SIGNAL RECOVERY FROM RANDOM MEASUREMENTS 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 $\mathrm{O}(m \ln d)$ random linear measurements of that signal. This is a massive improvement over previous results for OMP, which require $\mathrm{O}(m^2)$ measurements. The new results for OMP are comparable with recent results for another algorithm called Basis Pursuit (BP). The OMP algorithm is faster and 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 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ 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:

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

where $\langle \cdot,\ \cdot \rangle$ denotes the usual inner product. The problem of *signal recovery* 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, RKD99, CDS01, 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, drawn from papers of Candès–Tao [CT05] and of Rudelson–Vershynin [RV05], describes one example of this remarkable phenomenon.

Theorem 1. *Let $N \geq \mathrm{K}m\ln(d/m)$, and draw N vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ 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 $\{\langle \boldsymbol{s}, \ \boldsymbol{x}_n \rangle : n = 1, 2, \ldots, N\}$. The numbers K and k are universal constants.*

An important detail is that a particular choice of the Gaussian measurement vectors succeeds for *every* m-sparse signal with high probability. This theorem extends earlier results of Candès–Romberg–Tao [CRT06], Donoho [Don06a], and Candès–Tao [CT04].

All five of the papers [CRT06, Don06a, 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

$$\min_{\boldsymbol{f}} \ \|\boldsymbol{f}\|_1 \quad \text{subject to} \quad \langle \boldsymbol{f}, \ \boldsymbol{x}_n \rangle = \langle \boldsymbol{s}, \ \boldsymbol{x}_n \rangle \quad \text{for } n = 1, 2, \ldots, N. \tag{BP}$$

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 researchers 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 may take a long time to solve the linear program, even for signals of moderate length.[1] Furthermore, when off-the-shelf optimization software is not available, the implementation of optimization algorithms may demand serious effort.[2] Both these problems are being studied. In the meantime, one might wish to consider alternate methods for reconstructing sparse signals from random measurements.

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. On the other hand, conventional wisdom on OMP has been pessimistic about its performance outside the simplest regimes. 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 [CDS01, Sec. 2.3.2]. Other negative results, such as Theorem 3.10 of [Tro04] and Theorem 5 of [Don06b], 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, the following theorem holds.

Theorem 2 (OMP with Gaussian Measurements). *Fix $\delta \in (0, 0.36)$, and choose $N \geq \mathrm{K}m\ln(d/\delta)$. Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$. Draw N measurement vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ independently from the standard Gaussian distribution on $\mathbb{R}^d$. Given the data $\{\langle \boldsymbol{s}, \ \boldsymbol{x}_n \rangle : n = 1, 2, \ldots, N\}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $1 - 2\delta$. For this theoretical result, it suffices that $\mathrm{K} = 20$. When m is large, it suffices to take $\mathrm{K} \approx 4$.*

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

---

[1]Although this claim qualifies as folklore, the literature does not currently offer a refutation that we find convincing.

[2]The paper [GMS05] discusses the software engineering problems that arise in optimization.

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 distinctions may not be relevant 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 s is an arbitrary m-sparse signal in $\mathbb{R}^d$, and let $\{x_1, \ldots, x_N\}$ be a family of N measurement vectors. Form an $N \times d$ matrix Φ whose *rows* are the measurement vectors, and observe that the N measurements of the signal can be collected in an N-dimensional data vector $v = \Phi s$. We refer to Φ as the *measurement matrix* and denote its columns by $\varphi_1, \ldots, \varphi_d$.

As we mentioned, it is natural to think of signal recovery as a problem dual to sparse approximation. Since s has only m nonzero components, the data vector $v = \Phi s$ is a linear combination of m columns from Φ. In the language of sparse approximation, we say that v has an m-term representation over the dictionary Φ. 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 s, we need to determine *which* columns of Φ participate in the measurement vector v. The idea behind the algorithm is to pick columns in a greedy fashion. At each iteration, we choose the column of Φ that is most strongly correlated with the remaining part of v. Then we subtract off its contribution to 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 Φ*
- *An N-dimensional data vector v*
- *The sparsity level m of the ideal signal*

OUTPUT:

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

PROCEDURE:

(1) *Initialize the residual $r_0 = v$, the index set $\Lambda_0 = \emptyset$, and the iteration counter $t = 1$.*
(2) *Find the index λ_t that solves the easy optimization problem*

$$\lambda_t = \arg\max_{j=1,\ldots,d} |\langle r_{t-1}, \varphi_j \rangle|.$$

If the maximum occurs for multiple indices, break the tie deterministically.
(3) *Augment the index set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$ and the matrix of chosen atoms $\Phi_t = \begin{bmatrix} \Phi_{t-1} & \varphi_{\lambda_t} \end{bmatrix}$. We use the convention that Φ_0 is an empty matrix.*
(4) *Solve a least-squares problem to obtain a new signal estimate:*

$$x_t = \arg\min_x \|\Phi_t x - v\|_2.$$

(5) *Calculate the new approximation of the data and the new residual:*

$$\boldsymbol{a}_t = \boldsymbol{\Phi}_t \, \boldsymbol{x}_t$$
$$\boldsymbol{r}_t = \boldsymbol{v} - \boldsymbol{a}_t.$$

(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 value of the estimate $\widehat{\boldsymbol{s}}$ in component λ_j equals the jth component of $\boldsymbol{x}_t$.*

Steps 4, 5, and 7 have been written to emphasize the conceptual structure of the algorithm; they can be implemented more efficiently. It is important to recognize that the residual $\boldsymbol{r}_t$ is always orthogonal to the columns of $\boldsymbol{\Phi}_t$. Therefore, the algorithm always selects a new atom at each step, and $\boldsymbol{\Phi}_t$ has full column rank.

