Signal Recovery from Random Measurements via Orthogonal Matching Pursuit

Joel A. Tropp, *Member, IEEE*, 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, which require $\mathrm{O}(m^2)$ measurements. The new results for OMP are comparable with recent results for another approach called Basis Pursuit (BP). In some settings, the OMP algorithm is faster and easier to implement, so it is an attractive alternative to BP for signal recovery problems.**

***Index Terms*—Algorithms, approximation, Basis Pursuit, Compressed Sensing, group testing, Orthogonal Matching Pursuit, signal recovery, sparse approximation**

I. 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 [1], [2], [3], [4], [5].

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 the nonzero 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 [6] for a survey.

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 the papers of Candès–Tao [7] and Rudelson–Vershynin [8], 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 $\{\langle \boldsymbol{s},\ \boldsymbol{x}_n \rangle : n = 1, 2, \ldots, N\}$.

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

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 [9], Donoho [10], and Candès–Tao [11].

All five of the papers [9], [10], [11], [8], [7] offer constructive demonstrations of the recovery phenomenon by proving that the original signal $\boldsymbol{s}$ is the unique solution to the mathematical 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 can be recast as an ordinary linear program using standard transformations, and it suggests an answer to our second question about algorithms for reconstructing 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. In reality, commercial optimization packages tend not to work very well for sparse signal recovery because the solution vector is sparse and the measurement

Date: 11 April 2005. Revised 8 November 2006 and 15 August 2007.

JAT was with the Department of Mathematics, The University of Michigan at Ann Arbor, 530 Church St., Ann Arbor, MI 48109-1043. He is now with Applied and Computational Mathematics, MC 217-50, The California Institute of Technology, Pasadena, CA 91125. E-mail: jtropp@acm.caltech.edu.

ACG is with the Department of Mathematics, The University of Michigan at Ann Arbor, 530 Church St., Ann Arbor, MI 48109-1043. E-mail: annacg@umich.edu.

JAT was supported by NSF DMS 0503299 and ACG was supported by NSF DMS 0354600.

0000–0000/00$0.00 © 2007 IEEE

matrix is dense. Instead it is necessary to apply specialized techniques.

The literature describes a bewildering variety of algorithms that perform signal recovery by solving (BP) or a related problem. These methods include [3], [12], [13], [14], [15], [16]. The algorithms range widely in effectiveness, (empirical) computational cost, and implementation complexity. Unfortunately, there is little guidance available on choosing a good technique for a given parameter regime.

As a result, it seems valuable to explore alternative approaches that are not based on optimization. Thus, we adapted a sparse approximation algorithm called Orthogonal Matching Pursuit (OMP) [17], [18] to handle the signal recovery problem. The major advantages of this algorithm are its speed and its ease of implementation. On the other hand, conventional wisdom on OMP has been pessimistic about its performance outside the simplest settings. A notable instance of this complaint appears in a 1996 paper of DeVore and Temlyakov [19]. 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 [3, Sec. 2.3.2]. Other negative results, such as Theorem 3.10 of [20] and Theorem 5 of [21], echo this concern.

But these negative results about OMP are deceptive. Indeed, the empirical evidence suggests that OMP can recover an m-sparse signal when the number of measurements N is nearly proportional to 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 $1 - 2\delta$. The constant satisfies $\mathrm{K} \leq 20$. For large values of m, it can be reduced to $\mathrm{K} \approx 4$.

In comparison, earlier positive results, such as Theorem 3.6 from [20], only demonstrate that Orthogonal Matching Pursuit can recover m-sparse signals when the number of measurements N is roughly 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. We argue in Section VI that Orthogonal Matching Pursuit remains nevertheless a valuable tool. Indeed, we believe that the advantages of Orthogonal Matching Pursuit make Theorem 2 extremely compelling.

II. Orthogonal Matching Pursuit for Signal Recovery

This section describes how to apply a fundamental algorithm from sparse approximation to the signal recovery problem. Suppose that $\boldsymbol{s}$ is an arbitrary m-sparse signal in $\mathbb{R}^d$, and let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ 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 $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_d$.

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 $\boldsymbol{\Phi}$.

Therefore, sparse approximation algorithms can be used for recovering sparse signals. 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 $\{1, \ldots, d\}$
- An N-dimensional approximation $\boldsymbol{a}_m$ of the data $\boldsymbol{v}$
- 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
$$\lambda_t \ = \ \arg\max_{j=1,\ldots,d} \left|\langle \boldsymbol{r}_{t-1},\ \boldsymbol{\varphi}_j \rangle\right|.$$
If the maximum occurs for multiple indices, break the tie deterministically.
3) Augment the index set and the matrix of chosen atoms: $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$ and $\boldsymbol{\Phi}_t = \begin{bmatrix}\boldsymbol{\Phi}_{t-1} & \boldsymbol{\varphi}_{\lambda_t}\end{bmatrix}$. We use the convention that $\boldsymbol{\Phi}_0$ is an empty matrix.
4) Solve a least-squares problem to obtain a new signal estimate:
$$\boldsymbol{x}_t \ = \ \arg\min_{\boldsymbol{x}} \left\|\boldsymbol{v} - \boldsymbol{\Phi}_t\, \boldsymbol{x}\right\|_2.$$
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 $\boldsymbol{\Phi}_t$. Provided that the residual $\boldsymbol{r}_{t-1}$ is nonzero, the algorithm selects a new atom at iteration t and the matrix $\boldsymbol{\Phi}_t$ has full column rank. In which case the solution $\boldsymbol{x}_t$ to the least-squares problem in Step 4 is unique. (It should be noted that the approximation and residual calculated in Step 5 are always uniquely determined.)

