frac{1}{2}$. The independence and normalization requirements, (M0) and (M1), are straightforward to verify. To check the other two properties, we leverage results from the literature. The joint correlation property (M2) follows easily from standard Gaussian tail bounds and Boole's inequality.

Proposition 4. *Let $\{\boldsymbol{u}_t\}$ be a sequence of m vectors whose ℓ_2 norms do not exceed one. Independently, choose $\boldsymbol{z}$ to be a random vector with iid $\text{NORMAL}(0, N^{-1})$ entries. Then*

$$\text{Prob}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| \ \le\ \varepsilon\right\} \ \ge\ 1 \ -\ m\exp\{-\tfrac{1}{2}N\varepsilon^2\}.$$

Note that this result is already twice as strong as (M2) requires, and it can be honed even sharper with Sidak's Lemma [Bal02, Lemma 2.1] or with Slepian's Lemma [LT91, Sec. 3.3].

Proof. Observe that the probability only decreases as the length of each vector $\boldsymbol{u}_t$ increases. Therefore, we may assume that $\|\boldsymbol{u}_t\|_2 = 1$ for each t. Suppose that $\boldsymbol{z}$ is a random vector with iid $\text{NORMAL}(0, N^{-1})$ entries. Then the random variable $\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle$ also has the $\text{NORMAL}(0, N^{-1})$ distribution. A well-known Gaussian tail bound yields

$$\text{Prob}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| \ >\ \varepsilon\right\} \ =\ \sqrt{\frac{2}{\pi}}\int_{\varepsilon\sqrt{N}}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \ \le\ \exp\{-\tfrac{1}{2}N\varepsilon^2\}. \tag{3.1}$$

Using Boole's inequality,

$$\text{Prob}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| \ >\ \varepsilon\right\} \ \le\ m\exp\{-\tfrac{1}{2}N\varepsilon^2\}.$$

This bound is complementary to the one stated. □

The singular value property (M3) is contained in Theorem 2.13 of [DS02].

Proposition 5 (Davidson–Szarek)**.** *Suppose that $\boldsymbol{Z}$ is a tall $N \times m$ matrix whose entries are iid $\text{NORMAL}(0, N^{-1})$. Then its m-th singular value $\sigma_{\min}$ satisfies*

$$\text{Prob}\left\{\sigma_{\min}(\boldsymbol{Z}) \ \ge\ 1 - \sqrt{m/N} - \varepsilon\right\} \ \ge\ 1 \ -\ \exp\{-\tfrac{1}{2}N\varepsilon^2\}.$$

It is a standard consequence of measure concentration that the minimum singular value of a Gaussian matrix clusters around its expected value (see [Led01, Sec. 8.5], for example). Estimating the expectation, however, involves much more ingenuity. Davidson and Szarek produce their result with a clever application of the Slepian–Gordon Lemma. Sadly, their method does not apply to other types of matrices.

3.3. **Bernoulli Measurements.** This subsection sketches the argument that the Bernoulli ensemble yields admissible measurement matrices. The independence and normalization properties, (M0) and (M1), are both straightforward to check.

The joint correlation property (M2) with constant $\mathrm{C}_2 = \frac{1}{2}$ can be established by an argument analogous to that of Proposition 4. In this case, we replace the Gaussian tail bound with

$$\text{Prob}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| \ >\ \varepsilon\right\} \ \le\ 2\exp\{-\tfrac{1}{2}N\varepsilon^2\}. \tag{3.2}$$

It is easy to develop this estimate with the Laplace transform method [LT91, Sec. 4.1].

The singular value property (M3) seems to require some heavy machinery, and we have not been able to obtain a precise value for the constant C_3. We refer the reader to Section 4.2 of [CT04], which establishes that, for large m and N, there is a number η for which

$$\text{Prob}\left\{\sigma_{\min}(\boldsymbol{Z}) \ <\ 1 - \sqrt{m/N} - \eta - \varepsilon\right\} \ \le\ \exp\{-\tfrac{1}{16}N\varepsilon^2\}.$$

As m and N tend to infinity, the number $\eta \to 0$. It seems as if this statement could be replaced by something more concrete, but an extensive literature search did not yield fruit.

3.4. **Other Types of Measurement Ensembles.** It is valuable to compare admissible measurement matrices with the measurement ensembles introduced in other works on signal recovery. Here is a short summary of the types of measurement matrices that have appeared in the literature.

- In their first paper [CT04], Candès and Tao define random matrices that satisfy the Uniform Uncertainty Principle and the Exact Reconstruction Principle. Gaussian and Bernoulli matrices both meet these requirements. In a subsequent paper [CT05], they develop a deterministic class of matrices whose "restricted isometry constants" are under control. They show that Gaussian random matrices satisfy this property with high probability.
- Donoho introduces the deterministic class of compressed sensing (CS) matrices [Don04b]. He shows that Gaussian random matrices fall in this class with high probability.
- The approach in Rudelson and Vershynin's paper [RV05] is more direct. They prove that, if the measurement matrix spans a random subspace, then (BP) succeeds with high probability. Their method relies on the geometry of random slices of a high-dimensional cube. As such, their measurement ensembles are described intrinsically, in contrast with the extrinsic definitions of the other ensembles.

The extrinsic classes of measurement matrices involve conditions of two different flavors. The first flavor places a control on the singular values of submatrices. This property appears in our condition (M3) and in Donoho's condition (CS1). It also appears in the Uniform Uncertainty Principle (SC1) and the restricted isometry constants of Candès and Tao.

The second flavor is related to the type of reconstruction algorithm employed. Our algorithmic condition (M2) is couched in terms of inner products since the OMP algorithm relies on inner products during the greedy step. The other authors all use ℓ_1-minimization to reconstruct the input signal, which leads to ℓ_1-type conditions on their measurement matrices. Donoho uses conditions (CS2) and (CS3) to relate the ℓ_2 and ℓ_1 norms of certain vectors supported both on and off the measurement set. Candès and Tao enforce a similar condition in their Exact Reconstruction Principle (SC2).

4. Signal Recovery with Orthogonal Matching Pursuit

If we take random measurements of a sparse signal using an admissible measurement matrix, then OMP can be used to recover the original signal with high probability.

Theorem 6 (OMP with Admissible Measurements)**.** *Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$, and draw a random $N \times d$ admissible measurement matrix $\boldsymbol{\Phi}$ independent from the signal. Given the data $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding*

$$\sup_{\varepsilon \in [0, \sqrt{N/m}-1]} \left[1 \;-\; 2\,m \exp\{-\mathrm{C}_2\,(\sqrt{N/m} - 1 - \varepsilon)^2\}\right]^{d-m} \left[1 \;-\; \exp\{-\mathrm{C}_3\,m\,\varepsilon^2\}\right].$$

The success probability here is best calculated numerically. Some analysis yields a weaker but more concrete corollary. We defer the proofs until we have discussed the results.

