Average-Case Analysis of Greedy Pursuit

Joel A. Tropp[a]

[a]Mathematics Department, The University of Michigan,
530 Church St., Ann Arbor, MI 48109-1043, USA

ABSTRACT

Recent work on sparse approximation has focused on the theoretical performance of algorithms for random inputs. This average-case behavior is typically far better than the behavior of the algorithm for the worst inputs. Moreover, an average-case analysis fits naturally with the type of signals that arise in certain applications, such as wireless communications. This paper describes what is currently known about the performance of greedy pursuit algorithms with random inputs. In particular, it gives a new result for the performance of Orthogonal Matching Pursuit (OMP) for sparse signals contaminated with random noise, and it explains recent work on recovering sparse signals from random measurements via OMP. The paper also provides a list of open problems to stimulate further research.

Keywords: Compressed sensing, Orthogonal Matching Pursuit, signal recovery, sparse approximation

1. INTRODUCTION

A sparse approximation problems requests an approximation of some signal (or function) as a short linear combination of elementary signals (i.e., basis functions). It is valuable to solve this problem because the approximation can identify complex structures that are latent in the signal. Another perspective is that the approximation provides a compressed version of the signal. Sparse approximation also has significant technological applications.

To solve sparse approximation problems, researchers have proposed many different algorithms, most of which fall into two basic categories: greedy pursuit[1] and ℓ_1 minimization.[2] The ℓ_1 methods are more powerful than greedy methods. Nevertheless, engineers tend to prefer greedy algorithms because they are faster and easier to implement.

The last year has witnessed an enormous amount of progress on the theoretical behavior of ℓ_1 minimization algorithms. Whereas earlier results for these methods had concentrated on how the algorithms perform for the worst inputs, the new results focused on the performance of algorithms for random inputs. In many cases, the average-case behavior outstrips the worst-case behavior by a large margin. By now, there is a long sequence of papers on this subject.[3–11]

Numerical experiments indicate that greedy pursuit methods also exhibit average-case behavior far superior to the worst case behavior. At present, there is little theoretical explanation for this success. The first goal of this article is to provide a few rigorous results on the average-case performance of greedy algorithms. The second goal is to present some open questions that would lead toward a more complete understanding of greedy pursuit.

The next section continues with a more formal statement of the sparse approximation problems that we are considering, and it motivates the idea of average-case analysis by reference to some applications. Afterward, we present a formal statement of a specific greedy pursuit algorithm, Orthogonal Matching Pursuit. The remaining three sections discuss the average-case analysis of three different sparse approximation problems.

E-mail: jtropp@umich.edu

2. SPARSE APPROXIMATION

We will be studying a specific type of sparse approximation problem called *sparse signal recovery* or *sparse signal identification*. Suppose that we measure a signal $\boldsymbol{s}$ of the form

$$\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{c}_{\text{opt}} + \boldsymbol{\nu} \tag{1}$$

where $\boldsymbol{\Phi}$ is a known matrix, $\boldsymbol{c}_{\text{opt}}$ is an unknown sparse coefficient vector, and $\boldsymbol{\nu}$ is an unknown noise vector.

> *Given the signal $\boldsymbol{s}$, our goal is to find a sparse coefficient vector $\widehat{\boldsymbol{c}}$ that approximates $\boldsymbol{c}_{\text{opt}}$. In particular, we hope that $\widehat{\boldsymbol{c}}$ correctly identifies the nonzero entries of $\boldsymbol{c}_{\text{opt}}$ since they determine which columns of $\boldsymbol{\Phi}$ participate in the signal.*

We implicitly assume that the matrix $\boldsymbol{\Phi}$ is fat (more columns than rows), which implies that $\boldsymbol{\Phi}$ has a nontrivial null space. Elementary linear algebra seems to preclude a solution to the signal recovery problem. Nevertheless, in certain cases, the sparsity requirement regularizes the problem enough that we can approximate the ideal coefficients well.

The signal model (1) contains three different pieces: the matrix, the coefficients, and the noise. In applications, it may be natural to view one or more of these components as random. As examples:

1. (Random noise). Suppose that $\boldsymbol{\Phi}$ is fixed, $\boldsymbol{c}_{\text{opt}}$ is an sparse coefficient vector, and $\boldsymbol{\nu}$ is a random vector. Then the model describes an ideal sparse signal that is contaminated with additive noise. In this case, the goal of the signal recovery problem is to identify the ideal signal from the noisy observation.[12]
2. (Random coefficients). Suppose that $\boldsymbol{\Phi}$ is fixed, $\boldsymbol{c}_{\text{opt}}$ is a sparse vector whose nonzero entries are random, and that the noise vector $\boldsymbol{\nu}$ is zero. Now the model describes the problem of identifying a sparse linear combination with random coefficients. This challenge can arise in wireless communications, where a sparse signal is transmitted and the coefficients are corrupted by multiplicative white noise.
3. (Random matrix). Suppose that $\boldsymbol{\Phi}$ is random, $\boldsymbol{c}_{\text{opt}}$ is a fixed sparse vector, and the noise vector $\boldsymbol{\nu}$ is zero. Then our goal is to identify a sparse signal from random measurements. This problem has recently received a significant amount of attention.[3, 7, 13]

For signal identification problems, many algorithms perform well on random instances with high probability, even though the algorithms may fail completely for the worst instances. Studying the probabilistic performance of an algorithm with respect to a random signal model can be viewed as an average-case analysis of the algorithm. The applications we have described provide an impetus for attempting such an average-case analysis.