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 [22] 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 [23] for one example.

According to [24], there are algorithms that can solve (BP) with a dense, unstructured measurement matrix in time $\mathrm{O}(N^2d^{3/2})$. We focus on the case where d is much larger than m or N, so there is a substantial gap between the theoretical cost of OMP and the cost of BP. We compare their empirical costs in Section VI

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 [1], [17], [18] and approximation theory [25], [5] literatures.

III. 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 p of this distribution is
$$p(x) \;=\; \frac{1}{\sqrt{2\pi N}}\,\mathrm{e}^{-x^2N/2} \quad \text{for } x \in \mathbb{R}.$$
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 abstract these properties even though we are primarily interested in the foregoing examples.

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

(M3) Smallest singular value: For a given $N \times m$ submatrix $\boldsymbol{Z}$ from $\boldsymbol{\Phi}$, the mth largest singular value $\sigma_m(\boldsymbol{Z})$ satisfies
$$\mathbb{P}\left\{\sigma_m(\boldsymbol{Z}) \ge 0.5\right\} \;\ge\; 1 - \mathrm{e}^{-\mathrm{c}N}.$$

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 tail behavior of the random variables $\|\boldsymbol{\varphi}_j\|_2$. Property (M3) controls how much the matrix is likely to shrink a sparse vector.

In the two subsections below, 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 [26] 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.

B. 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. $\textsc{normal}(0, N^{-1})$ entries. Then
$$\mathbb{P}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| \le \varepsilon\right\} \;\ge\; 1 - m\,\mathrm{e}^{-\varepsilon^2 N/2}.$$

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 i.i.d. $\textsc{normal}(0, N^{-1})$ entries. Then the random variable $\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle$ also has the $\textsc{normal}(0, N^{-1})$ distribution. A well-known Gaussian tail bound (see [27, p. 118] for example) yields
$$\mathbb{P}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| > \varepsilon\right\} \;=\; \sqrt{\frac{2}{\pi}}\int_{\varepsilon\sqrt{N}}^{\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \;\le\; \mathrm{e}^{-\varepsilon^2 N/2}.$$
Owing to Boole's inequality,
$$\mathbb{P}\left\{\max_t |\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| > \varepsilon\right\} \;\le\; m\,\mathrm{e}^{-\varepsilon^2 N/2}.$$
This bound is complementary to the one stated. ∎

For Bernoulli measurements, we simply replace the Gaussian tail bound with
$$\mathbb{P}\left\{|\langle \boldsymbol{z},\ \boldsymbol{u}_t\rangle| > \varepsilon\right\} \;\le\; 2\,\mathrm{e}^{-\varepsilon^2 N/2}. \tag{III.1}$$

This is a direct application of the Hoeffding inequality. (See [28] 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.

C. Smallest Singular Value

It requires more sophistication to develop the lower singular value property. Using a clever combination of classical arguments, Baraniuk et al. establish the following result [29].

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

$$0.5 \|\boldsymbol{x}\|_2 \;\le\; \|\boldsymbol{Z}\boldsymbol{x}\|_2 \;\le\; 1.5 \|\boldsymbol{x}\|_2 \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^m$$

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 $N \ge \mathrm{C}m$.

D. 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 one of their papers [11], 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 another paper [7], 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 [10]. He shows that Gaussian random matrices fall in this class with high probability.
- The approach in Rudelson and Vershynin's paper [8] 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.

IV. 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 $\delta \in (0, 0.36)$, and choose $N \ge \mathrm{K}m\log(d/\delta)$ 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 $\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{s}$, 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 with $\mathrm{K} \le 20$. Moreover, when the number m of nonzero components approaches infinity, it is possible to take $\mathrm{K} \le 4 + \eta$ for any positive number η. See the technical report [26] 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 $N \ge 2m$. Suppose that $\boldsymbol{\Phi}$ is an $N \times d$ 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, 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 $\widehat{\boldsymbol{s}} \ne \boldsymbol{s}$, 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 $2m \le N$, this event occurs with probability zero. Geometrically, this observation is equivalent with the fact that independent Gaussian vectors lie in general position with probability one. This claim follows from the zero–one law for Gaussian processes [30, Sec. 1.2]. The kernel of our argument originates in [21, 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 [31].

A. 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:

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

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 $\mu \le \sqrt{4N^{-1}\ln(d/\delta)}$ with probability exceeding $1 - \delta^2$.

Proof: Suppose that $\boldsymbol{\Phi}$ is an $N \times d$ Bernoulli measurement matrix. For each $j < k$, the Bernoulli tail bound (III.1) establishes that

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

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

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

This estimate completes the proof. ∎

A typical coherence result for OMP, such as Theorem 3.6 of [20], 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 $1 - \delta^2$. OMP can reconstruct every m-sparse signal $\boldsymbol{s}$ in $\mathbb{R}^d$ from the data $\boldsymbol{v} = \mathbf{\Phi}\boldsymbol{s}$.

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 $\mu \geq \sqrt{\mathrm{c}N^{-1} \ln d}$ 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 $\mu \geq \sqrt{\mathrm{c}N^{-1} \ln d}$ with high probability. This argument proceeds from lower bounds for Gaussian tail probabilities.

B. Proof of Theorem 6

Most of the argument follows the approach developed in [20]. The main difficulty here is to deal with the nasty independence issues that arise in the random 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 data 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 the vector $\boldsymbol{v} = \mathbf{\Phi}\boldsymbol{s}$ is statistically independent from the random matrix $\mathbf{\Psi}$.