The running time of the OMP algorithm is dominated by Step 2, whose total cost is $\mathrm{O}(mNd)$. At iteration t, the least-squares problem can be solved with marginal cost $\mathrm{O}(tN)$. To do so, we maintain a $\boldsymbol{QR}$ factorization of $\boldsymbol{\Phi}_t$. Our implementation uses the Modified Gram–Schmidt (MGS) algorithm because the measurement matrix is unstructured and dense. The book [Bjö96] provides extensive details and a survey of alternate approaches. When the measurement matrix is structured, more efficient implementations of OMP are possible; see the paper [KR06] for one example.

According to [NN94], there are algorithms that can solve (BP) with a dense, unstructured measurement matrix in time $\mathrm{O}(N^2 d^{3/2})$. We are focused on the case where d is much larger than m or N, so there is a substantial gap between the cost of OMP and the cost of BP.

A prototype of the OMP algorithm first appeared in the statistics community at some point in the 1950s, where it was called stagewise regression. The algorithm later 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 NORMAL$(0, N^{-1})$ distribution. For reference, the density function of this distribution is

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

(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 $j = 1, \ldots, d$.
(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

$$\mathbb{P}\left\{\max_t |\langle \boldsymbol{\varphi}, \, \boldsymbol{u}_t \rangle| \le \varepsilon\right\} \; \ge \; 1 - 2m \, \mathrm{e}^{-c\varepsilon^2 N}.$$

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

$$\mathbb{P}\left\{\sigma_{\min}(\boldsymbol{Z}) \geq 0.5\right\} \geq 1 - \mathrm{e}^{-\mathrm{c}N}.$$

We follow the analysts' convention that upright letters $(\mathrm{c}, \mathrm{C}, \mathrm{K}, \text{etc.})$ indicate positive, universal constants that may vary at each appearance.

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 $\|\boldsymbol{\varphi}_j\|_2$. Property (M3) controls how much the matrix can shrink a sparse vector.

In the two subsections, we explain why the Gaussian and Bernoulli ensembles both yield admissible measurement matrices. We make no effort to determine the precise value of the constants. See the technical report [TG06] for a detailed treatment of the Gaussian case, including explicit constants. Afterward, we compare admissible measurement matrices with other types of measurement ensembles that have appeared in the literature.

3.2. **Joint Correlation.** The joint correlation property (M2) is essentially a large deviation bound for sums of random variables. For the Gaussian and Bernoulli measurement ensembles, we can leverage classical concentration inequalities to establish this property.

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 i.i.d. NORMAL$(0, N^{-1})$ entries. Then*

$$\mathbb{P}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t \rangle| \leq \varepsilon\right\} \geq 1 - m\, \mathrm{e}^{-\varepsilon^2 N/2}.$$

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 NORMAL$(0, N^{-1})$ entries. Then the random variable $\langle \boldsymbol{z},\ \boldsymbol{u}_t \rangle$ also has the NORMAL$(0, N^{-1})$ distribution. A well-known Gaussian tail bound yields

$$\mathbb{P}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t \rangle| > \varepsilon\right\} = \sqrt{\frac{2}{\pi}} \int_{\varepsilon\sqrt{N}}^{\infty} \mathrm{e}^{-x^2/2}\, \mathrm{d}x \leq \mathrm{e}^{-\varepsilon^2 N/2}.$$

Using Boole's inequality,

$$\mathbb{P}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t \rangle| > \varepsilon\right\} \leq m\, \mathrm{e}^{-\varepsilon^2 N/2}.$$

This bound is complementary to the one stated. $\qquad\square$

For Bernoulli measurements, we simply replace the Gaussian tail bound with

$$\mathbb{P}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t \rangle| > \varepsilon\right\} \leq 2\, \mathrm{e}^{-\varepsilon^2 N/2}. \tag{3.1}$$

This is a direct application of the Hoeffding inequality. (See [Lug05], for example). For other types of measurement matrices, it may take some effort to obtain the quadratic dependence on ε. We omit a detailed discussion.

3.3. **Smallest Singular Value.** It requires more sophistication to develop the lower singular value property. We sketch a new approach due to Baraniuk et al. that combines classical arguments in a clever fashion [BDDW06].

First, observe that the Gaussian and Bernoulli ensembles both admit a Johnson–Lindenstrauss lemma, as follows. Suppose that $\mathscr{A}$ is a collection of points in $\mathbb{R}^m$. Let $\boldsymbol{Z}$ be a $N \times m$ matrix whose entries are all i.i.d. NORMAL$(0, N^{-1})$ or else i.i.d. uniform on $\{\pm 1/\sqrt{N}\}$. Then

$$0.75 \, \|\boldsymbol{a}\|_2^2 \ \leq \ \|\boldsymbol{Z}\boldsymbol{a}\|_2^2 \ \leq \ 1.25 \, \|\boldsymbol{a}\|_2^2 \qquad \text{for all } \boldsymbol{a} \in \mathscr{A}$$

except with probability

$$2 \, |\mathscr{A}| \, \mathrm{e}^{-\mathrm{c}N}.$$

This statement follows from the arguments in [Ach03].

Lemma 4.16 of [Pis89] shows that there is a subset $\mathscr{A}$ of the unit sphere in $\mathbb{R}^m$ such that

  (1) each vector on the sphere lies within ℓ_2 distance 0.125 of some point in $\mathscr{A}$ and
  (2) the cardinality of $\mathscr{A}$ does not exceed 24^m.