2.1. Notation and Terminology

We provide a very brief introduction to the notation and terminology of sparse approximation. For a more detailed explanation, see other recent papers of the author.[12]

The symbol $\langle \cdot,\ \cdot \rangle$ indicates the usual bilinear inner product, and $\|\cdot\|_2$ refers to the associated norm. The ℓ_∞ norm $\|\boldsymbol{x}\|_\infty = \max_k |x_k|$. The ℓ_0 quasi-norm is $\|\boldsymbol{x}\|_0 = \#\{x_k : x_k \neq 0\}$. For $0 < p < 1$, the ℓ_p quasi-norm is calculated as

$$\|\boldsymbol{x}\|_p = \sum\nolimits_k |x_k|^p .$$

The symbol $\boldsymbol{\Phi}$ will denote a $d \times N$ matrix, called the *dictionary*. Its columns are sometimes called *atoms*, and they are denoted by $\boldsymbol{\varphi}_\omega$, where ω is drawn from the index set $\Omega = \{1, 2, \ldots, N\}$. The *support* of a vector $\boldsymbol{c} \in \mathbb{R}^\Omega$ is the set of indices of its nonzero entries:

$$\operatorname{supp}(\boldsymbol{c}) = \{\omega : c_\omega \neq 0\}.$$

We will employ the symbol Λ to indicate a subset of Ω. The *coherence* μ of the matrix $\mathbf{\Phi}$ is the absolute inner product between distinct columns:

$$\mu = \max_{\lambda \neq \omega} |\langle \boldsymbol{\varphi}_\lambda, \boldsymbol{\varphi}_\omega \rangle| .$$

Finally, we write $\text{NORMAL}(0, \mathbf{\Sigma})$ to denote the normal probability distribution with zero mean and positive-definite covariance matrix $\mathbf{\Sigma}$. In $\mathbb{R}^d$, this distribution has density function

$$f(\boldsymbol{x}) = \frac{\det(\mathbf{\Sigma})^{1/2}}{(2\pi)^{d/2}} \exp\{-\tfrac{1}{2} \langle \mathbf{\Sigma}^{-1} \boldsymbol{x}, \boldsymbol{x} \rangle\}.$$

3. GREEDY PURSUIT

This paper concentrates on greedy algorithms for sparse approximation. Greedy methods are popular in the engineering community because they are reasonably fast, powerful, and implementable. Unfortunately, their theoretical properties are not well understood in many régimes of interest. This section describes the generic structure of a greedy pursuit, and it presents a specific greedy algorithm that we will analyze in more detail.

3.1. Overview

A *greedy pursuit* method for sparse approximation is an iterative algorithm that consists of two basic steps and a criterion for halting.

The first step of the iteration is called the *greedy selection*. The reason for the unflattering sobriquet "greedy" is that the algorithm picks the atom that will provide the biggest improvement in the current iteration. The hope is that these locally optimal choices will lead to a globally optimal (or near-optimal) sparse approximation.

The second step of the iteration is called the *greedy update*. Here, the algorithm uses the new atom and the current approximation to determine a new approximation to the input signal. The simplest method simply adds a scalar multiple of the new atom to the current approximation, while more sophisticated techniques may recalculate all the coefficients in the current approximation.

After each iteration, the algorithm decides whether or not to proceed. It may just stop after a fixed number of iterations, or it may use current information more subtly to determine whether another iteration would be beneficial.

3.2. Orthogonal Matching Pursuit

We focus on a particular greedy algorithm called *Orthogonal Matching Pursuit (OMP)*. This algorithm was developed by the statistics community during the 1950s, where it was called *stagewise regression*.[14] In the 1990s, it reappeared in the signal processing literature (as OMP) and in the literature on approximation theory as the *Orthogonal Greedy Algorithm*.[15]

ALGORITHM 1 (ORTHOGONAL MATCHING PURSUIT).
INPUT:

- *A $d \times N$ matrix $\mathbf{\Phi}$*
- *A d-dimensional input signal $\boldsymbol{s}$*
- *A stopping criterion*

OUTPUT:

- *The total number T of iterations*
- *A d-dimensional approximation $\widehat{\boldsymbol{s}}$ of the input signal*
- *A d-dimensional residual $\boldsymbol{r} = \boldsymbol{s} - \widehat{\boldsymbol{s}}$*