Consider the event E_{succ} where the algorithm correctly identifies the signal $\boldsymbol{s}$ after m iterations. We only decrease the probability of success if we impose the additional requirement that the smallest singular value of $\mathbf{\Phi}_{\mathrm{opt}}$ meet a lower bound. To that end, define the event

$$\Sigma \stackrel{\mathrm{def}}{=} \{\sigma_m(\mathbf{\Phi}_{\mathrm{opt}}) \geq 0.5\}.$$

Applying the definition of conditional probability, we reach

$$\begin{aligned} \mathbb{P}\left(E_{\mathrm{succ}}\right) &\geq \mathbb{P}\left(E_{\mathrm{succ}} \cap \Sigma\right) \\ &= \mathbb{P}\left(E_{\mathrm{succ}} \mid \Sigma\right) \cdot \mathbb{P}\left(\Sigma\right). \end{aligned} \tag{IV.1}$$

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

To prove that E_{succ} occurs conditional on Σ, it suffices to check that the algorithm correctly identifies the columns of $\mathbf{\Phi}_{\mathrm{opt}}$. These columns determine *which* entries of the signal are nonzero. The *values* of the nonzero entries are determined by solving a least-squares problem, which has a unique solution because the event Σ implies that $\mathbf{\Phi}_{\mathrm{opt}}$ has full column rank. In other words, there is just one explanation for the signal $\boldsymbol{s}$ using the columns in $\mathbf{\Phi}_{\mathrm{opt}}$.

Now we may concentrate on showing that the algorithm locates the columns of $\mathbf{\Phi}_{\mathrm{opt}}$. For a vector $\boldsymbol{r}$ in $\mathbb{R}^N$, define the *greedy selection ratio*

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

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 $\mathbf{\Phi}_{\mathrm{opt}}$ whenever $\rho(\boldsymbol{r}) < 1$. In case $\rho(\boldsymbol{r}) = 1$, an optimal and a nonoptimal column both achieve the maximum inner product. The algorithm has no cause to prefer one over the other, so we cannot be sure it chooses correctly. The greedy selection ratio was first isolated and studied in [20].

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 statistically 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 $\lambda_1, \lambda_2, \ldots, \lambda_m$. Conditional on Σ, OMP succeeds in reconstructing $\boldsymbol{s}$ after m iterations if and only if the algorithm selects the m columns of $\mathbf{\Phi}_{\mathrm{opt}}$ in some order. We use induction to prove that this situation occurs when $\rho(\boldsymbol{q}_t) < 1$ for each $t = 0, 1, \ldots, m-1$.

The statement of the algorithm ensures that the initial residuals satisfy $\boldsymbol{q}_0 = \boldsymbol{r}_0$. Clearly, the condition $\rho(\boldsymbol{q}_0) < 1$ ensures $\rho(\boldsymbol{r}_0) < 1$. It follows that the actual invocation chooses the column λ_1 from $\mathbf{\Phi}_{\mathrm{opt}}$ whose inner product with $\boldsymbol{r}_0$ has the largest magnitude (ties broken deterministically). Meanwhile, the imaginary invocation chooses the column ω_1 from $\mathbf{\Phi}_{\mathrm{opt}}$ whose inner product with $\boldsymbol{q}_0$ has largest magnitude. Evidently, $\lambda_1 = \omega_1$. This observation completes the base case.

Suppose that, during the first k iterations, the actual execution of OMP chooses the same columns as the imaginary execution. That is, $\lambda_j = \omega_j$ for $j = 1, 2, \ldots, k$. Since the algorithm calculates the new residual as the (unique) best approximation of the signal $\boldsymbol{s}$ from the span of the chosen columns, the actual and imaginary residuals must be identical at the beginning of iteration k. In symbols, $\boldsymbol{r}_k = \boldsymbol{q}_k$. An obvious consequence is that $\rho(\boldsymbol{q}_k) < 1$ implies $\rho(\boldsymbol{r}_k) < 1$. Repeat the argument of the last paragraph to establish that $\lambda_{k+1} = \omega_{k+1}$.

We conclude that the conditional probability satisfies

$$\mathbb{P}\left(E_{\mathrm{succ}} \mid \Sigma\right) \geq \mathbb{P}\left\{\max_t \rho(\boldsymbol{q}_t) < 1 \mid \Sigma\right\} \tag{IV.2}$$

where $\{\boldsymbol{q}_t\}$ is a sequence of m random vectors that fall in the column span of $\mathbf{\Phi}_{\mathrm{opt}}$ and that are statistically independent from $\mathbf{\Psi}$.

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

$$\rho(\boldsymbol{q}_t) \;\le\; \frac{\sqrt{m}\, \max_{\boldsymbol{\psi}} |\langle \boldsymbol{\psi},\ \boldsymbol{q}_t \rangle|}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{\mathrm{T}}\, \boldsymbol{q}_t \right\|_{2}}.$$

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

$$\frac{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{\mathrm{T}}\, \boldsymbol{q} \right\|_{2}}{\|\boldsymbol{q}\|_{2}} \;\ge\; \sigma_m(\boldsymbol{\Phi}_{\mathrm{opt}}) \;\ge\; 0.5$$

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

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

for each index t. On account of this fact,

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

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

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

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

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

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

$$\mathbb{P}(\Sigma) \;=\; \mathbb{P}\{\sigma_m(\boldsymbol{\Phi}_{\mathrm{opt}}) \ge 0.5\} \;\ge\; 1 - \mathrm{e}^{-\mathrm{c}N}.$$

Introduce the latter two bounds into (IV.2), then substitute the result into (IV.1) to reach

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

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

$$\mathbb{P}(E_{\mathrm{succ}}) \;\ge\; 1 - 2m(d-m)\, \mathrm{e}^{-\mathrm{c}N/4m} - \mathrm{e}^{-\mathrm{c}N}.$$

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

$$\mathbb{P}(E_{\mathrm{succ}}) \;\ge\; 1 - d^2\, \mathrm{e}^{-\mathrm{c}N/m}.$$

In conclusion, the choice $N \ge \mathrm{K} m \log(d/\delta)$ is sufficient to reduce the failure probability below δ. To ensure that the logarithm exceeds one for all values of d, we require that $\delta < 0.36$.