Consider such a covering, and apply the Johnson–Lindenstrauss lemma to it. This procedure shows that $\boldsymbol{Z}$ cannot distort the points in the covering $\mathscr{A}$ very much. A simple approximation argument shows that $\boldsymbol{Z}$ cannot distort any other point on the sphere. Working out the details leads to the following result.

Proposition 5 (Baraniuk et al.)**.** *Suppose that $\boldsymbol{Z}$ is a $N \times m$ matrix whose entries are all i.i.d. NORMAL$(0, N^{-1})$ or else i.i.d. uniform on $\{\pm 1/\sqrt{N}\}$. Then*

$$0.5 \, \|\boldsymbol{x}\|_2 \ \leq \ \|\boldsymbol{Z}\boldsymbol{x}\|_2 \ \leq \ 1.5 \, \|\boldsymbol{x}\|_2 \qquad \textit{for all } \boldsymbol{x} \in \mathbb{R}^m$$

with probability at least

$$1 - 2 \cdot 24^m \cdot \mathrm{e}^{-\mathrm{c}N}.$$

We conclude that Property (M3) holds for Gaussian and Bernoulli measurement ensembles, provided that $N \geq \mathrm{C}m$.

3.4. **Other Types of Measurement Ensembles.** It may be interesting 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 study a class of matrices whose "restricted isometry constants" are under control. They show that both Gaussian and Bernoulli matrices satisfy this property with high probability.
  • Donoho introduces the deterministic class of compressed sensing (CS) matrices [Don06a]. 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 rows of the measurement matrix span 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.

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). *Fix $\delta \in (0, 1)$, and choose $N \geq \mathrm{K}m \log(d/\delta)$ where K is an absolute constant. 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 $\boldsymbol{v} = \boldsymbol{\Phi s}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $1 - \delta$.*

For Gaussian measurements, we have obtained more precise estimates for the constant. In this case, a very similar result (Theorem 2) holds when $\mathrm{K} = 20$. Moreover, as the number m of nonzero components approaches infinity, it is possible to take $\mathrm{K} = 4 + \eta$ for any positive number η. See the technical report [TG06] for a detailed proof of these estimates.

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

Proposition 7. *Choose an arbitrary m-sparse signal $\boldsymbol{s}$ from $\mathbb{R}^d$, and let $N \geq 2m$. Suppose that $\boldsymbol{\Phi}$ is an $N \times d$ Gaussian measurement ensemble, and execute OMP with the data $\boldsymbol{v} = \boldsymbol{\Phi s}$. 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 $\boldsymbol{r}_m = \boldsymbol{0}$ 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 $2m$ columns from $\boldsymbol{\Phi}$. Since $\boldsymbol{\Phi}$ is an $N \times d$ Gaussian matrix and $2m \leq N$, 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 [Don06b, Lemma 2.1]. □

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

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

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

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

Lemma 8. *Fix $\delta \in (0, 1)$. For an $N \times d$ Bernoulli measurement matrix, the coherence statistic $\mu \leq \sqrt{4N^{-1} \ln(d/\delta)}$ with probability exceeding $1 - \delta$.*

Proof of Lemma 8. Suppose that $\boldsymbol{\Phi}$ is an $N \times d$ Bernoulli measurement matrix. For each $j < k$, the Bernoulli tail bound (3.1) establishes that

$$\mathbb{P}\{|\langle \boldsymbol{\varphi}_j, \, \boldsymbol{\varphi}_k \rangle| > \varepsilon\} \; \leq \; 2\,\mathrm{e}^{-\varepsilon^2 N/2}.$$

Choosing $\varepsilon^2 = 4N^{-1} \ln(d/\delta)$ and applying Boole's inequality,

$$\mathbb{P}\left\{\max_{j<k} |\langle \boldsymbol{\varphi}_j, \, \boldsymbol{\varphi}_k \rangle| > \sqrt{4N^{-1} \ln(d/\delta)}\right\} \; < \; d^2\,\mathrm{e}^{-2\ln(d/\delta)}.$$

This estimate completes the proof. □

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.

Proposition 9. *Fix $\delta \in (0,1)$. Let $N \geq 16m^2 \ln(d/\delta)$, and draw an $N \times d$ Bernoulli measurement matrix $\mathbf{\Phi}$. The following statement holds with probability at least $1 - \delta$. OMP can reconstruct every m-sparse signal $\boldsymbol{s}$ in $\mathbb{R}^d$ from the data $\boldsymbol{v} = \mathbf{\Phi}\boldsymbol{s}$.*

A proof sketch appears at the end of the section. Very similar coherence results hold for Gaussian matrices, but they are messier because the columns of a Gaussian matrix do not have identical norms. We prefer to omit a detailed discussion.

There are several important differences between Proposition 9 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.

Remark 10. *It is impossible to develop stronger results by way of the coherence statistic on account of the following observations. First, the coherence of a Bernoulli matrix satisfies $\mu \geq \mathrm{c}\sqrt{N^{-1}\ln d}$ with high probability. One may check this statement by using standard estimates for the size of a Hamming ball. Meanwhile, the coherence of a Gaussian matrix also satisfies $\mu \geq \mathrm{c}\sqrt{N^{-1}\ln d}$ with high probability. This argument proceeds from lower bounds for Gaussian tail probabilities.*

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 $\mathbf{\Phi} = [\mathbf{\Phi}_{\mathrm{opt}} \mid \mathbf{\Psi}]$ so that $\mathbf{\Phi}_{\mathrm{opt}}$ 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*

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

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 $\mathbf{\Phi}_{\mathrm{opt}}$ if and only if $\rho(\boldsymbol{r}) < 1$. In case $\rho(\boldsymbol{r}) = 1$, an optimal and a nonoptimal column both achieve the maximum inner product. The algorithm has no provision for choosing one instead of the other, so we assume that the algorithm always fails. The greedy selection ratio was first identified and studied in [Tro04].