- *An N-dimensional T-sparse coefficient vector $\widehat{\boldsymbol{c}}$ such that $\widehat{\boldsymbol{s}} = \boldsymbol{\Phi}\widehat{\boldsymbol{c}}$*

PROCEDURE:

1. *Initialize the residual $\boldsymbol{r}_0 = \boldsymbol{s}$, the index set $\Lambda_0 = \emptyset$, and the iteration counter $t = 0$.*
2. *Increment the counter t.*
3. *Find the index λ_t that solves the easy optimization problem*

$$\lambda_t \;=\; \arg\max_{\omega \in \Omega} \; |\langle \boldsymbol{r}_{t-1},\; \boldsymbol{\varphi}_\omega \rangle| \,.$$

   If the maximum occurs for multiple indices, break the tie deterministically. Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$.

4. *Calculate the new approximation and residual:*

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

   where $\boldsymbol{P}_t$ is the orthogonal projector onto $\operatorname{span}\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_t\}$.

5. *Return to Step 2 unless the stopping criterion is met.*
6. *The approximation $\widehat{\boldsymbol{s}} = \boldsymbol{a}_t$. The coefficient vector $\widehat{\boldsymbol{c}}$ has nonzero entries at the indices listed in Λ_t; its values in these components appear in the series expansion*

$$\widehat{\boldsymbol{s}} \;=\; \sum\nolimits_{\lambda \in \Lambda_t} \widehat{c}_\lambda \, \boldsymbol{\varphi}_\lambda .$$

It is important to note that, at each iteration, the residual is orthogonal to the atoms that have already been chosen. Therefore, the algorithm selects a new atom in every iteration, so the index set Λ_t always contains t distinct indices.

The most significant time cost in the algorithm is the calculation of the inner products in Step 3. In general, this step requires $O(dN)$ time per iteration, but it can be reduced when the dictionary admits a fast transform.[1] It may also be possible to use approximate nearest-neighbor data structures to reduce the cost farther,[16] although this approach has not been implemented.

The calculation of the orthogonal projection in Step 4 can be accomplished quickly using standard methods of numerical linear algebra. The marginal cost in iteration t is $O(td)$, so the overall cost is $O(T^2 d)$, where T is the total number of iterations.

3.3. Stopping Criteria

There are several natural methods for deciding when to halt the iteration.

1. Halt the algorithm after a fixed number of iterations.
2. Halt the algorithm when $\|\boldsymbol{r}_t\|_2$, the norm of the residual, declines below a specific tolerance.
3. Halt the algorithm when $\|\boldsymbol{\Phi}^* \boldsymbol{r}_t\|_\infty$, the maximum correlation between the residual and the dictionary, drops below some threshold.

Naturally, the appropriate criterion depends on the application domain. In the sequel, we will indicate which criterion is appropriate for the problem we are discussing.

3.4. Why is Average-Case Analysis Hard?

The major challenge in producing an average-case analysis of Orthogonal Matching Pursuit is that the residuals are not stochastically independent of each other or the input signal. So far, successful proofs of average-case results for OMP have skirted this issue. To answer the open questions stated in this paper, it may be necessary to deal with the correlations directly.

4. RANDOM NOISE

In this section, we present a result on the performance of OMP for sparse signals contaminated with noise. In short, the algorithm can tolerate random noise much more easily than deterministic noise. The intuition behind this result is that deterministic noise can be aligned with one of the nonoptimal atoms, while it is very probable that the random noise is almost orthogonal to all the nonoptimal atoms (provided that there are not too many).

Let us pose a concrete signal model.

The matrix:	$\mathbf{\Phi}$	Dimensions $d \times N$, unit-norm columns, coherence μ
The coefficients:	$\boldsymbol{c}_{\text{opt}}$	Sparsity level $m = \|\boldsymbol{c}_{\text{opt}}\|_0$ where $m\mu \leq \frac{1}{3}$
The noise:	$\boldsymbol{\nu}$	Distributed as $\text{NORMAL}(0, \sigma^2 \mathbf{I})$

We consider the following sparse approximation problem.

Observed Signal:	$\boldsymbol{s} = \mathbf{\Phi}\boldsymbol{c}_{\text{opt}} + \boldsymbol{\nu}$
Goal:	Given $\boldsymbol{s}$ and $\mathbf{\Phi}$, find the support of $\boldsymbol{c}_{\text{opt}}$

Now, let us describe how to adapt Orthogonal Matching Pursuit to solve the problem. Naturally, we cannot expect to recover entries of the coefficient vector that are very small because the noise may drown them out. So we must develop a stopping criterion that depends on the size of the smallest coefficient. Define $c_{\min}$ to be the absolute value of the smallest nonzero entry of $\boldsymbol{c}_{\text{opt}}$. That is,

$$c_{\min} = \min\nolimits_{\lambda \in \text{supp}(\boldsymbol{c}_{\text{opt}})} |\boldsymbol{c}_{\text{opt}}(\lambda)| \,.$$

Choose a threshold τ that satisfies $\tau < c_{\min}/3$. We halt OMP when the correlation between the residual and the dictionary becomes sufficiently small.