V. 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 $\widehat{\boldsymbol{s}}$ is identical with the original signal $\boldsymbol{s}$ by comparing their supports. Proposition 7 implies that, if the supports match then the algorithm has succeeded with probability one. For each triple (m, N, d), we perform 1000 independent trials.

The first plot, Fig. 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.

Fig. 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. Fig. 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 I 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, our best theoretical bound for the Gaussian case is about $N \ge 4m(\ln d + 4.6)$ if we want a 99% probability of success [26].

Fig. 4 provides a graphical comparison between the empirical results and theoretical bounds from the technical report [26]. 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.

[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.

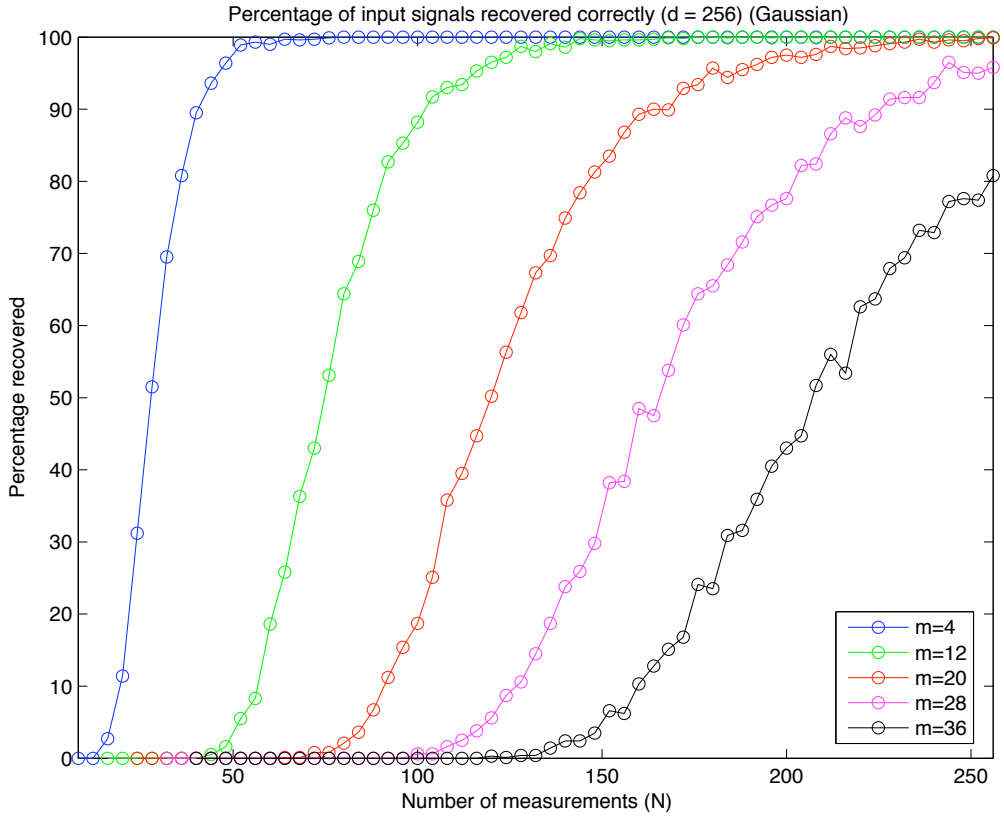

Fig. 1. The percentage of 1000 input signals correctly recovered as a function of the number N of measurements for different sparsity levels m in dimension $d = 256$.