Imagine that we could execute m iterations of OMP with the input signal $\boldsymbol{s}$ and the restricted measurement matrix $\mathbf{\Phi}_{\mathrm{opt}}$ to obtain a sequence of residuals $\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_{m-1}$ and a sequence of column indices $\omega_1, \omega_2, \ldots, \omega_m$. The algorithm is deterministic, so these sequences are both functions of $\boldsymbol{s}$ and $\mathbf{\Phi}_{\mathrm{opt}}$. In particular, the residuals are stochastically independent from $\mathbf{\Psi}$. It is also evident that each residual lies in the column span of $\mathbf{\Phi}_{\mathrm{opt}}$.

Execute OMP with the input signal $\boldsymbol{s}$ and the full matrix $\mathbf{\Phi}$ to obtain the actual sequence of residuals $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ and the actual sequence of column indices $\lambda_1, \lambda_2, \ldots, \lambda_m$. Observe that OMP succeeds in reconstructing $\boldsymbol{s}$ after m iterations if and only if the algorithm selects the first m columns of $\mathbf{\Phi}$ in some order. We use induction to prove that success occurs if and only if $\rho(\boldsymbol{q}_t) < 1$ for each $t = 0, 1, \ldots, m - 1$.

In the first iteration, OMP chooses one of the optimal columns if and only if $\rho(\boldsymbol{r}_0) < 1$. The algorithm sets the initial residual equal to the input signal, so $\boldsymbol{q}_0 = \boldsymbol{r}_0$. Therefore, the success criterion is identical with $\rho(\boldsymbol{q}_0) < 1$. It remains to check that λ_1, the actual column chosen,

matches ω_1, the column chosen in our thought experiment. Because $\rho(\boldsymbol{r}_0) < 1$, the algorithm selects the index λ_1 of the column from $\boldsymbol{\Phi}_{\mathrm{opt}}$ whose inner product with $\boldsymbol{r}_0$ is largest (ties being broken deterministically). Meanwhile, ω_1 is defined as the column of $\boldsymbol{\Phi}_{\mathrm{opt}}$ whose inner product with $\boldsymbol{q}_0$ is largest. This completes the base case.

Suppose that, during the first k iterations, the actual execution of OMP chooses the same columns as our imaginary invocation of OMP. That is, $\lambda_j = \omega_j$ for $j = 1, 2, \ldots, k$. Since the residuals are calculated using only the original signal and the chosen columns, it follows that $\boldsymbol{r}_k = \boldsymbol{q}_k$. Repeating the argument in the last paragraph, we conclude that the algorithm identifies an optimal column if and only if $\rho(\boldsymbol{q}_k) < 1$. Moreover, it must select $\lambda_{k+1} = \omega_{k+1}$.

In consequence, the event on which the algorithm succeeds is

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

where $\{\boldsymbol{q}_t\}$ is a sequence of m random vectors that fall in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$ and that are stochastically independent from $\boldsymbol{\Psi}$. We can decrease the probability of success by placing the additional requirement that the smallest singular value of $\boldsymbol{\Phi}_{\mathrm{opt}}$ meet a lower bound:

$$\mathbb{P}\left(E_{\mathrm{succ}}\right) \geq \mathbb{P}\{\max_t \rho(\boldsymbol{q}_t) < 1 \quad \text{and} \quad \sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}}) \geq 0.5\}.$$

We will use Σ to abbreviate the event $\{\sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}}) \geq 0.5\}$. Applying the definition of conditional probability, we reach

$$\mathbb{P}\left(E_{\mathrm{succ}}\right) \geq \mathbb{P}\{\max_t \rho(\boldsymbol{q}_t) < 1 \mid \Sigma\} \cdot \mathbb{P}\left(\Sigma\right). \tag{4.1}$$

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

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

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

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

$$\rho(\boldsymbol{q}_t) \leq \frac{\sqrt{m} \, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi}, \ \boldsymbol{q}_t \rangle|}{\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^{\mathrm{T}} \boldsymbol{q}_t\right\|_2}.$$

To simplify this expression, define the vector

$$\boldsymbol{u}_t \stackrel{\mathrm{def}}{=} \frac{0.5 \, \boldsymbol{q}_t}{\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^{\mathrm{T}} \boldsymbol{q}_t\right\|_2}.$$

The basic properties of singular values furnish the inequality

$$\frac{\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^{\mathrm{T}} \boldsymbol{q}\right\|_2}{\left\|\boldsymbol{q}\right\|_2} \geq \sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}}) \geq 0.5$$

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

$$\rho(\boldsymbol{q}_t) \leq 2\sqrt{m} \, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi}, \ \boldsymbol{u}_t \rangle|$$

for each index t. On account of this fact,

$$\mathbb{P}\{\max_t \rho(\boldsymbol{q}_t) < 1 \mid \Sigma\} \geq \mathbb{P}\left\{\max_t \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi}, \ \boldsymbol{u}_t \rangle| < \frac{1}{2\sqrt{m}} \, \middle| \, \Sigma\right\}.$$