Stopping Criterion:	$\|\mathbf{\Phi}^* \boldsymbol{r}_t\|_\infty \leq 2\tau$

In this setting, we have the following new result. This theorem is a consequence of a general result on OMP[17] combined with some probability estimates developed for an ℓ_1 minimization algorithm.[12] A sketch of the argument appears later in this section.

THEOREM 1 (RANDOM NOISE). *Execute OMP on the observed signal $\boldsymbol{s}$ to obtain a coefficient vector $\widehat{\boldsymbol{c}}$. Then the support of $\widehat{\boldsymbol{c}}$ equals the support of $\boldsymbol{c}_{\text{opt}}$ with probability (over the noise) exceeding*

$$\left[1 - \exp\{-\tfrac{1}{2}\tau^2/\sigma^2\}\right]^{N-m} \left[1 - \exp\{-\tfrac{1}{3}(c_{\min} - 3\tau)^2/\sigma^2\}\right]^m .$$

The success probability involves a delicate interplay between the noise level σ^2, the size $c_{\min}$ of the smallest coefficient, and the threshold τ that we use to halt the algorithm. Note that we may vary the threshold τ between zero and $c_{\min}/3$ to maximize the probability of success. If we choose τ to equate the two terms in the probability bound, we reach a cleaner result.

COROLLARY 2. *Select* $\tau = 0.236\, c_{\min}$. *Then the lower bound on the success probability in Theorem 1 exceeds*

$$\left[1 - \exp\{-(0.167\, c_{\min}/\sigma)^2\}\right]^N.$$

In particular, to obtain a failure probability of δ, *it suffices that*

$$\sigma < \frac{0.167}{\ln^{1/2}(N/\delta)}\, c_{\min}.$$

In contrast, suppose that we have no statistical model for the noise vector $\boldsymbol{\nu}$ but only a bound on its norm: $\|\boldsymbol{\nu}\|_2 \le \varepsilon$. To ensure that we recover the entire support of $\boldsymbol{c}_{\text{opt}}$, it suffices that $\varepsilon \le c_{\min}/3$. Moreover, one can probably contrive examples where the algorithm fails to tolerate noise above this level. These claims follow from Theorem 3 of the sequel and its proof.

Let us develop a concrete numerical example based on calculations in the literature.[12] Suppose that the dimension $d = 1024$ and the size of the dictionary is $N = 4096$. If the dictionary has coherence $\mu = 1/32$, our signal model permits a 10-sparse signal. Suppose that we wish to recover this signal with 99% probability of success over the noise. In this case, we can tolerate noise with $\sigma = 0.0491\, c_{\min}$, which means that the total noise power is $\mathbb{E}\,\|\boldsymbol{\nu}\|_2^2 = 2.21\, c_{\min}^2$. If the ten nonzero coefficients in the ideal signal all have absolute value one, the signal-to-noise ratio (SNR) is no greater than 7.81 dB. In contrast, if the noise is deterministic, we must require that $\|\boldsymbol{\nu}\|_2^2 \le 0.111\, c_{\min}^2$, which is twenty times smaller than the power in the random noise.

For comparison, here is a slightly larger example. Let $d = 4096$ and $N = 16384$. If the coherence is $\mu = 1/64$, we can allow 21-sparse signals. To achieve at least 99% success, we may set $\sigma = 0.0441\, c_{\min}$. If the ideal coefficients have constant absolute value, the SNR does not exceed 5.45 dB. In summary, quadrupling the length of the signal doubles the number of atoms, while the noise variance decreases just slightly and the SNR goes down by 2.36 dB, which is a factor of 1.72.

4.1. Open Questions

It is unlikely that Theorem 1 can be strengthened substantially. But it can certainly be refined to give a sharper reflection of several factors:

- Other types of distributions for the noise $\boldsymbol{\nu}$.
- The relative magnitude of the nonzero entries in $\boldsymbol{c}_{\text{opt}}$.
- The cumulative coherence μ_1 of the matrix.[18]

These extensions could be important for understanding the performance of OMP in applications.