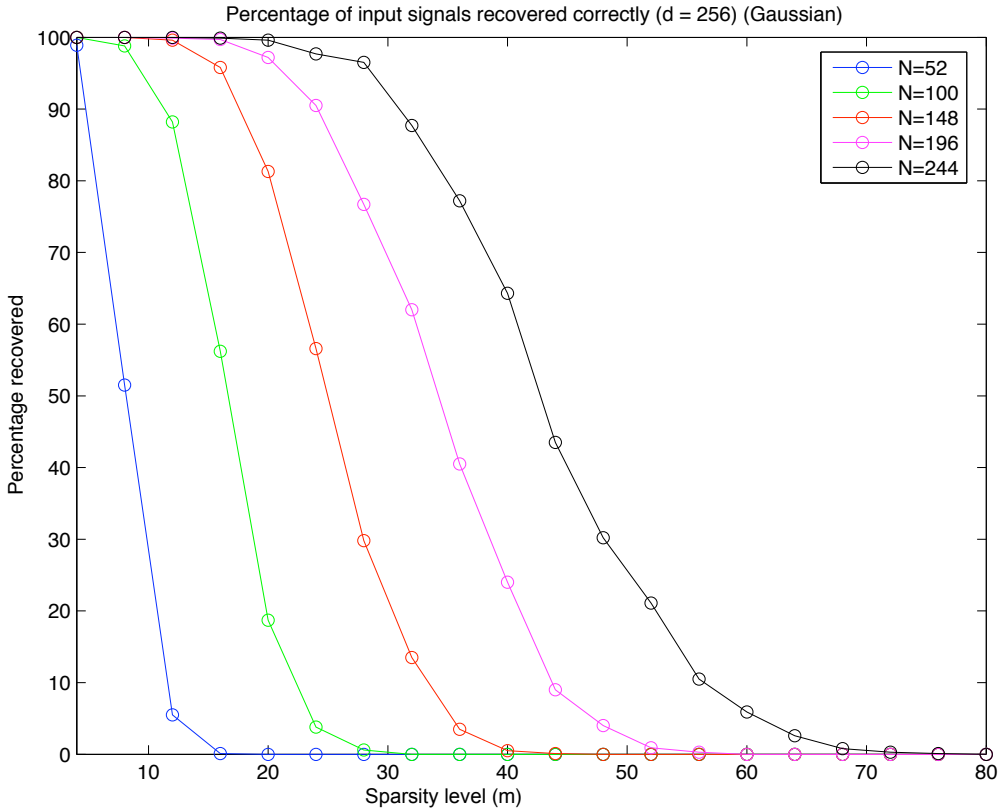

Fig. 2. The percentage of 1000 input signals correctly recovered as a function of the sparsity level m for different numbers N of measurements in dimension $d = 256$.

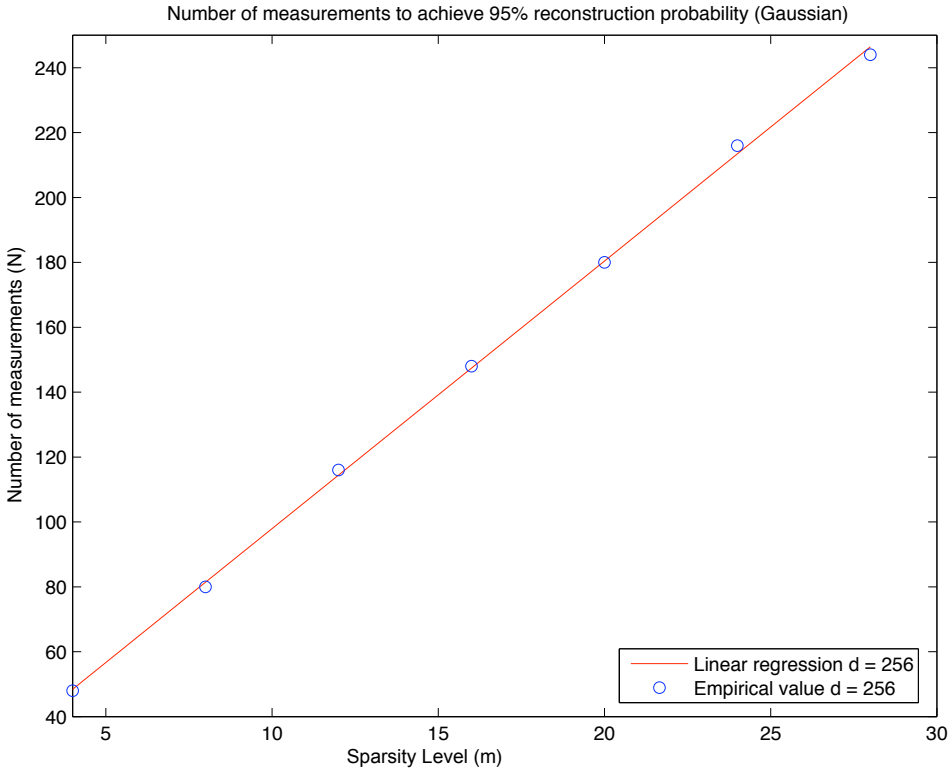

Fig. 3. The number N of measurements necessary to recover an m-sparse signal in dimension $d = 256$ at least 95% of the time. The regression line has equation $N = 1.5\, m \ln 256 + 15.4$.