Corollary 7. *Fix a positive number p, and let $N \geq \mathrm{K}_{p,m}\, m \ln d$. Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$, and draw a random $N \times d$ admissible measurement matrix $\boldsymbol{\Phi}$ independent from the signal. Given the data $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $(1 - 2\,d^{-p})$. For this theoretical result to hold, it suffices that*

$$\mathrm{K}_{p,m} \;\geq\; \left[\sqrt{\frac{p+2}{\mathrm{C}_2}} \;+\; \sqrt{\frac{p}{\mathrm{C}_3\,m}}\right]^2.$$

This paragraph describes the situation for the Gaussian case, which relies on the fact that $\mathrm{C}_2 = \mathrm{C}_3 = \frac{1}{2}$. To reach Theorem 2 of the Introduction, just define the constant

$$\mathrm{K}_p \stackrel{\text{def}}{=} \mathrm{K}_{p,1} = \sup_{m \geq 1} \mathrm{K}_{p,m}.$$

The inequality $\sqrt{a+b}\sqrt{a-b} \leq a$ yields

$$\begin{aligned} \mathrm{K}_p &= \left[\sqrt{2p+4} + \sqrt{2p}\right]^2 \\ &\leq (2p+4) + 2\,(2p+2) + 2p \\ &= 8\,(p+1). \end{aligned}$$

Of course, the sharper result in Corollary 7 provides much better bounds. As $m \to \infty$, we have $\mathrm{K}_{p,m} \to 2\,(p+2)$. Therefore, K_p badly overstates the number of measurements necessary for most values of m. For example, when m and d are both large, the theory yields a failure probability around d^{-1} if we choose $N = 6\,m\,\ln d$. As we will see, the empirical results indicate that fewer measurements are necessary in practice.

Even though OMP may fail, it is important to recognize that the user can detect a success or failure in the present setting. We state a simple result for Gaussian measurements.

Proposition 8. *Choose an arbitrary m-sparse signal $\boldsymbol{s}$ from $\mathbb{R}^d$, and let $N > 2\,m$. Suppose that $\boldsymbol{\Phi}$ is a $d \times N$ Gaussian measurement ensemble, and execute OMP with the data $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$. If the residual $\boldsymbol{r}_m$ after m iterations is zero after m iterations, then OMP has correctly identified $\boldsymbol{s}$ with probability one. Conversely, if the residual after m iterations is nonzero, then OMP has failed.*

Proof. The converse is obvious, so we concentrate on the forward direction. If $\boldsymbol{r}_m = \mathbf{0}$ but $\widehat{\boldsymbol{s}} \neq \boldsymbol{s}$, then it is possible to write the data vector $\boldsymbol{v}$ as a linear combination of m columns from $\boldsymbol{\Phi}$ in two different ways. In consequence, there is a linear dependence among $2\,m$ columns from $\boldsymbol{\Phi}$. Since $\boldsymbol{\Phi}$ is an $N \times d$ Gaussian matrix and $2\,m < N$, this event occurs with probability zero. Geometrically, this observation is equivalent to the fact that independent Gaussian vectors lie in general position with probability one. This claim follows from the zero-one law for Gaussian processes [Fer74, Sec. 1.2]. The kernel of our argument originates in [Don04a, Lemma 2.1]. □

For Bernoulli measurements, a similar result holds with probability exponentially close to one. This claim follows from the fact that an exponentially small fraction of Bernoulli matrices are singular [KKS95].

4.1. **Comparison with Prior Work.** Most results on OMP rely on the *coherence parameter* μ of the matrix $\boldsymbol{\Phi}$. This summary statistic measures the correlation between distinct columns of the matrix:

$$\mu \stackrel{\text{def}}{=} \max_{j \neq k} \left|\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k \rangle\right|.$$

The next lemma shows that the coherence of a Gaussian or Bernoulli matrix is fairly small.

Lemma 9. *For an $N \times d$ Gaussian or Bernoulli measurement matrix, the coherence parameter $\mu \leq \ln d / \sqrt{N}$ with probability exceeding $(1 - d^{2-(\ln d)/2})$.*

The probability estimate in this lemma approaches one at a superpolynomial rate. It can also be established that $\mu \geq 1/\sqrt{N}$ with probability exponentially close to one as d grows.

A typical coherence result for OMP, such as Theorem 3.6 of [Tro04], shows that the algorithm can recover any m-sparse signal provided that $m\,\mu \leq \frac{1}{2}$. This theorem applies immediately to the Bernoulli case. The Gaussian case requires some work because the columns of $\boldsymbol{\Phi}$ no longer have unit norm. We have the following result.

Proposition 10. *Let $N \geq (2\,m\,\ln d)^2$, and choose an $N \times d$ Gaussian or Bernoulli measurement matrix $\mathbf{\Phi}$. The following statement holds with probability superpolynomially close to one. It is possible to reconstruct every m-sparse signal in $\mathbb{R}^d$ from the data $\boldsymbol{v} = \mathbf{\Phi}\,\boldsymbol{s}$.*

A proof sketch appears at the end of the section.

There are several important differences between Proposition 10 and Theorem 6. The proposition shows that a particular choice of the measurement matrix succeeds for *every* m-sparse signal. In comparison with our new results, however, it requires an enormous number of measurements. Since the coherence parameter is unlikely to be smaller than $1/\sqrt{N}$, it is impossible to develop substantially stronger results with coherence.

4.2. **Proof of Theorem 6.** Most of the argument follows the approach developed in [Tro04]. The main difficulty here is to deal with the nasty independence issues that arise in the stochastic setting. The primary novelty is a route to avoid these perils.

We begin with some notation and simplifying assumptions. Without loss of generality, assume that the first m entries of the original signal $\boldsymbol{s}$ are nonzero, while the remaining $(d - m)$ entries equal zero. Therefore, the measurement vector $\boldsymbol{v}$ is a linear combination of the first m columns from the matrix $\mathbf{\Phi}$. Partition the matrix as $\mathbf{\Phi} = [\mathbf{\Phi}_{\text{opt}} \mid \mathbf{\Psi}]$ so that $\mathbf{\Phi}_{\text{opt}}$ has m columns and $\mathbf{\Psi}$ has $(d - m)$ columns. Note that $\boldsymbol{v}$ is stochastically independent from the random matrix $\mathbf{\Psi}$.

For a vector $\boldsymbol{r}$ in $\mathbb{R}^N$, define the *greedy selection ratio*

$$\rho(\boldsymbol{r}) \stackrel{\text{def}}{=} \frac{\left\|\mathbf{\Psi}^T \boldsymbol{r}\right\|_\infty}{\left\|\mathbf{\Phi}_{\text{opt}}^T \boldsymbol{r}\right\|_\infty} = \frac{\max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\ \boldsymbol{r}\rangle|}{\left\|\mathbf{\Phi}_{\text{opt}}^T \boldsymbol{r}\right\|_\infty}$$

where the maximization takes place over the columns of $\mathbf{\Psi}$. If $\boldsymbol{r}$ is the residual vector that arises in Step 2 of OMP, the algorithm picks a column from $\mathbf{\Phi}_{\text{opt}}$ provided $\rho(\boldsymbol{r}) < 1$. The greedy selection ratio was first identified and studied in [Tro04].