4.2. Proof

Theorem 1 follows from a general theorem on the performance of Orthogonal Matching Pursuit,[17] combined with some estimates on the probability that the hypotheses of this theorem hold.[12]

THEOREM 3 (TROPP–GILBERT–STRAUSS). *Suppose that* Λ *lists* m *atoms, where* $m\mu \le \frac{1}{3}$. *Let* $\boldsymbol{s}$ *be an arbitrary input signal,* $\boldsymbol{a}_\Lambda$ *its best approximation over the atoms listed in* Λ, *and* $\boldsymbol{c}_\Lambda$ *the coefficient vector that synthesizes* $\boldsymbol{a}_\Lambda$. *Finally, assume we have the bound*

$$\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty \le \tau.$$

After iteration t *of Orthogonal Matching Pursuit, halt the algorithm if*

$$\|\boldsymbol{\Phi}^* \boldsymbol{r}_t\|_\infty \le 2\,\tau.$$

It follows that

- *the algorithm has chosen t indices from Λ, and*
- *the algorithm has chosen every index λ from Λ for which $|\boldsymbol{c}_\Lambda(\lambda)| > 3\tau$.*

In particular, if $|\boldsymbol{c}_\Lambda(\lambda)| > 3\tau$ for each index λ in Λ, then the algorithm identifies Λ and recovers the best approximation of the signal over Λ. That is, $\widehat{\boldsymbol{s}} = \boldsymbol{a}_\Lambda$ and $\widehat{\boldsymbol{c}} = \boldsymbol{c}_\Lambda$.

We wish to apply this theorem to the random signal model described in the last subsection. To that end, suppose that $\boldsymbol{c}_{\text{opt}}$ is the ideal coefficient vector, and let $\Lambda = \text{supp}(\boldsymbol{c}_{\text{opt}})$. Suppose that we draw a random signal $\boldsymbol{s}$ according to the model. Denote by $\boldsymbol{a}_\Lambda$ the best approximation of $\boldsymbol{s}$ over the atoms listed in Λ, and let $\boldsymbol{c}_\Lambda$ be the coefficients that synthesize $\boldsymbol{s}$. Note that $\boldsymbol{c}_\Lambda$ does not equal $\boldsymbol{c}_{\text{opt}}$ because of the influence of the noise.

To determine the probability that we may invoke the theorem, we must bound the probability over the noise that $\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty \le \tau$. To ensure that the algorithm identifies Λ, the entire support of $\boldsymbol{c}_{\text{opt}}$, we must also bound the probability that $|\boldsymbol{c}_\Lambda(\lambda)| > 3\tau$ for each λ in Λ. These calculations can be extracted from Appendix IV-A of a recent paper.[12]

LEMMA 4 (TROPP). *Suppose that the hypotheses of the last subsection are in force. Write $\Lambda = \text{supp}(\boldsymbol{c}_{\text{opt}})$. Let $\boldsymbol{a}_\Lambda$ be the best approximation of $\boldsymbol{s}$ over the atoms in Λ, and suppose that $\boldsymbol{c}_\Lambda$ is the coefficient vector that synthesizes $\boldsymbol{a}_\Lambda$. We have*

$$\text{Prob}\left\{\|\boldsymbol{\Phi}^*(\boldsymbol{s} - \boldsymbol{a}_\Lambda)\|_\infty \le \tau\right\} \ge \left[1 - \exp\{-\tfrac{1}{2}\tau^2/\sigma^2\}\right]^{N-m}.$$

and also

$$\text{Prob}\left\{|\boldsymbol{c}_\Lambda(\lambda)| > 3\tau \text{ for all } \lambda \in \Lambda\right\} \ge \left[1 - \exp\{-\tfrac{1}{3}(c_{\min} - 3\tau)^2/\sigma^2\}\right]^m$$

Moreover, these two events are stochastically independent.

We obtain Theorem 1 by applying the Lemma to estimate the probability that Theorem 3 yields the desired conclusions.

5. RANDOM COEFFICIENTS

Let us turn to the problem of perfectly recovering a sparse signal whose nonzero coefficients are random. For simplicity, we assume that this signal is not contaminated with additive noise. There are applications (such as wireless communications) in which sparse signals with random coefficients arise. Understanding the noiseless case is an important step toward understanding the general case.

This signal model is intriguing because experiments show that OMP can identify these signals much more easily than arbitrary sparse signals. In fact, numerical results indicate that we can reliably recover random superpositions of $O(d)$ atoms, whereas the algorithm can fail for the worst superposition of $O(\sqrt{d})$ atoms. Unfortunately, we have no theoretical explanation for this behavior. This section describes these observations in more detail, which leads to a collection of important open problems.

We consider a simple signal model.

The matrix:	$\boldsymbol{\Phi}$	Dimensions $d \times N$, unit-norm columns, coherence μ
The coefficients:	$\boldsymbol{c}_{\text{opt}}$	Sparsity level $m = \|\boldsymbol{c}_{\text{opt}}\|_0$
		Each nonzero entry of $\boldsymbol{c}_{\text{opt}}$ is iid NORMAL$(0, 1)$

And we attempt to solve the following sparse approximation problem.

Observed Signal:	$\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{c}_{\text{opt}}$
Goal:	Given $\boldsymbol{s}$ and $\boldsymbol{\Phi}$, determine $\boldsymbol{c}_{\text{opt}}$

To solve this problem, we halt OMP after it has chosen m atoms.