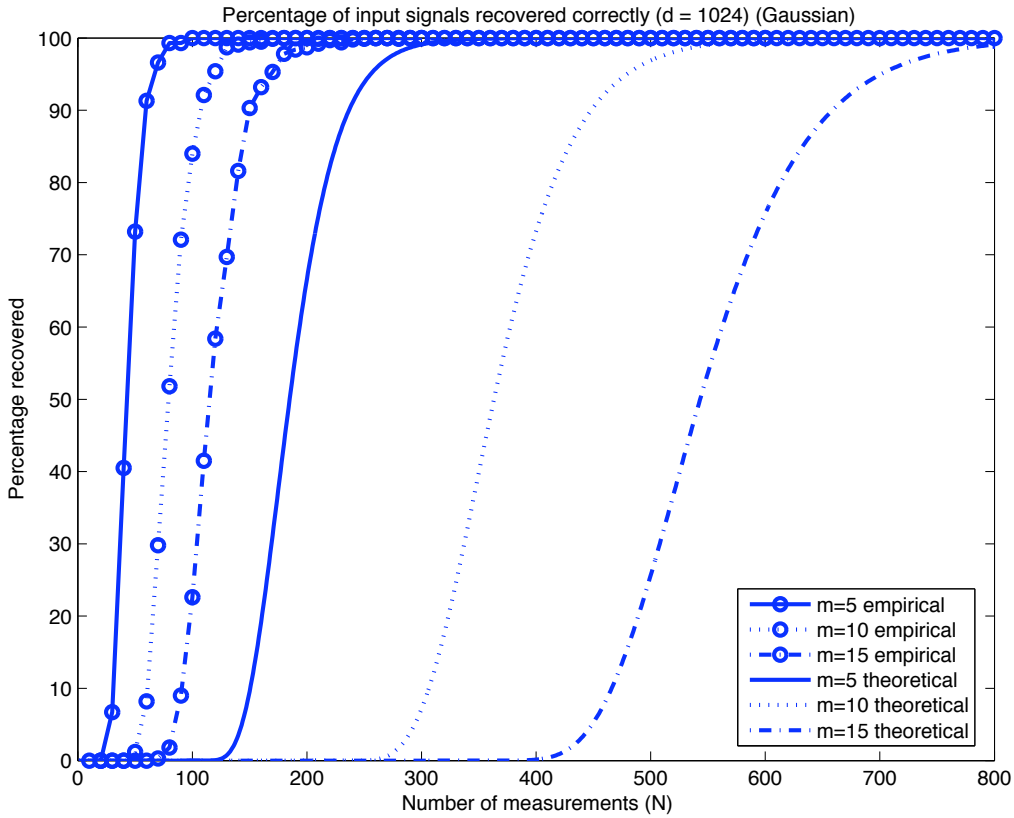

Fig. 4. The probability of recovering an m-sparse signal in dimension $d = 1024$ from N measurements. The marked lines display empirical data, while the unmarked lines show the theoretical bounds from Theorem 6 of [26].

$d = 256$			$d = 1024$		
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 I
THE NUMBER N OF MEASUREMENTS NECESSARY TO RECOVER AN m-SPARSE SIGNAL AT LEAST 99% OF THE TIME IN DIMENSIONS $d = 256, 1024$.

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. Fig. 5 shows the number of Bernoulli measurements necessary for OMP to recover an m-sparse signal in dimension $d = 256$. Comparing this chart with Fig. 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 Fig. 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.

VI. 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 describe another basic type of measurement ensemble. Finally, we discuss the relationship between our work and results on the linear program (BP).

A. 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 $\rho(\boldsymbol{q}_t)$ can be improved to $\sqrt{m - t}$. The effect of this change, however, seems to be minimal.

B. 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 [11] 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 [32] 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 $m^2 \ln d$. We consider it 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.

C. 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 statistically 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

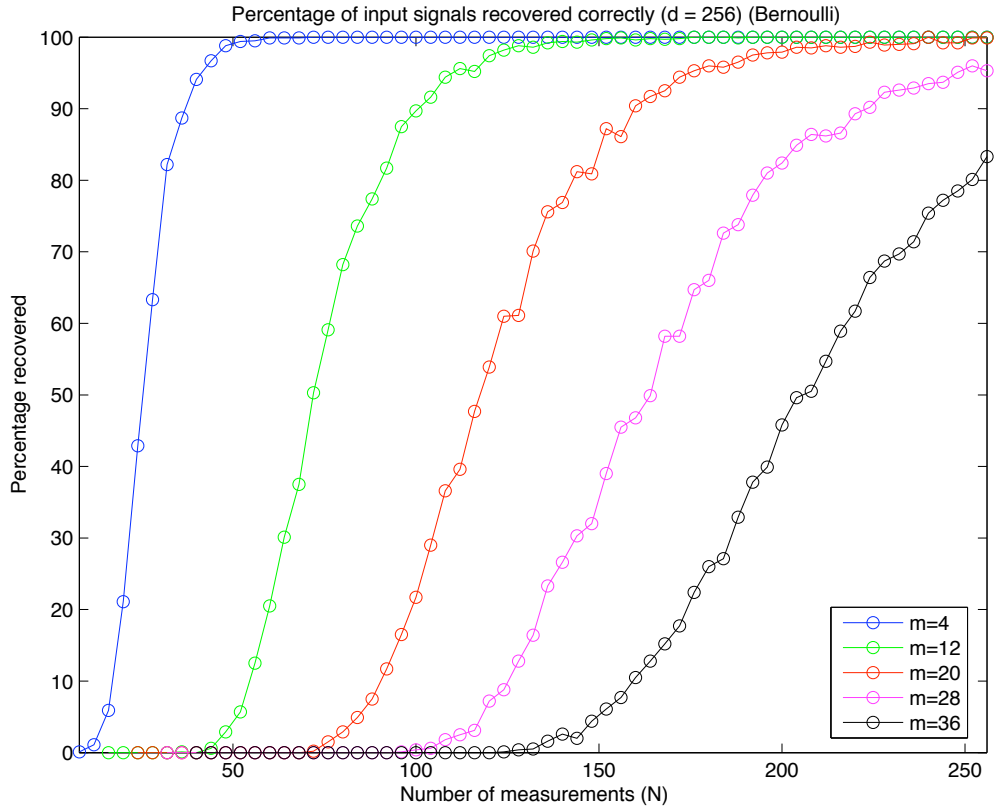

Fig. 5. The percentage of 1000 input signals correctly recovered as a function of the number N of Bernoulli measurements for different sparsity levels m in dimension $d = 256$.

set of input signals. Still, this amount of randomness is far from ideal.

D. 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 is preferable to employ a structured measurement ensemble that can be stored and processed efficiently. For this reason, the literature on (BP) advocates the use of random frequency measurements. That is, the N rows of the measurement matrix $\mathbf{\Phi}$ are drawn at random from the rows of the d-dimensional DFT matrix.

For OMP, random frequency measurements offer several specific advantages. Most significantly, it is possible to compute the maximum correlation between a signal and the columns of the matrix in time $\mathrm{O}(d \ln d)$ using an FFT. Second, the matrix can be constructed and stored using only $\mathrm{O}(N \ln d)$ bits because it is only necessary to choose N rows from a d-row matrix.

