



**SIGNAL RECOVERY FROM RANDOM MEASUREMENTS
VIA ORTHOGONAL MATCHING PURSUIT:
THE GAUSSIAN CASE**

JOEL A. TROPP AND ANNA C. GILBERT

Technical Report No. 2007-01
August 2007

APPLIED & COMPUTATIONAL MATHEMATICS
CALIFORNIA INSTITUTE OF TECHNOLOGY
mail code 217-50 · pasadena, ca 91125

SIGNAL RECOVERY FROM RANDOM MEASUREMENTS VIA ORTHOGONAL MATCHING PURSUIT: THE GAUSSIAN CASE

JOEL A. TROPP AND ANNA C. GILBERT

ABSTRACT. This report demonstrates theoretically and empirically that a greedy algorithm called Orthogonal Matching Pursuit (OMP) can reliably recover a signal with m nonzero entries in dimension d given $O(m \ln d)$ random linear measurements of that signal. This is a massive improvement over previous results, which require $O(m^2)$ 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.

1. INTRODUCTION

Let \mathbf{s} be a d -dimensional real signal with at most m nonzero components. This type of signal is called m -sparse. Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ 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:

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

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product. The problem of *signal recovery* asks two distinct questions:

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

As we will see, signal recovery is dual to sparse approximation, a problem of significant interest [MZ93, RKD99, CDS01, Mil02, Tem02].

To the first question, we can immediately respond that no fewer than m measurements will do. Even if the measurements were adapted to the signal, it would still take m pieces of information to determine 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 [DH93] for a survey.

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

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

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

JAT is 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 has been supported by NSF DMS 0503299 and ACG has been supported by NSF DMS 0354600.

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 [CT05] and Rudelson–Vershynin [RV05], describes one example of this remarkable phenomenon.

Theorem 1. *Let $N \geq Km \ln(d/m)$, and draw N vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ independently from the standard Gaussian distribution on \mathbb{R}^d . The following statement is true with probability exceeding $1 - e^{-kN}$. It is possible to reconstruct every m -sparse signal \mathbf{s} in \mathbb{R}^d from the data $\{\langle \mathbf{s}, \mathbf{x}_n \rangle : n = 1, 2, \dots, N\}$.*

We follow the analysts’ convention that upright letters (c, C, 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 [CRT06], Donoho [Don06a], and Candès–Tao [CT06].

All five of the papers [CRT06, Don06a, CT06, RV05, CT05] offer constructive demonstrations of the recovery phenomenon by proving that the original signal \mathbf{s} is the unique solution to the mathematical program

$$\min_{\mathbf{f}} \|\mathbf{f}\|_1 \quad \text{subject to} \quad \langle \mathbf{f}, \mathbf{x}_n \rangle = \langle \mathbf{s}, \mathbf{x}_n \rangle \quad \text{for } n = 1, 2, \dots, N. \quad (\text{BP})$$

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 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 [CDS01, EHJT04, DDM04, MÇW05, KKL⁺07, FNW07]. 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) [PRK93, DMA97] 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 [DT96]. Pursuing their reasoning leads to an example of a nonrandom ensemble of measurement vectors and a sparse signal that OMP cannot identify without d measurements [CDS01, Sec. 2.3.2]. Other negative results, such as Theorem 3.10 of [Tro04] and Theorem 5 of [Don06b], echo this concern.

But these negative results about OMP are 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 technical report is to establish the following theorem in detail.

Theorem 2 (OMP with Gaussian Measurements). *Fix $\delta \in (0, 0.36)$, and choose $N \geq Km \ln(d/\delta)$. Suppose that \mathbf{s} is an arbitrary m -sparse signal in \mathbb{R}^d . Draw N measurement vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ independently from the standard Gaussian distribution on \mathbb{R}^d . Given the data $\{\langle \mathbf{s}, \mathbf{x}_n \rangle : n = 1, 2, \dots, N\}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $1 - 2\delta$. The constant satisfies $K \leq 20$. For large values of m , it can be reduced to $K \approx 4$.*

In comparison, earlier positive results, such as Theorem 3.6 from [Tro04], only demonstrate that Orthogonal Matching Pursuit can recover m -sparse signals when the number of measurements N is 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. Nevertheless, we believe that the advantages of Orthogonal Matching Pursuit make Theorem 2 extremely compelling.