Now imagine that we could execute OMP with the input signal $\boldsymbol{s}$ and the matrix $\mathbf{\Phi}_{\text{opt}}$ to obtain a sequence of residuals: $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{m-1}$. The algorithm is deterministic, so this sequence is simply a function of $\boldsymbol{s}$ and $\mathbf{\Phi}_{\text{opt}}$. In particular, the residuals are stochastically independent from $\mathbf{\Psi}$. It is also evident that each residual lies in the column span of $\mathbf{\Phi}_{\text{opt}}$.

Instead, suppose that we execute OMP with the input signal $\boldsymbol{s}$ and the full matrix $\mathbf{\Phi}$. Observe that, if OMP should succeed in reconstructing $\boldsymbol{s}$ after m iterations, then the algorithm must choose a column from $\mathbf{\Phi}_{\text{opt}}$ in each iteration. The matrix $\mathbf{\Psi}$ affects neither the greedy selection nor the formation of the residuals. Therefore, the algorithm must generate the *same* sequence of residuals as in our mental experiment: $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{m-1}$.

In consequence, the event on which the algorithm succeeds is

$$E_{\text{succ}} \stackrel{\text{def}}{=} \{\ \max_t\ \rho(\boldsymbol{r}_t)\ <\ 1\ \}$$

where $\{\boldsymbol{r}_t\}$ is a sequence of m random variables that depends only on $\mathbf{\Phi}_{\text{opt}}$. We can decrease the probability of success by placing the additional requirement that the smallest singular value of $\mathbf{\Phi}_{\text{opt}}$ meet a lower bound:

$$\text{Prob}\,\{E_{\text{succ}}\}\ \geq\ \text{Prob}\,\{\max_t\ \rho(\boldsymbol{r}_t)\ <\ 1\ \wedge\ \sigma_{\min}(\mathbf{\Phi}_{\text{opt}}) \geq \sigma\}\,.$$

We will use Σ to abbreviate the event $\{\sigma_{\min}(\mathbf{\Phi}_{\text{opt}}) \geq \sigma\}$. Applying the definition of conditional probability, we reach

$$\text{Prob}\,\{E_{\text{succ}}\}\ \geq\ \text{Prob}\,\{\max_t\ \rho(\boldsymbol{r}_t)\ <\ 1 \mid \Sigma\}\ \text{Prob}\,\{\Sigma\}\,. \tag{4.1}$$

Property (M3) controls $\text{Prob}\,\{\Sigma\}$, so it remains to develop a lower bound on the conditional probability.

Assume that Σ occurs. For each index $t = 0, \dots, m-1$, we have

$$\rho(\boldsymbol{r}_t) \;=\; \frac{\max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\, \boldsymbol{r}_t \rangle|}{\left\| \boldsymbol{\Phi}_{\text{opt}}^T \boldsymbol{r}_t \right\|_\infty}.$$

Since $\boldsymbol{\Phi}_{\text{opt}}^T \boldsymbol{r}_t$ is an m-dimensional vector,

$$\rho(\boldsymbol{r}_t) \;\le\; \frac{\sqrt{m}\, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\, \boldsymbol{r}_t \rangle|}{\left\| \boldsymbol{\Phi}_{\text{opt}}^T \boldsymbol{r}_t \right\|_2}.$$

To simplify this expression, define the vector

$$\boldsymbol{u}_t \;\stackrel{\text{def}}{=}\; \frac{\sigma\, \boldsymbol{r}_t}{\left\| \boldsymbol{\Phi}_{\text{opt}}^T \boldsymbol{r}_t \right\|_2}.$$

The basic properties of singular values furnish the inequality

$$\frac{\left\| \boldsymbol{\Phi}_{\text{opt}}^T \boldsymbol{r} \right\|_2}{\|\boldsymbol{r}\|_2} \;\ge\; \sigma_{\min}(\boldsymbol{\Phi}_{\text{opt}}) \;\ge\; \sigma$$

for any vector $\boldsymbol{r}$ in the range of $\boldsymbol{\Phi}_{\text{opt}}$. The residual $\boldsymbol{r}_t$ falls in this subspace, so $\|\boldsymbol{u}_t\|_2 \le 1$. In summary,

$$\rho(\boldsymbol{r}_t) \;\le\; \frac{\sqrt{m}}{\sigma}\, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\, \boldsymbol{u}_t \rangle|$$

for each index t. On account of this fact,

$$\text{Prob}\left\{ \max_t\, \rho(\boldsymbol{r}_t) \;<\; 1 \mid \Sigma \right\} \;\ge\; \text{Prob}\left\{ \max_t\, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\, \boldsymbol{u}_t \rangle| \;<\; \frac{\sigma}{\sqrt{m}} \;\middle|\; \Sigma \right\}.$$

Exchange the two maxima and use the independence of the columns of $\boldsymbol{\Psi}$ to obtain

$$\text{Prob}\left\{ \max_t\, \rho(\boldsymbol{r}_t) \;<\; 1 \mid \Sigma \right\} \;\ge\; \prod\nolimits_{\boldsymbol{\psi}} \text{Prob}\left\{ \max_t\, |\langle \boldsymbol{\psi},\, \boldsymbol{u}_t \rangle| \;<\; \frac{\sigma}{\sqrt{m}} \;\middle|\; \Sigma \right\}.$$

Since every column of $\boldsymbol{\Psi}$ is independent from $\{\boldsymbol{u}_t\}$ and from Σ, Property (M2) of the measurement matrix yields a lower bound on each of the $(d-m)$ terms appearing in the product. It emerges that

$$\text{Prob}\left\{ \max_t\, \rho(\boldsymbol{r}_t) \;<\; 1 \mid \Sigma \right\} \;\ge\; \left[1 \;-\; 2\, m \exp\{-\text{C}_2\, N\, \sigma^2/m\} \right]^{d-m}.$$

We may choose the parameter

$$\sigma \;=\; 1 - \sqrt{m/N} - \varepsilon\, \sqrt{m/N}$$

where ε ranges between zero and $\sqrt{N/m} - 1$. This substitution delivers

$$\text{Prob}\left\{ \max_t\, \rho(\boldsymbol{r}_t) \;<\; 1 \mid \Sigma \right\} \;\ge\; \left[1 \;-\; 2\, m \exp\{-\text{C}_2\, (\sqrt{N/m} - 1 - \varepsilon)^2\} \right]^{d-m}. \tag{4.2}$$

With the foregoing choice of σ, Property (M3) shows that

$$\text{Prob}\left\{ \sigma_{\min}(\boldsymbol{\Phi}_{\text{opt}}) \;\ge\; \sigma \right\} \;\ge\; 1 \;-\; \exp\{-\text{C}_3\, m\, \varepsilon^2\}.$$

Introduce this fact and (4.2) into the inequality (4.1). This action yields

$$\text{Prob}\{E_{\text{succ}}\} \;\ge\; \left[1 \;-\; 2\, m \exp\{-\text{C}_2\, (\sqrt{N/m} - 1 - \varepsilon)^2\} \right]^{d-m} \left[1 \;-\; \exp\{-\text{C}_3\, m\, \varepsilon^2\} \right]. \tag{4.3}$$

In principle, we could maximize the right-hand side over ε to obtain the best lower bound on the probability of success. This optimization is best performed numerically.