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

$$\mathbb{P}\{\max_t \rho(\boldsymbol{q}_t) < 1 \mid \Sigma\} \geq \prod_{\boldsymbol{\psi}} \mathbb{P}\left\{\max_t |\langle \boldsymbol{\psi}, \ \boldsymbol{u}_t \rangle| < \frac{1}{2\sqrt{m}} \, \middle| \, \Sigma\right\}.$$

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

$$\mathbb{P}\{\max_t \rho(\boldsymbol{q}_t) < 1 \mid \Sigma\} \ \geq \ \left[1 - 2m\,\mathrm{e}^{-cN/4m}\right]^{d-m}.$$

Property (M3) furnishes a bound on $\mathbb{P}(\Sigma)$, namely

$$\mathbb{P}\{\sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}}) \geq 0.5\} \ \geq \ 1 - \mathrm{e}^{-cN}.$$

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

$$\mathbb{P}(E_{\mathrm{succ}}) \ \geq \ \left[1 \ - \ 2m\,\mathrm{e}^{-cN/4m}\right]^{d-m} \left[1 \ - \ \mathrm{e}^{-cN}\right].$$

To complete the argument, we need to make some estimates. Apply the inequality $(1-x)^k \geq 1 - kx$, valid for $k \geq 1$ and $x \leq 1$. This step delivers

$$\mathbb{P}(E_{\mathrm{succ}}) \ \geq \ 1 - 2m(d-m)\,\mathrm{e}^{-cN/4m} - \mathrm{e}^{-cN}.$$

It holds that $m(d-m) \leq d^2/4$. Next, absorb the third term into the second term, altering the constants if necessary. We see that

$$\mathbb{P}(E_{\mathrm{succ}}) \ \geq \ 1 - d^2\,\mathrm{e}^{-cN/m}.$$

In conclusion, the choice $N \geq \mathrm{K}m\log(d/\delta)$ is sufficient to reduce the failure probability below δ.

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.[3] We draw an $N \times d$ Gaussian measurement matrix $\boldsymbol{\Phi}$ and execute OMP with the data vector $\boldsymbol{v} = \boldsymbol{\Phi s}$. 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 another 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 $N \approx 1.5m \ln 256$. 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 $N \approx 2m \ln d$ in both cases. In comparison,

---

[3]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.
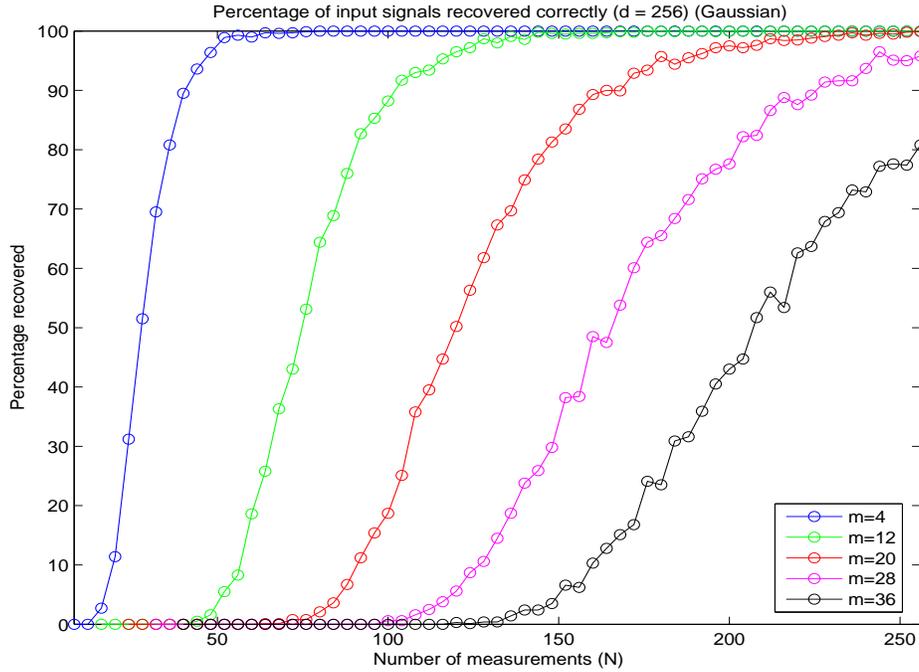
FIGURE 1. The percentage of 1000 input signals correctly recovered as a function of the number N of measurements for different sparsity levels m in dimension $d = 256$.
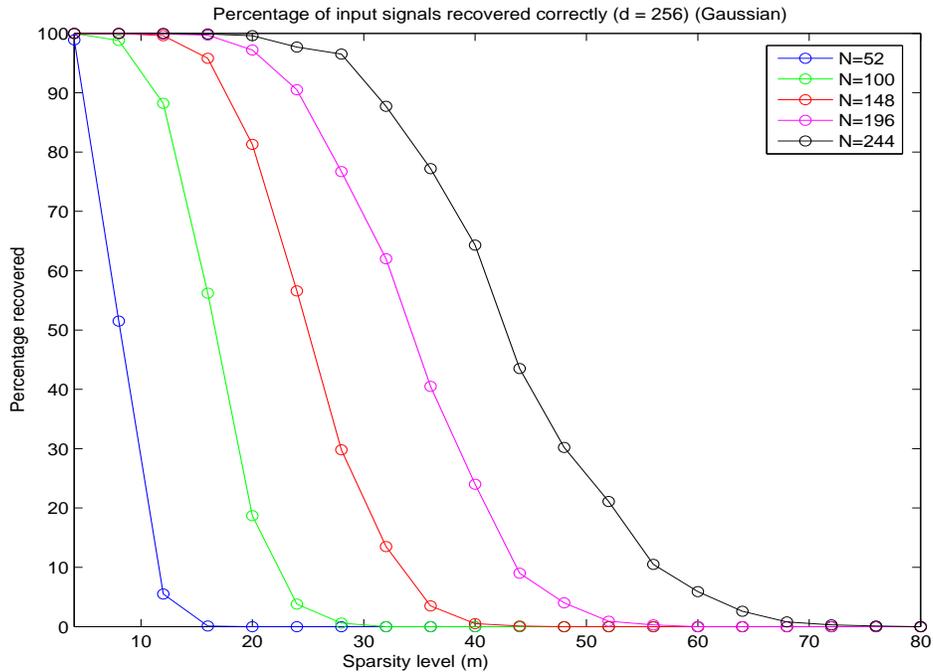