Stopping Criterion: $t = m$

If we do not take into account the statistical distribution of the coefficients, a worst case analysis[18] leads to a strong condition on the sparsity level, $m < \frac{1}{2}(\mu^{-1} + 1)$, to ensure that we recover the ideal coefficients. For most dictionaries, the coherence μ is at least $d^{-1/2}$. Therefore, the worst-case analysis suggests that we cannot recover signals unless the sparsity level $m = O(\sqrt{d})$.

On the other hand, consider the following experiment documented by Tropp et al.[17] Let the dictionary $\mathbf{\Phi} = [\mathbf{I}\ \boldsymbol{F}]$, the concatenation of the 128×128 identity matrix and the 128×128 (unitary) discrete Fourier transform matrix. This matrix has coherence $\mu = 1/\sqrt{128}$. At each sparsity level m, we generate 1000 signals with random coefficients, and we execute OMP. For each sparsity level, we calculate the average percentage of the optimal support that the algorithm determines correctly. The worst-case analysis suggests that we should only be able to recover about $0.5\,(\sqrt{128} + 1) \approx 6$ atoms. And yet the algorithm reliably recovers the *entire* support for sparsity levels up to $m = 64$. See Figure 1 for a plot of the data. We have no rigorous explanation for this phenomenon.

5.1. Open Questions

The experimental evidence makes it clear that OMP performs far better for sparse signals with random coefficients than the worst-case analysis indicates.

Open Question 1. *Fix a matrix $\mathbf{\Phi}$. Suppose that $\boldsymbol{c}_{\text{opt}}$ is an m-sparse vector whose nonzero entries are chosen at random. What is the probability that OMP will correctly determine $\boldsymbol{c}_{\text{opt}}$ given the input signal $\boldsymbol{s} = \mathbf{\Phi}\boldsymbol{c}_{\text{opt}}$? What properties of the matrix $\mathbf{\Phi}$ determine this success probability?*

We feel that the latter problem represents one of the most important challenges in the theory of greedy pursuit algorithms. A second (more difficult) question concerns the performance of OMP for random sparse signals that have been contaminated with noise. Here is one reasonable formulation.

Open Question 2. *Fix a matrix $\mathbf{\Phi}$. Suppose that $\boldsymbol{c}_{\text{opt}}$ is an m-sparse vector whose nonzero entries are chosen at random, and let $\boldsymbol{\nu}$ be a random noise vector. If we execute OMP with the input signal $\boldsymbol{s} = \mathbf{\Phi}\boldsymbol{c}_{\text{opt}} + \boldsymbol{\nu}$, what is the probability that it returns a coefficient vector $\widehat{\boldsymbol{c}}$ that satisfies $\|\widehat{\boldsymbol{c}} - \boldsymbol{c}_{\text{opt}}\|_2 \leq \varepsilon$? What is the probability that the supports of $\widehat{\boldsymbol{c}}$ and $\boldsymbol{c}_{\text{opt}}$ match?*

It is also possible to extend these types of questions to coefficient vectors that are well approximated by sparse vectors.

Open Question 3. *Fix a matrix $\mathbf{\Phi}$. Suppose that $\boldsymbol{c}_{\text{opt}}$ is chosen randomly from the ℓ_p unit ball, where $0 < p < 1$. Let $\boldsymbol{c}_m$ be the best m-sparse approximation of $\boldsymbol{c}_{\text{opt}}$. Suppose that we execute OMP with the input signal $\boldsymbol{s} = \mathbf{\Phi}\boldsymbol{c}_{\text{opt}}$. What is the probability that it returns an m-sparse coefficient vector $\widehat{\boldsymbol{c}}$ that satisfies $\|\widehat{\boldsymbol{c}} - \boldsymbol{c}_m\|_2 \leq \varepsilon$?*

This last problem can also be extended to the case of noisy observations, but we omit a rigorous statement of this question.

6. RANDOM DICTIONARIES

This section discusses the problem of recovering a sparse signal from random measurements. This setup leads to a signal model where the coefficient vector is fixed and the dictionary is a random matrix. The literature contains fairly comprehensive results on the performance of ℓ_1 minimization in this context,[3,7] but there is limited information about the behavior of OMP.[13] We describe the work that has been done on this problem, and we list a series of open questions.

To develop a theoretical result, we place the following assumptions on the signal model.