4.3. **Proof of Corollary 7.** We need to show that it is possible to choose the number of measurements N on the order of $m \ln d$ while maintaining an error polynomially small in d. Begin with (4.3), and apply the inequality $(1-x)^q \geq 1 - qx$, valid for $q \geq 1$ and $x \leq 1$. This yields

$$\operatorname{Prob}\{E_{\text{succ}}\} \;\geq\; \left[1 \;-\; 2\,m\,(d-m)\,\exp\{-\mathrm{C}_2\,(\sqrt{N/m} - 1 - \varepsilon)^2\}\right]\;\left[1 \;-\; \exp\{-\mathrm{C}_3\,m\,\varepsilon^2\}\right].$$

Next, use the fact that $m\,(d-m) \leq d^2/4$ to reach

$$\operatorname{Prob}\{E_{\text{succ}}\} \;\geq\; \left[1 \;-\; \tfrac{1}{2}\,d^2\,\exp\{-\mathrm{C}_2\,(\sqrt{N/m} - 1 - \varepsilon)^2\}\right]\;\left[1 \;-\; \exp\{-\mathrm{C}_3\,m\,\varepsilon^2\}\right]. \tag{4.4}$$

We develop separate bounds for the two terms on the right-hand side.

Fix a positive number p, and select $N \geq \mathrm{K}_{p,m}\,m\,\ln d$ with the constant $\mathrm{K}_{p,m}$ to be determined momentarily. Choose

$$\varepsilon \;=\; 1 \;+\; \left[\frac{p\,\ln d}{\mathrm{C}_3\,m}\right]^{1/2}.$$

Clearly $\varepsilon \geq 0$. Our selection of $\mathrm{K}_{p,m}$ will ensure that ε is also smaller than $\sqrt{N/m} - 1$, as required. By definition of N and ε, we have

$$\sqrt{N/m} - 1 - \varepsilon \;\geq\; \left[\sqrt{\mathrm{K}_{p,m}} - \sqrt{\frac{p}{\mathrm{C}_3\,m}}\right]\,\sqrt{\ln d}.$$

This inequality implies

$$d^2\,\exp\{-\mathrm{C}_2\,(\sqrt{N/m} - 1 - \varepsilon)^2\} \;\leq\; d^{2-\mathrm{C}_2\,(\sqrt{\mathrm{K}_{p,m}} - \sqrt{p/\mathrm{C}_3 m})^2}.$$

To keep the error probability on the order of d^{-p}, we set

$$2 - \mathrm{C}_2\,(\sqrt{\mathrm{K}_{p,m}} - \sqrt{p/\mathrm{C}_3 m})^2 \;\leq\; -p.$$

Solve this inequality to find that

$$\mathrm{K}_{p,m} \;\geq\; \left[\sqrt{\frac{p+2}{\mathrm{C}_2}} \;+\; \sqrt{\frac{p}{\mathrm{C}_3\,m}}\right]^2.$$

With these choices, the first term of the probability bound (4.4) is no smaller than $(1 - \frac{1}{2}\,d^{-p})$.

Let us move to the second term. We can bound ε below using the inequality $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$.

$$\varepsilon \;=\; [\ln \mathrm{e}]^{1/2} \;+\; \left[\frac{p\,\ln d}{\mathrm{C}_3\,m}\right]^{1/2} \;\geq\; \left[\ln\left(\mathrm{e}\,d^{p/\mathrm{C}_3 m}\right)\right]^{1/2}.$$

It follows immediately that

$$\exp\{-\mathrm{C}_3\,m\,\varepsilon^2\} \;\leq\; (\mathrm{e}\,d^{p/\mathrm{C}_3 m})^{-\mathrm{C}_3\,m} \;\leq\; \mathrm{e}^{-\mathrm{C}_3\,m}\,d^{-p}.$$

So the second term of the bound (4.4) is no smaller than $(1 - \mathrm{e}^{-\mathrm{C}_3\,m}\,d^{-p})$. Multiply the bounds on the first and second term together to conclude that

$$\begin{aligned}
\operatorname{Prob}\{E_{\text{succ}}\} \;&\geq\; (1 - \tfrac{1}{2}\,d^{-p})\,(1 - \mathrm{e}^{-\mathrm{C}_3\,m}\,d^{-p}) \\
&\geq\; 1 \;-\; (\tfrac{1}{2} + \mathrm{e}^{-\mathrm{C}_3\,m})\,d^{-p} \\
&\geq\; 1 \;-\; \tfrac{3}{2}\,d^{-p}.
\end{aligned}$$

In the statement of the theorem, we have replaced the factor of 3/2 with a factor of two for simplicity.

4.4. **Proof of Coherence Results.** The section concludes with a sketch of the coherence results. First we demonstrate that the coherence parameter is probably not too large.

Proof of Lemma 9. Suppose that $\mathbf{\Phi}$ is an $N \times d$ Gaussian measurement matrix. For all indices $j < k$, the random variables $\{\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k \rangle\}$ are $\text{NORMAL}(0, N^{-1})$. Using the Gaussian tail bound (3.1) and Boole's inequality,

$$\text{Prob}\left\{\max_{j<k} |\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k \rangle| \ >\ \varepsilon\right\} \ \le\ \tfrac{1}{2}\, d\,(d-1) \exp\{-\tfrac{1}{2}\,\varepsilon^2\, N\} \ \le\ d^2\, \exp\{-\tfrac{1}{2}\,\varepsilon^2\, N\}.$$

Choose $\varepsilon = \ln d / \sqrt{N}$ and observe that $\mu = \max_{j<k} |\langle \boldsymbol{\varphi}_j,\ \boldsymbol{\varphi}_k \rangle|$ to conclude that

$$\text{Prob}\left\{\mu \ >\ \frac{\ln d}{\sqrt{N}}\right\} \ \le\ d^{2-(\ln d)/2}.$$

Using (3.2), we see that the same estimate holds for a Bernoulli measurement matrix. □

Using this coherence bound, we can develop a positive result for signal recovery via OMP. As we mentioned, the Bernoulli case follows directly from results in the literature. Therefore, we only establish the Gaussian case.