FIGURE 2. The percentage of 1000 input signals correctly recovered as a function of the sparsity level m for different numbers N of measurements in dimension $d = 256$.

our best theoretical bound for the Gaussian case is about $N \geq 4m(\ln d + 4.6)$ if we want a 99%
probability of success.

$d = 256$			$d = 1024$		
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 graphical comparison between the empirical results and theoretical bounds
from the technical report [TG06]. 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.
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.

To deliver some intuition about the execution cost off running OMP, 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 0.20 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 linearly 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 several 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. Next, we comment on the role of randomness in our theory. Then, we present a
recent result of Kunis and Rauhut that addresses some of the shortcomings of the present paper.
Finally, we discuss the relationship between our work and results on the linear program (BP).
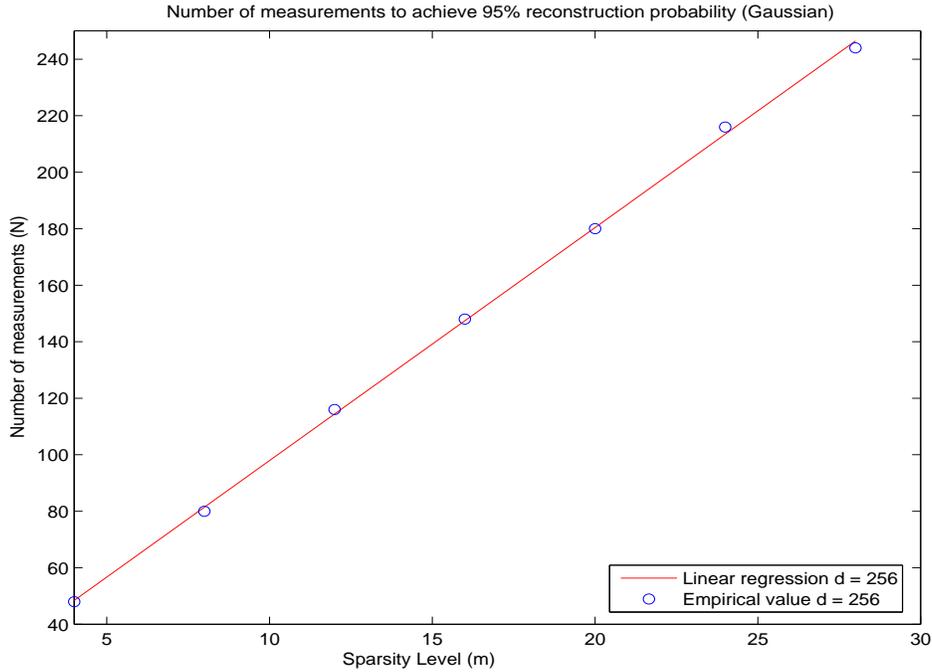
FIGURE 3. The number N of measurements necessary to recover an m-sparse signal in dimension $d = 256$ at least 95% of the time. The regression line has equation $N = 1.5\, m \ln 256 + 15.4$.
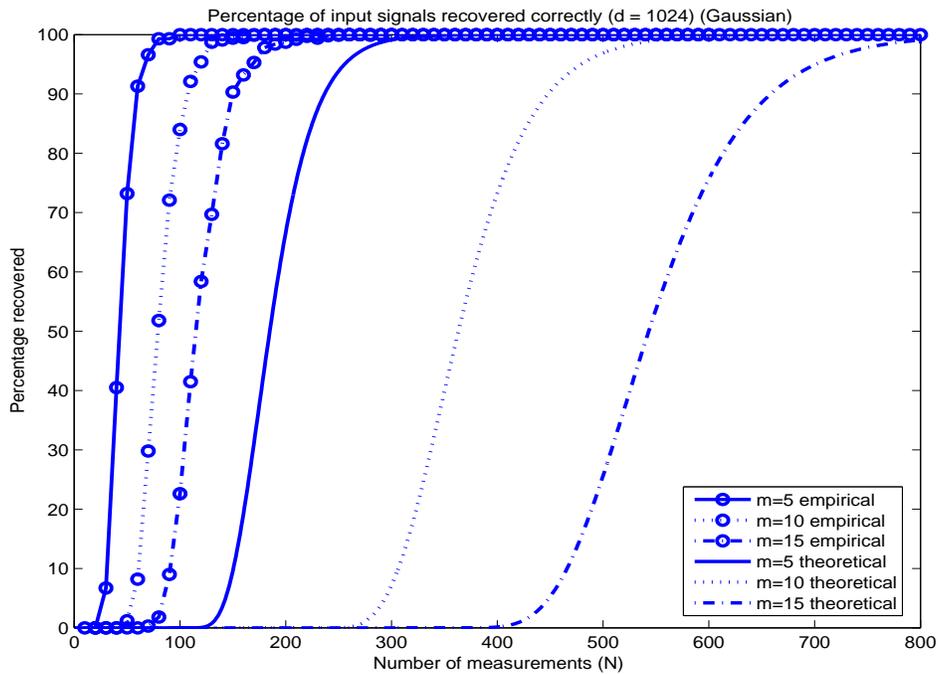