Kunis and Rauhut have studied the performance of OMP for signal recovery from random frequency measurements [23]. Their empirical work suggests that $\mathrm{O}(m \ln d)$ measurements are sufficient for OMP to recover an m-sparse signal in $\mathbb{R}^d$. Moreover, OMP often produces signal approximations that are superior to (BP). They also find that OMP executes faster than several algorithms for solving (BP).

Kunis and Rauhut were able to provide a partial theoretical explanation of their empirical work [23]. In particular, they show that the first iteration of OMP is likely to choose a correct column from the measurement matrix, given $\mathrm{O}(m \ln d)$ measurements of an m-sparse signal in $\mathbb{R}^d$. Unfortunately, since the columns of the measurement matrix are no longer statistically independent, it is difficult to analyze subsequent iterations of the algorithm. It remains an open problem to ascertain whether a result analogous to Theorem 6 holds for random frequency measurements.

Extremely recently, Needell and Vershynin have shown that a variant of OMP, called Regularized OMP (ROMP), can, with high probability, recover all m-sparse signals from $\mathrm{O}(m \ln^5 d)$ random frequency measurements [33]. This development is based on the Restricted Isometry Property [11] of random frequency measurements, and it should be considered a major step forward.

E. Comparison with Basis Pursuit

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

First, we note that there are situations where (BP) is provably more powerful than OMP. For example, with a Gaussian or Bernoulli measurement matrix, (BP) can, with high probability, recover all sparse signals. In the same setting, OMP recovers each sparse signal with high probability but with high probability fails to recover all sparse signals. One may infer the latter statement from Theorem 3.10 of [20] along with a somewhat involved probability estimate.

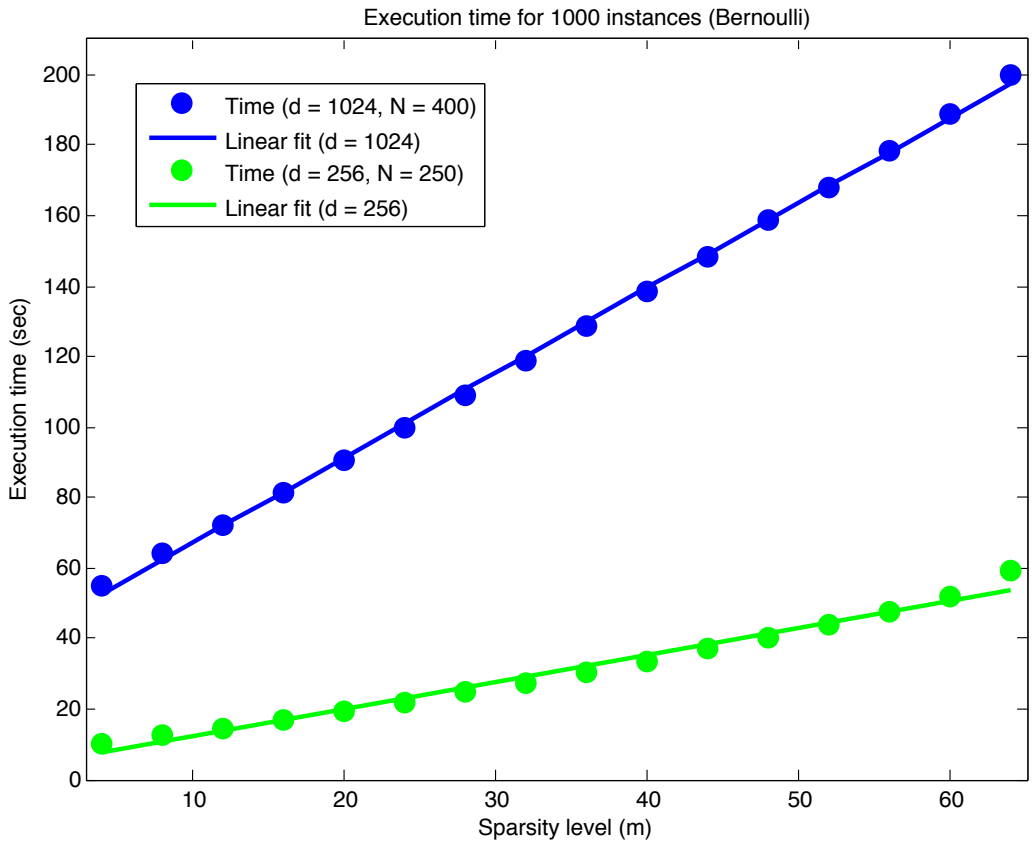

Fig. 6. The processor time, as a function of the sparsity level m, for 1000 complete trials in dimension $d = 256, 1024$ with $N = 250, 400$ Bernoulli measurements. The regression curves are linear polynomials calculated with least-squares.

Since OMP is inherently more difficulty to analyze than (BP), the literature on the convex relaxation 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 some advantages when we ask about computational cost. In certain parameter regimes, OMP is faster than standard approaches for completing the minimization (BP). OMP is especially efficient when the signal is highly sparse although homotopy methods for (BP) are competitive here [14]. When the signal is not very sparse, OMP may be a poor choice because the cost of orthogonalization increases quadratically with the number of iterations. In this setting, other types of greedy algorithms, such as StOMP [34] or ROMP [33], can alleviate the computational burden by reducing the number of least-squares problems that need to be solved.

Since we first announced our results on OMP in April 2005, there has been a significant amount of work on algorithms for (BP) and related problems. In consequence, it appears that the performance differences between the greedy approach and the optimization approach are becoming smaller.