The matrix:	$\mathbf{\Phi}$	Dimensions $d \times N$, entries iid $\text{NORMAL}(0, 1)$
The coefficients:	$\boldsymbol{c}_{\text{opt}}$	Sparsity level $m = \|\boldsymbol{c}_{\text{opt}}\|_0$

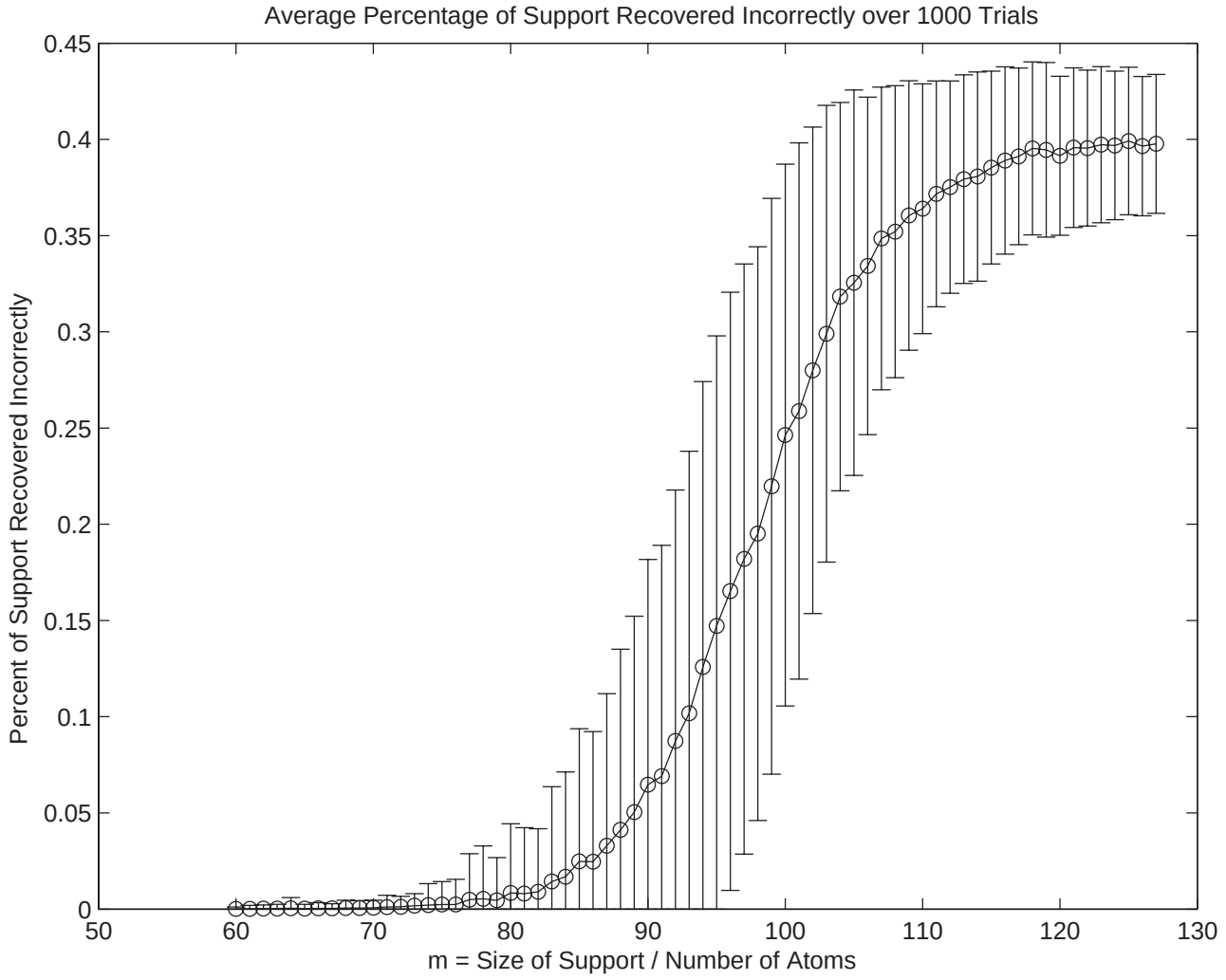

Figure 1. At each sparsity level, OMP is invoked 1000 times with random coefficients. The sparsity level is marked on the horizontal axis. The vertical axis shows what percentage of the support the algorithm *fails* to identify on average. The error bars mark one standard deviation in each direction. We have omitted the part of the curve for $m < 60$ because it coincides with the horizontal axis. This figure is reproduced from the paper.[17]

Here is the statement of the sparse approximation problem.

Observed Signal:	$\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{c}_{\text{opt}}$
Goal:	Given $\boldsymbol{s}$ and $\boldsymbol{\Phi}$, determine $\boldsymbol{c}_{\text{opt}}$

To solve this problem, we halt OMP after choosing m atoms.

Stopping Criterion:	$t = m$

In this setting, we have the following result, which can be extracted from the work of Tropp and Gilbert.[13]

THEOREM 5 (TROPP–GILBERT). *Execute OMP on the observed signal $\boldsymbol{s}$ to obtain a coefficient vector $\widehat{\boldsymbol{c}}$. Then $\widehat{\boldsymbol{c}} = \boldsymbol{c}_{\text{opt}}$ with probability (over the matrix) exceeding*

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

A corollary of this result gives an explicit bound on the size of m that suffices to obtain a good success probability.

COROLLARY 6 (TROPP–GILBERT). *Fix a positive number p. In the preceding theorem, to obtain a success probability exceeding $(1 - 2N^{-p})$, it suffices that the sparsity level*

$$m \leq \frac{d}{8\,(p+1)\ln N}.$$

We see that the algorithm succeeds at sparsity levels far beyond $m = O(\sqrt{d})$.