FIGURE 4. The probability of recovering an m-sparse signal in dimension $d = 1024$ from N measurements. The marked lines display empirical data, while the unmarked lines show the theoretical bounds from Theorem 6 of [TG06].
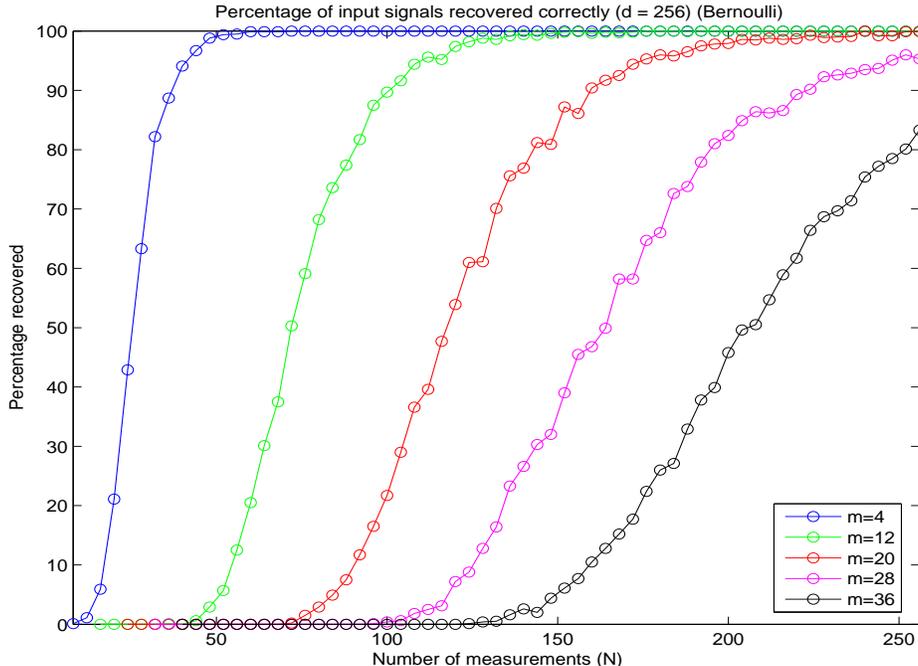
FIGURE 5. The percentage of 1000 input signals correctly recovered as a function of the number N of Bernoulli measurements for different sparsity levels m in dimension $d = 256$.

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 $\boldsymbol{\Phi}_{\mathrm{opt}}$. It does not seem easy to account for these factors. In addition, the $\sqrt{m}$ term in the estimate for $\rho(\boldsymbol{q}_t)$ 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 $\boldsymbol{q}_t$ generated by the imaginary invocation of the algorithm. This fact creates serious difficulties in the analysis.

The literature does contain a body of results on the stability of OMP for nonsparse signals. For example, Theorem 5.3 of [TGS06] can be used to establish that OMP identifies signal components above the noise level, provided that the number of measurements N is on the order of $m^2 \ln d$. It
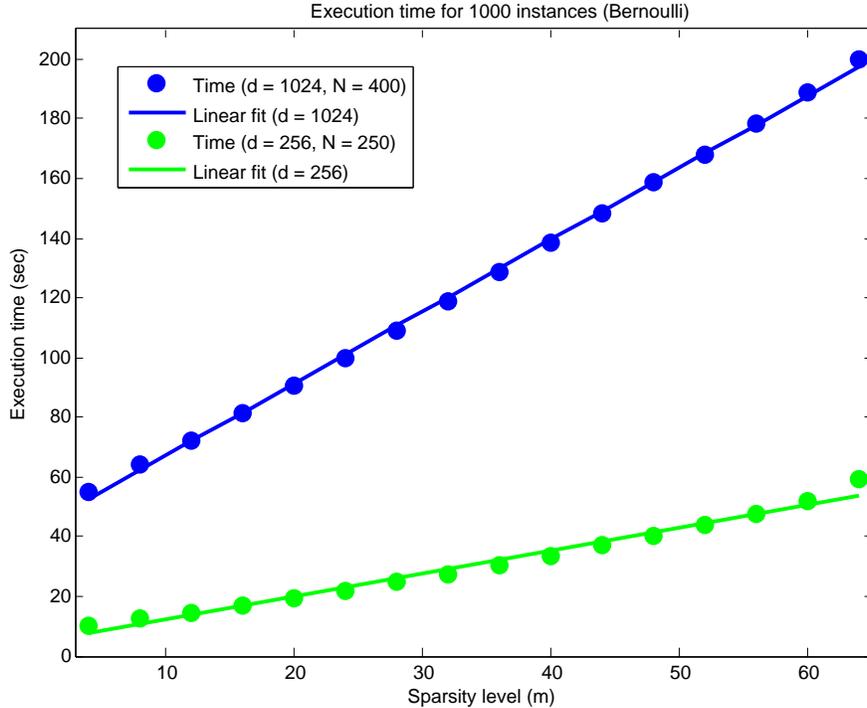
FIGURE 6. The processor time, as a function of the sparsity level m, for 1000 complete trials in dimension $d = 256, 1024$ with $N = 250, 400$ Bernoulli measurements. The regression curves are linear polynomials calculated with least-squares.

is likely that a stability result also holds in the same regime as Theorem 6. At present, we do not know exactly what such a result should look like, nor do we know a promising method of proof.