Proposition 10, Proof Sketch. The first part of argument is essentially identical with Theorem 3.3 from [Tro04], so we omit the details. Write the matrix $(\mathbf{\Phi}_{\text{opt}}^T\, \mathbf{\Phi}_{\text{opt}}) = \boldsymbol{D} - \boldsymbol{A}$, where $\boldsymbol{D}$ is the diagonal part and $\boldsymbol{A}$ the off-diagonal part. We obtain a sufficient condition for OMP to recover all m-sparse signals:

$$m \ <\ \frac{1}{2}\,\left[\left\|\boldsymbol{D}^{-1}\right\|_{1,1}^{-1}\, \mu^{-1} + 1\right] \tag{4.5}$$

where $\|\cdot\|_{1,1}$ returns the maximum absolute column sum of a matrix. Observe that the diagonal entries of $\boldsymbol{D}$ are the squared ℓ_2 norms of columns from $\mathbf{\Phi}$. Therefore,

$$\left\|\boldsymbol{D}^{-1}\right\|_{1,1}^{-1} \ \ge\ \min_j \|\boldsymbol{\varphi}_j\|_2^2\,. \tag{4.6}$$

Let us see how (4.5) applies when $\mathbf{\Phi}$ is a Gaussian measurement matrix. The random variable $N\,\|\boldsymbol{\varphi}_j\|_2^2$ has the χ^2 distribution with N degrees of freedom. Using the Laplace transform method, it can be shown that

$$\text{Prob}\left\{\|\boldsymbol{\varphi}_j\|_2^2 \ \ge\ 1 - \frac{2\,\varepsilon}{\sqrt{N}}\right\} \ \ge\ 1 - \mathrm{e}^{-\varepsilon^2}.$$

Since the columns of $\mathbf{\Phi}$ are stochastically independent,

$$\text{Prob}\left\{\min_j \|\boldsymbol{\varphi}_j\|_2^2 \ \ge\ 1 - \frac{2\,\varepsilon}{\sqrt{N}}\right\} \ \ge\ \left[1 - \mathrm{e}^{-\varepsilon^2}\right]^d.$$

Apply (4.6), and choose $\varepsilon = \frac{1}{2} \ln d$ to reach

$$\text{Prob}\left\{\left\|\boldsymbol{D}^{-1}\right\|_{1,1}^{-1} \ \ge\ 1 - \frac{\ln d}{\sqrt{N}}\right\} \ \ge\ 1 - d^{1-(\ln d)/4}.$$

According to Lemma 9, the coherence parameter $\mu^{-1} \ge \sqrt{N}/\ln d$ with probability no less than $(1 - d^{2-(\ln d)/2})$. Introducing these calculations into (4.5), we find that the condition

$$m \ \le\ \frac{\sqrt{N}}{2 \ln d} \tag{4.7}$$

is probably sufficient for OMP to recover all m-sparse signals using the measurement matrix $\mathbf{\Phi}$. The probability over $\mathbf{\Phi}$ that (4.7) does not suffice is superpolynomially small in d. □

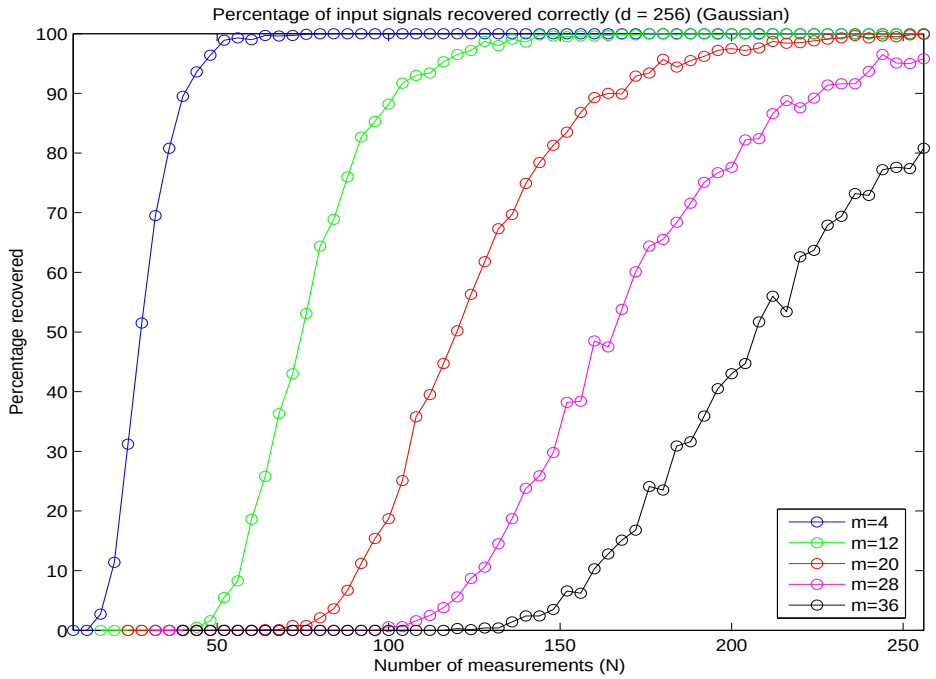

FIGURE 1. The percentage of 1000 input signals correctly recovered as a function of the number N of measurements for different sparsity levels m in dimension $d = 256$.

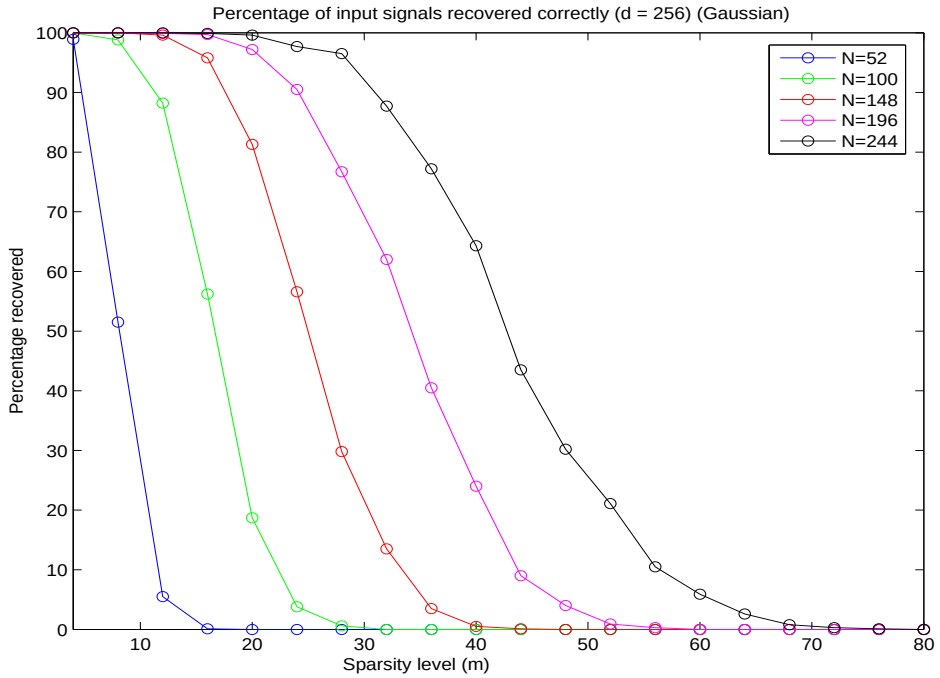

FIGURE 2. The percentage of 1000 input signals correctly recovered as a function of the sparsity level m for different numbers N of measurements in dimension $d = 256$.

5. Experiments

This section illustrates experimentally that OMP is a powerful algorithm for signal recovery. It also shows that the theoretical bounds of the last section are qualitatively correct even though they are slightly pessimistic.