6.1. Open Questions

The major shortcoming in Theorem 5 is that it only applies to signals that are exactly sparse. It is important to extend this result to signals that have a good sparse approximation.

OPEN QUESTION 4. *Suppose that $\boldsymbol{c}_{\text{opt}}$ is an arbitrary signal in the ℓ_p unit ball for $0 < p < 1$. Suppose that $\boldsymbol{\Phi}$ is a random matrix. If we execute OMP with the input signal $\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{c}_{\text{opt}}$, what is the probability that it returns a coefficient vector $\widehat{\boldsymbol{c}}$ for which $\|\widehat{\boldsymbol{c}} - \boldsymbol{c}_{\text{opt}}\|_2 \leq \varepsilon$? For a given number m, how many rows must $\boldsymbol{\Phi}$ have to ensure that the success probability is close to one?*

It is also natural to ask about the performance of OMP when the matrix is random and the signal is also corrupted with random noise, i.e., the measurements are imperfect.

OPEN QUESTION 5. *Suppose that $\boldsymbol{c}_{\text{opt}}$ is an arbitrary m-sparse vector. Suppose that $\boldsymbol{\Phi}$ is a random matrix and $\boldsymbol{\nu}$ is random noise. If we execute OMP with the input signal $\boldsymbol{s} = \boldsymbol{\Phi}\boldsymbol{c}_{\text{opt}} + \boldsymbol{\nu}$, what is the probability that it returns a coefficient vector $\widehat{\boldsymbol{c}}$ for which $\|\widehat{\boldsymbol{c}} - \boldsymbol{c}_{\text{opt}}\|_2 \leq \varepsilon$?*

Another valuable direction would be to understand the performance of the algorithm for other types of random matrices. The case of subgaussian (including ± 1) random matrices has already been developed in unpublished work.[19] But numerical evidence suggests that orthogonalizing the rows of $\boldsymbol{\Phi}$ leads to a significant improvement in the behavior of the algorithm.

OPEN QUESTION 6. *Suppose that $\boldsymbol{\Phi}$ is a matrix whose rows form an orthonormal basis for a random subspace. For N-dimensional coefficient vectors with a given sparsity level m, how many rows d must the matrix $\boldsymbol{\Phi}$ have to ensure that OMP succeeds with probability close to one?*

Answers to these questions could have a significant technological impact.

ACKNOWLEDGMENTS

I wish to thank Anna C. Gilbert, Martin J. Strauss, and Roman Vershynin for discussions of the work described in this article. The writing of this manuscript has been supported by the Erwin Schrödinger Institute and by NSF Grant No. DMS-0503299.

REFERENCES

1. G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *J. Constr. Approx.* **13**, pp. 57–98, 1997.
2. S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comp.* **20**(1), pp. 33–61, 1999.
3. E. J. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?." Submitted for publication, Nov. 2004.
4. E. Candès, J. Romberg, and T. Tao, "Exact signal reconstruction from highly incomplete frequency information." Submitted for publication, June 2004.
5. D. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," Statistics Department Technical Report, Stanford Univ., Sept. 2004.
6. D. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution approximates the sparsest near-solution," Statistics Department Technical Report, Stanford Univ., Sept. 2004.
7. D. L. Donoho, "Compressed sensing." Unpublished manuscript, Oct. 2004.
8. D. L. Donoho, "Neighborly polytopes and sparse solutions of underdetermined linear equations," Statistics Department Technical Report, Stanford Univ., Jan. 2005.
9. E. J. Candès and T. Tao, "Decoding by linear programming." Available from `arXiv:math.MG/0502327`, Feb. 2005.
10. M. Rudelson and R. Veshynin, "Geometric approach to error correcting codes and reconstruction of signals." Available from `arXiv:math.MG/0502299`, Feb. 2005.
11. D. L. Donoho and J. Tanner, "Sparse nonnegative solutions of underdetermined linear equations by linear programming," Statistics Department Technical Report, Stanford Univ., April 2005.
12. J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals." Submitted for publication June 2004; revised Feb. 2005.
13. J. A. Tropp and A. C. Gilbert, "Signal recovery from partial information via Orthogonal Matching Pursuit." Submitted for publication, April 2005.
14. A. J. Miller, *Subset Selection in Regression*, Chapman and Hall, London, 2nd ed., 2002.
15. V. Temlyakov, "Nonlinear methods of approximation," *Foundations of Comp. Math.* , July 2002.
16. A. C. Gilbert, M. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, Jan. 2003.
17. J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *EURASIP J. Signal Processing* , 2005. To appear.
18. J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory* **50**, pp. 2231–2242, Oct. 2004.
19. A. C. Gilbert, J. A. Tropp, and R. Vershynin, "Instant algorithms for signal recovery." Working draft.