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. For now, it is an open question whether a truly deterministic measurement matrix exists for any (stable) recovery algorithm.

Our result for OMP, Theorem 6, requires that the measurement matrix be stochastically independent from the signal. Unfortunately, it takes dN random bits to select a Bernoulli measurement ensemble, and a Gaussian measurement ensemble demands even more. 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. Still, this amount of randomness is far from ideal.

6.4. **OMP with Frequency Measurements.** The work in this paper focuses on rather generic measurement ensembles, such as Bernoulli and Gaussian matrices. From an algorithmic point of view, it would be preferable to design a structured measurement ensemble that could be stored and processed efficiently. For this reason, the literature on BP promotes the use of random frequency measurements. Very recently, Kunis and Rauhut have demonstrated that OMP can also identify sparse signals from random frequency data [KR06].

Theorem 11 (Kunis–Rauhut). *Fix $\delta \in (0, 1)$. Let $N \geq \mathrm{K}m \ln(d/\delta)$, and suppose that Ω is a subset of $\{1, 2, \ldots, d\}$ with cardinality m. Independent from Ω, draw a matrix $\boldsymbol{\Phi}$ consisting of N random rows from the d-dimensional DFT matrix. With probability at least $1 - \delta$, OMP can recover every vector $\boldsymbol{s}$ supported on Ω from the data $\boldsymbol{v} = \boldsymbol{\Phi s}$.*

This result is substantially stronger than Theorem 6 because it demonstrates that OMP can recover an entire m-dimensional subspace of signals with high probability. No analogous theorem holds for Bernoulli or Gaussian measurements, as one may infer from Theorem 3.10 of [Tro04]. Another advantage of this result is that the matrix $\mathbf{\Phi}$ requires only $\mathrm{O}(N \ln d)$ random bits to construct and a similar amount of storage space. Moreover, the recovery algorithm can be accelerated substantially using fast Fourier transforms.

The proof of Theorem 11 combines delicate moment calculations inspired by [CRT06] with some simple observations about OMP from [Tro04]. As an added bonus, this type of argument also yields some stability results [Rau06]. It remains possible that random frequency measurements allow us to recover all m-sparse signals via OMP. If so, the argument is likely to involve profound methods from Banach space geometry.

6.5. **Basis Pursuit.** This subsection offers a brief comparison between known results for the greedy algorithm and results for the convex relaxation approach.

First, note that there are situations where (BP) is provably more powerful than OMP. For example, with a Gaussian or Bernoulli measurement matrix, (BP) can recover all sparse signals simultaneously with high probability. In the same setting, OMP recovers each sparse signal with high probability but with high probability fails to recover all sparse signals.

It is not clear, however, that convex relaxation is always more powerful. OMP is especially effective at recovering signals from random frequency measurements. The theoretical results for (BP) are still stronger, but it remains possible that we can close this gap.

Since OMP is inherently more difficult to analyze than (BP), the literature on the convex relaxation approach also contains a richer variety of results. Right now, we understand the stability of (BP) much better than the stability of OMP. More research in this direction would be valuable.

Greedy pursuit gains the advantage when we ask about computational cost. In most parameter regimes, OMP is faster than standard approaches for completing the minimization (BP). We hope that new algorithms will eventually allow us to solve the ℓ_1 problem more rapidly.

6.6. **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 significantly narrow the gap between the theoretical performance of the greedy algorithm and the linear programming approach. 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

References

[Ach03]   D. Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comp. Sys. Sci.*, 66(4):671–687, 2003. Special issue of invited papers from PODS'01.

[BDDW06] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. The Johnson–Lindenstrauss lemma meets Compressed Sensing. Submitted for publication, May 2006.

[Bjö96]   Å. Björck. *Numerical Methods for Least Squares Problems.* SIAM, 1996.

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

[CDS01]   S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[CRT06]   E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Trans. Info. Theory*, 52(2):489–509, Feb. 2006.

[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. *IEEE Trans. Info. Theory*, 51(12):4203–4215, Dec. 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.

[Don06a]   D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, Apr. 2006.

[Don06b]   D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.

[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.

[GMS05]   P. E. Gill, W. Murray, and M. A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Rev.*, 47(1):99–132, March 2005.

[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.

[KR06]   S. Kunis and H. Rauhut. Random sampling of sparse trigonometric polynomials II: Orthogonal Matching Pursuit versus Basis Pursuit. Submitted for publication, 2006.

[Lug05]   G. Lugosi. Concentration of measure inequalities. Lecture notes, 2005.

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

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

[NN94]   Y. E. Nesterov and A. S. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[Pis89]   G. Pisier. *The Volume of Convex Bodies and the Geometry of Banach Spaces*. Cambridge Univ. Press, 1989.

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

[Rau06]   H. Rauhut. Stability results for random sampling of sparse trigonometric polynomials. Submitted for publication, 2006.

[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.

[RV05]   M. Rudelson and R. Veshynin. Geometric approach to error correcting codes and reconstruction of signals. *Int. Math. Res. Not.*, 64:4019–4041, 2005.

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

[TG06]   J. A. Tropp and A. C. Gilbert. Signal recovery from random measurement via Orthogonal Matching Pursuit: The gaussian case. Mathematics dept. technical report, Univ. Michigan at Ann Arbor, Nov. 2006.

[TGS06]   J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *J. Signal Process.*, 86:572–588, April 2006. Special issue, "Sparse approximations in signal and image processing".

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