The main empirical question is to determine how many measurements N are necessary to recover an m-sparse signal in $\mathbb{R}^d$ with high probability. Let us describe the experimental setup. In each trial, we generate an m-sparse signal $\boldsymbol{s}$ by choosing m components (out of d) at random and setting them equal to one[1]. We draw an $N \times d$ Gaussian measurement matrix $\boldsymbol{\Phi}$ and execute OMP with the data vector $\boldsymbol{v} = \boldsymbol{\Phi}\,\boldsymbol{s}$. Finally, we check whether the recovered signal is identical with the original signal. For each triple (m, N, d), we perform 1000 independent trials.

The first plot, Figure 1, describes the situation in dimension $d = 256$. It shows what percentage (of the 1000 trial signals) were recovered correctly as a function of N, the number of measurements. Each curve represents a different sparsity level m. As expected, when the number of nonzero components increases, more measurements are necessary to guarantee signal recovery.

Figure 2 presents a different view of the same data. It displays the percentage of signals recovered correctly as a function of the sparsity level. We discover that, for a fixed sparsity level, the recovery probability increases as we take more measurements. This figure also exhibits a point that is important in applications. Suppose that we have only enough space to store $N = 100$ measurements or we have only enough time to measure and process $N = 100$ pieces of data. In dimension $d = 256$, we should expect to recover a signal with 16 terms in 90% of instances and a signal with 20 terms in about 50% of instances.

Pursuing this idea, let us see how many measurements are required to identify a sparse signal with a fixed rate of success. Figure 3 displays the relationship between N and m necessary to achieve a recovery probability of 95% in dimension $d = 256$. The data exhibit a clear trend $N \approx 1.5\, m \ln 256$. Table 1 examines the relationship between N and m to achieve a recovery probability of 99% in dimensions $d = 256, 1024$. For this error rate, we have $N \approx 2\, m \ln d$ in both cases. In comparison, the theoretical bound in Corollary 7 can never be better than $N \geq 4\, m \ln d$ in the Gaussian case.

$d = 256$			$d = 1024$		
m	N	$N/(m \ln d)$	m	N	$N/(m \ln d)$
4	56	2.52	5	80	2.31
8	96	2.16	10	140	2.02
12	136	2.04	15	210	2.02
16	184	2.07			
20	228	2.05			

Table 1. The number N of measurements necessary to recover an m-sparse signal in dimensions $d = 256, 1024$ at least 99% of the time.

Figure 4 provides a more graphical comparison between the theoretical bound of Theorem 6 and the empirical results. This chart matches three theoretical error curves against the corresponding empirical curves in dimension $d = 1024$. Observe that the shape of the theoretical curves is very similar to the shape of the empirical curves, even though the theoretical bounds are somewhat too pessimistic.

In the first set of experiments, we used Gaussian measurement matrices. We repeated the same body of experiments with Bernoulli measurement matrices and obtained strikingly similar results.

[1]The analysis suggests that this is a challenging case for OMP, and our experience has shown that other methods for choosing coefficients lead to similar results.

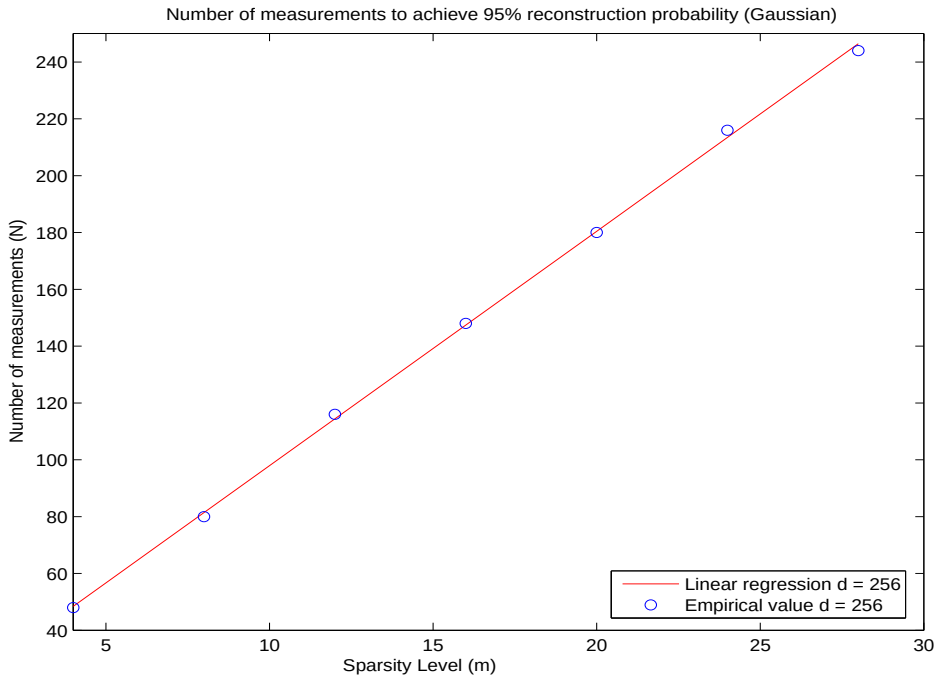

FIGURE 3. The number N of measurements necessary to recover an m-sparse signal in dimension $d = 256$ at least 95% of the time. The regression line has equation $N = 1.5\, m \ln 256 + 15.4$.

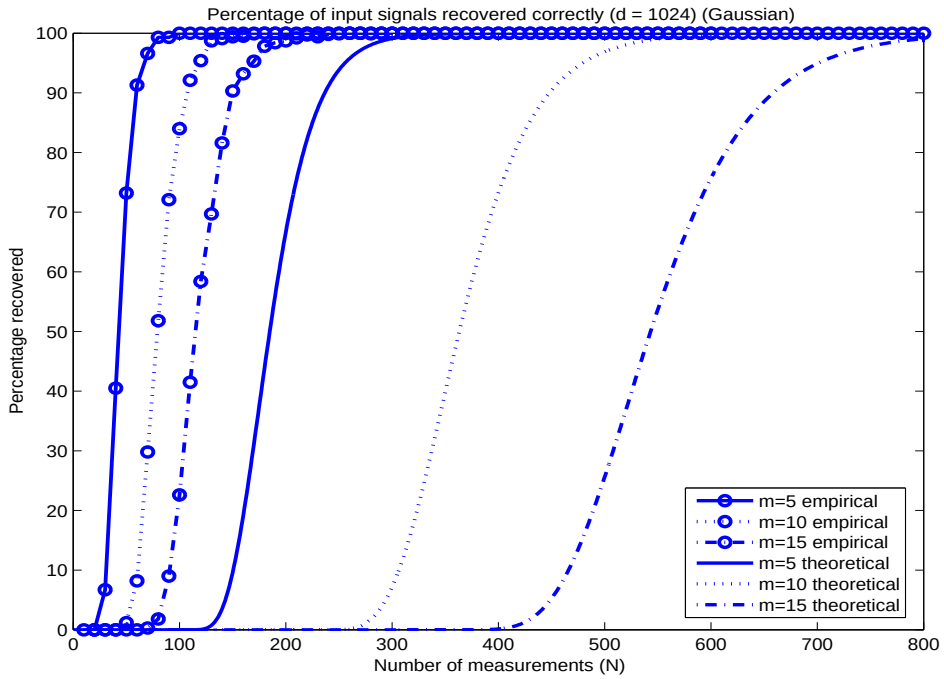

FIGURE 4. The probability of recovering an m-sparse signal in dimension $d = 1024$ from N measurements. The marked lines display empirical data, while the unmarked lines show the theoretical bounds from Theorem 6.