Finally, note that greedy algorithms, such as OMP, are typically much easier to implement than algorithms for (BP). One should not underestimate the difficulties inherent in software engineering, so implementation complexity is a relevant point of comparison between algorithms. (As evidence, see the paper [35], which discusses software engineering challenges in optimization.)

F. Conclusions

The theoretical and empirical work in this paper demonstrates that OMP is an effective alternative to (BP) for signal recovery from random measurements. 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 may be faster and easier to implement, it offers an attractive alternative. 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. We also appreciate the careful comments of the anonymous reviewers.

REFERENCES

[1] S. Mallat and Z. Zhang, "Matching Pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Processing*, vol. 47, no. 1, pp. 187–200, 1999.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.
[4] A. J. Miller, *Subset Selection in Regression*, 2nd ed. London: Chapman and Hall, 2002.
[5] V. Temlyakov, "Nonlinear methods of approximation," *Foundations of Comput. Math.*, July 2002.
[6] D.-Z. Du and F. K. Hwang, *Combinatorial group testing and its applications*. World Scientific, 1993.
[7] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[8] M. Rudelson and R. Veshynin, "Geometric approach to error correcting codes and reconstruction of signals," *Int. Math. Res. Not.*, vol. 64, pp. 4019–4041, 2005.
[9] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[10] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[11] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
[12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.
[13] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.
[14] D. Malioutov, M. Çetin, and A. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, Philadelphia, 2005, pp. 733–736.
[15] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A method for large-scale ℓ_1-regularized least-squares problems with applications in signal processing and statistics," 2007, submitted for publication.
[16] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to Compressed Sensing and other inverse problems," 2007, submitted for publication.
[17] 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.
[18] G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *Constr. Approx.*, vol. 13, pp. 57–98, 1997.
[19] R. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 5, pp. 173–187, 1996.
[20] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Info. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[21] 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.*, vol. 59, no. 6, pp. 797–829, 2006.
[22] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM, 1996.
[23] S. Kunis and H. Rauhut, "Random sampling of sparse trigonometric polynomials II: Orthogonal Matching Pursuit versus Basis Pursuit," 2007, to appear, *Foundations Comp. Math.*
[24] Y. E. Nesterov and A. S. Nemirovski, *Interior Point Polynomial Algorithms in Convex Programming*. Philadelphia: SIAM, 1994.
[25] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, pp. 51–150, 1998.
[26] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via Orthogonal Matching Pursuit: The Gaussian case," Caltech, ACM Technical Report 2007-01, Aug. 2007. [Online]. Available: http://www.acm.caltech.edu/~jtropp/reports/TG07-Signal-Recovery-TR.pdf
[27] K. Ball, "Convex geometry and functional analysis," in *Handbook of Banach Space Geometry*, W. B. Johnson and J. Lindenstrauss, Eds. Elsevier, 2002, pp. 161–193.
[28] G. Lugosi, "Concentration of measure inequalities," 2005, lecture notes. [Online]. Available: http://www.econ.upf.es/~lugosi/surveys.html
[29] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," 2007, to appear, *Constr. Approx.*
[30] X. Fernique, "Regularité des trajectoires des fonctions aléatoires gausiennes," in *Ecole d'Eté de Probabilités de Saint-Flour IV*, ser. Lecture Notes in Mathematics. Springer, 1974, no. 480, pp. 1–96.
[31] J. Kahn, J. Komlós, and E. Szemerédi, "On the probability that a random ± 1-matrix is singular," *J. Amer. Math. Soc.*, vol. 86, no. 1, pp. 223–240, Jan. 1995.
[32] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *J. Signal Processing*, vol. 86, pp. 572–588, April 2006, special issue, "Sparse approximations in signal and image processing".
[33] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via Regularized Orthogonal Matching Pursuit," July 2007, submitted for publication.
[34] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined linear equations by stagewise Orthogonal Matching Pursuit (StOMP)," 2007, submitted for publication.
[35] P. E. Gill, W. Murray, and M. A. Saunders, "SNOPT: An SQP algorithm for large-scale constrained optimization," *SIAM Rev.*, vol. 47, no. 1, pp. 99–132, March 2005.

Joel A. Tropp (S'03, M'04) earned the B.A. degree in Plan II Liberal Arts Honors and the B.S. degree in Mathematics from the University of Texas at Austin in 1999, graduating *magna cum laude*. He continued his graduate studies in Computational and Applied Mathematics (CAM) at UT-Austin where he completed the M.S. degree in 2001 and the Ph.D. degree in 2004. His graduate work was supported by an NSF Graduate Fellowship and a CAM Graduate Fellowship.

In 2004, Dr. Tropp joined the Mathematics Department of the University of Michigan at Ann Arbor as Research Assistant Professor. In 2005, he won an NSF Mathematical Sciences Postdoctoral Research Fellowship, and he was appointed T. H. Hildebrandt Research Assistant Professor.

Since August 2007, Dr. Tropp has been Assistant Professor of Computational and Applied Mathematics at the California Institute of Technology.

Anna C. Gilbert earned the S.B. degree from the University of Chicago and the Ph.D. degree from Princeton University, both in mathematics. In 1997, she was a postdoctoral fellow at Yale University and AT&T Labs–Research. From 1998 to 2004, she was a member of technical staff at AT&T Labs–Research in Florham Park, NJ. In 2004, she joined the mathematics faculty of the University of Michigan at Ann Arbor.

Dr. Gilbert's research interests include analysis, probability, networking, and algorithms. She is especially interested in randomized algorithms with applications to harmonic analysis, signal and image processing, networking, and massive data sets.