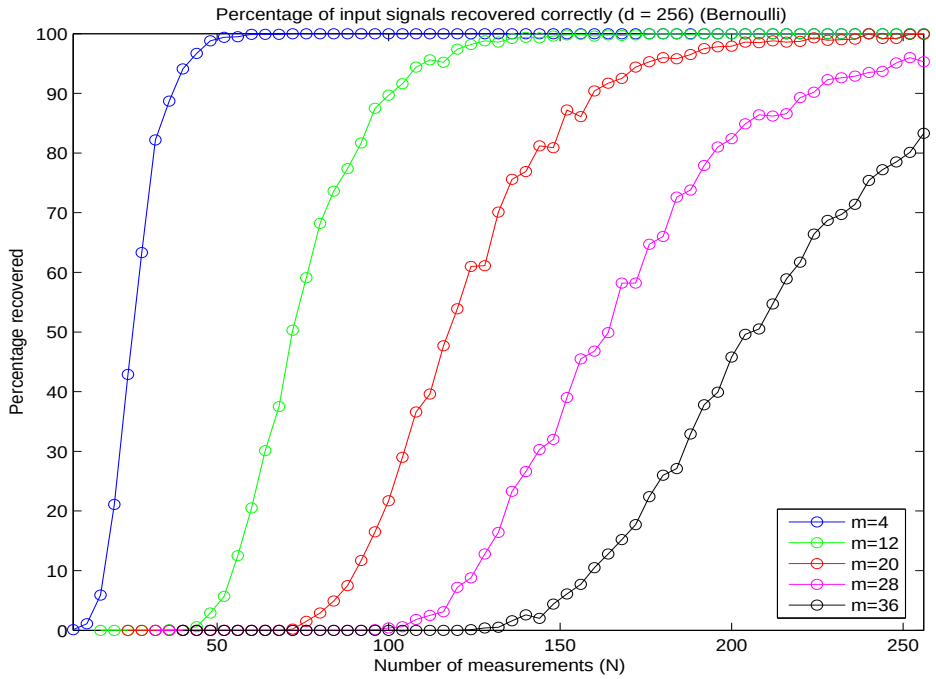

FIGURE 5. The percentage of 1000 input signals correctly recovered as a function of the number N of Bernoulli measurements for different sparsity levels m in dimension $d = 256$.

For the sake of brevity, we include just one graphic for Bernoulli measurements. Figure 5 shows the number of Bernoulli measurements necessary for OMP to recover an m-sparse signal in dimension $d = 256$. Comparing this chart with Figure 1, we discover that OMP performs almost identically with Gaussian and Bernoulli measurements.

The Introduction notes that OMP is a significantly faster algorithm than (BP). To justify this claim, we present Figure 6, which displays execution times (as opposed to processor times) for several experiments with Bernoulli measurement matrices. Timings for Gaussian matrices are similar. Let us emphasize that the chart displays the clock time required for 1000 complete trials, which includes the time to generate 1000 sparse signals and 1000 random measurement matrices *in addition to* the time required by 1000 invocations of the OMP algorithm. For the most computationally intensive experiment ($m = 64$, $N = 400$, and $d = 1024$), each trial takes an average of 0.3 seconds.

While the absolute execution time for a particular parameter setting is impossible for others to duplicate (nor is it especially meaningful), the asymptotic growth of execution time as a function of the sparsity level m, the number N of measurements, and the dimension d provides a useful and reproducible curve. The graph clearly demonstrates that the execution time grows quadratically with m. Unfortunately, we do not have enough data to determine the empirical dependence of the execution time on d and N.

6. DISCUSSION

This section addresses three major issues that arise from our work. First, we describe how the analysis might be extended to match the empirical results better. Afterward, we discuss more realistic models for input signals and the prospect of applying OMP to recover signals that are not perfectly sparse. Finally, we comment on the role of randomness in our theory and the relationship with work on the linear program (BP).

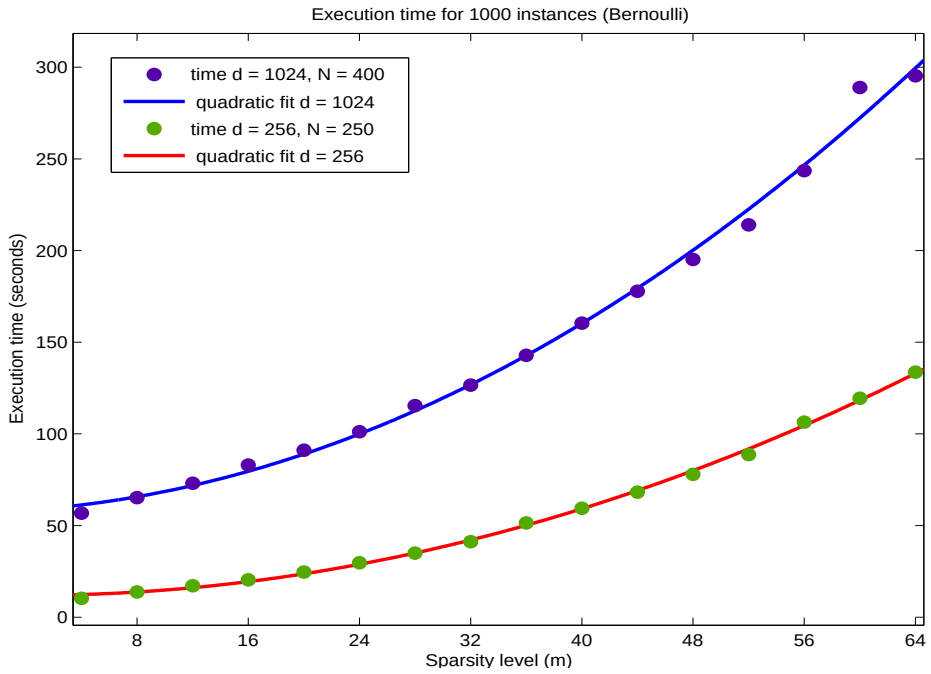

FIGURE 6. The processor time, as a function of the sparsity level m, for 1000 complete trials in dimension $d = 256, 1024$ with $N = 250, 400$ Bernoulli measurements. The regression curves are quadratic polynomials calculated with least-squares.

6.1. **Theory vs. Practice.** Although it appears that our theory correctly describes the qualitative performance of OMP for the signal recovery problem, our experiments demonstrate that the number of measurements required in practice is somewhat smaller than we predict.

Let us describe several technical reasons that the analysis is loose. The most significant problem is that the vectors $\boldsymbol{u}_t$ constructed during the analysis may have large mutual inner products. As a result, Property (M2) yields a pessimistic assessment of the maximum correlation with $\boldsymbol{\psi}$. A secondary issue is that $\|\boldsymbol{u}_t\|_2$ is somewhat smaller than one because these vectors are unlikely to be aligned with the smallest singular subspace of $\boldsymbol{\Phi}_{\text{opt}}$. It does not seem easy to account for these factors. In addition, the $\sqrt{m}$ term in the estimate for $\rho(\boldsymbol{r}_t)$ can be improved to $\sqrt{m-t}$. The effect of this change, however, seems to be minimal.

6.2. **Nonsparse Signals.** Our assumption that signals are precisely sparse is not likely to obtain in most applications. Therefore, it would be valuable to develop results for signals that are "nearly sparse" in some sense. One potential model contaminates the m-sparse signals with additive white noise. We might also consider signals whose sorted components decay in magnitude according to a power law. Candès and Tao [CT04] argue that the second model is appropriate for many types of natural signals. Of course, the correct model must match the application domain.

Unfortunately, the strategy we used to prove Theorem 6 depends heavily on the fact that the input signals are exactly sparse. When the ideal signals are not sparse, the nonoptimal columns of the matrix $\boldsymbol{\Phi}$ are statistically correlated with the residual vectors $\boldsymbol{r}_t$ generated by the algorithm. This fact creates serious difficulties in the analysis.

Nevertheless, we believe that Theorem 6 is "morally true" for nearly sparse signals. The basic reason is that the Orthogonal Matching Pursuit algorithm is stable under small perturbations of the input signal $\boldsymbol{s}$. At the moment, we have no theoretical justification of this point. In case the

number of measurements N is on the order of m^2, the claim can be established by appealing to Theorem 4.2 of [Tro04].

6.3. **Randomness.** Like computation time and storage space, randomness is an expensive resource that should be used sparingly. At present, all approaches to signal recovery using (BP) or OMP involve some degree of randomness. The goal of this subsection is to explain the role of randomness in different types of signal recovery algorithms and to discuss whether it can be evicted.

Our result for OMP, Theorem 6, requires that the measurement matrix be stochastically independent from the signal. Unfortunately, it takes $d\,N$ random bits to select a Bernoulli measurement ensemble, and a Gaussian measurement ensemble demands even more. Corollary 7 suggests one method for reducing the amount of randomness. Since the failure probability of OMP is polynomially small in the dimension d, it follows that a polynomially large collection of input signals can be recovered reliably with a single random measurement ensemble. Therefore, we can amortize the randomness over a moderately large set of input signals.

Another way to reduce the amount of randomness is to consider measurement ensembles more sophisticated than the Gaussian or Bernoulli examples. It might be possible to construct a small probability space that contains admissible measurement matrices. For example, it may be that $O(d \ln N)$ random bits would suffice to pick a satisfactory ensemble. This would represent a significant improvement over fully random measurements.

The strongest result for the linear programming method (BP) shows that a fixed measurement matrix with $N = O(m \ln(d/m))$ can be used to recover every m-sparse signal in $\mathbb{R}^d$ [RV05]. At present, all constructions of these "deterministic" ensembles are random, which is not entirely satisfactory. From an algorithmic point of view, it would be preferable to design a structured measurement ensemble that could be stored and processed efficiently. For now, it is an open question whether a truly deterministic measurement matrix exists. The literature on error-correcting codes and pseudorandom sequences contains many possibilities worth investigating.

In comparison, it is unclear whether the OMP algorithm can ensure success for a fixed measurement matrix when $N = o(m^2)$. In other words, we would like to know if an adversarial choice of the input signal can defeat OMP whenever $\mathbf{\Phi}$ is fixed and $N = o(m^2)$. The random constructions that have been developed for (BP) definitely fail on account of Theorem 3.10 of [Tro04]. Nevertheless, it remains possible that one could design a class of deterministic measurement matrices for OMP. This project would require deep facts about the geometry of convex polytopes.

Even if an adversary can defeat OMP for all deterministic measurement ensembles, this fact is unlikely to have much practical importance. We believe that, with an incoherent measurement matrix, the set of signals that confound OMP has exponentially small measure. As a result, the algorithm is unlikely to have any trouble recovering natural signals from deterministic measurements.

6.4. **Conclusions.** The theoretical and empirical work in this paper demonstrates that OMP is an effective alternative to (BP) for signal recovery from partial information. Our results offer a tremendous improvement over previous work on OMP, and they indicate that the greedy algorithm is nearly as powerful as the linear programming method. On account of the fact that OMP is faster and easier to implement, it is a natural choice for applications where storage space and processor time are limited. In other words, greed is still good.

Acknowledgments

We wish to thank Martin Strauss for his insight on algorithms, Justin Romberg for sharing his expertise on Basis Pursuit, and Michael Wakin for his perspective on the empirical results.

References

[Bal02] K. Ball. Convex geometry and functional analysis. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 161–193. Elsevier, 2002.

[BV04] S. Boyd and L. Vanderberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[CDS99] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[CRT04] E. Candès, J. Romberg, and T. Tao. Exact signal reconstruction from highly incomplete frequency information. Submitted for publication, June 2004.

[CT04] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? Submitted for publication, Nov. 2004.

[CT05] E. J. Candès and T. Tao. Decoding by linear programming. Available from `arXiv:math.MG/0502327`, Feb. 2005.

[DeV98] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[DH93] Ding-Zhu Du and Frank K. Hwang. *Combinatorial group testing and its applications*. World Scientific, 1993.

[DMA97] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *J. Constr. Approx.*, 13:57–98, 1997.

[Don04a] D. Donoho. For most large underdetermined systems of linear equations, the mimimal ℓ_1-norm solution is also the sparsest solution. Unpublished manuscript, Sept. 2004.

[Don04b] D. L. Donoho. Compressed sensing. Unpublished manuscript, Oct. 2004.

[DS02] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 317–366. Elsevier, 2002.

[DT96] R. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5:173–187, 1996.

[Fer74] X. Fernique. Regularité des trajectoires des fonctions aléatoires gausiennes. In *Ecole d'Eté de Probabilités de Saint-Flour IV*, number 480 in Lecture Notes in Mathematics, pages 1–96. Springer, 1974.

[FS81] J. H. Friedman and W. Stuetzle. Projection Pursuit Regressions. *J. Amer. Statist. Soc.*, 76:817–823, 1981.

[GVL96] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Univ. Press, 3rd edition, 1996.

[KKS95] J. Kahn, J. Komlós, and E. Szemerédi. On the probability that a random ± 1-matrix is singular. *J. Amer. Math. Soc.*, 86(1):223–240, Jan. 1995.

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, Providence, 2001.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[Mil02] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 2nd edition, 2002.

[MP67] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72:407–535, 1967. In Russian.

[MZ93] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[PRK93] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal Matching Pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proc. of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, Nov. 1993.

[RKD99] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Processing*, 47(1):187–200, 1999.

[Rom05] J. Romberg. Personal communication, Feb. 2005.

[RV05] M. Rudelson and R. Veshynin. Geometric approach to error correcting codes and reconstruction of signals. Available from `arXiv:math.MG/0502299`, Feb. 2005.

[Tem02] V. Temlyakov. Nonlinear methods of approximation. *Foundations of Comput. Math.*, July 2002.

[Tro04] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, Oct. 2004.