This section describes how to apply a fundamental algorithm from sparse approximation to the signal recovery problem. Suppose that \mathbf{s} is an arbitrary m -sparse signal in \mathbb{R}^d , and let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ be a family of N measurement vectors. Form an $N \times d$ matrix Φ whose *rows* are the measurement vectors, and observe that the N measurements of the signal can be collected in an N -dimensional data vector $\mathbf{v} = \Phi \mathbf{s}$. We refer to Φ as the *measurement matrix* and denote its columns by $\varphi_1, \dots, \varphi_d$.

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

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

Algorithm 3 (OMP for Signal Recovery).

INPUT:

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

OUTPUT:

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

PROCEDURE:

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

$$\lambda_t = \arg \max_{j=1, \dots, d} |\langle \mathbf{r}_{t-1}, \varphi_j \rangle|.$$

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 $\Phi_t = [\Phi_{t-1} \quad \varphi_{\lambda_t}]$. We use the convention that Φ_0 is an empty matrix.
- (4) Solve a least-squares problem to obtain a new signal estimate:

$$\mathbf{x}_t = \arg \min_{\mathbf{x}} \|\mathbf{v} - \Phi_t \mathbf{x}\|_2.$$

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

$$\begin{aligned}\mathbf{a}_t &= \mathbf{\Phi}_t \mathbf{x}_t \\ \mathbf{r}_t &= \mathbf{v} - \mathbf{a}_t.\end{aligned}$$

(6) Increment t , and return to Step 2 if $t < m$.

(7) The estimate $\hat{\mathbf{s}}$ for the ideal signal has nonzero indices at the components listed in Λ_m . The value of the estimate $\hat{\mathbf{s}}$ in component λ_j equals the j th component of \mathbf{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 \mathbf{r}_t is always orthogonal to the columns of $\mathbf{\Phi}_t$. Provided that the residual \mathbf{r}_{t-1} is nonzero, the algorithm selects a new atom at iteration t and the matrix $\mathbf{\Phi}_t$ has full column rank. In which case the solution \mathbf{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 $O(mNd)$. At iteration t , the least-squares problem can be solved with marginal cost $O(tN)$. To do so, we maintain a \mathbf{QR} factorization of $\mathbf{\Phi}_t$. Our implementation uses the Modified Gram-Schmidt (MGS) algorithm because the measurement matrix is unstructured and dense. The book [Bj96] 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 [KR07] for one example.

According to [NN94], there are algorithms that can solve (BP) with a dense, unstructured measurement matrix in time $O(N^2 d^{3/2})$. 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.

A prototype of the OMP algorithm first appeared in the statistics community at some point in the 1950s, where it was called stagewise regression. The algorithm later developed a life of its own in the signal processing [MZ93, PRK93, DMA97] and approximation theory [DeV98, Tem02] literatures.

2. GAUSSIAN MEASUREMENT ENSEMBLES

In this report, we are concerned with Gaussian measurements only. In this section, we identify the properties of this measurement ensemble that are used to prove that the algorithm succeeds. Since Gaussian matrices are so well studied, we can make much more precise claims about them than other types of random matrices.

A Gaussian measurement ensemble for m -sparse signals in \mathbb{R}^d is a $d \times N$ matrix $\mathbf{\Phi}$, whose entries are drawn independently from the $\text{NORMAL}(0, N^{-1})$ distribution. For reference, the density function p of this distribution is

$$p(x) = \frac{1}{\sqrt{2\pi N}} e^{-x^2 N/2} \quad \text{for } x \in \mathbb{R}.$$

As we will see, this matrix has the following four properties:

- (G0) Independence: The columns of $\mathbf{\Phi}$ are statistically independent.
- (G1) Normalization: $\mathbb{E} \|\boldsymbol{\varphi}_j\|_2^2 = 1$ for $j = 1, \dots, d$.
- (G2) Joint correlation: Let $\{\mathbf{u}_t\}$ be a sequence of m vectors whose ℓ_2 norms do not exceed one. Let $\boldsymbol{\varphi}$ be a column of $\mathbf{\Phi}$ that is independent from this sequence. Then

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

- (G3) Smallest singular value: Given an $N \times m$ submatrix \mathbf{Z} from $\mathbf{\Phi}$, the m th largest singular value $\sigma_m(\mathbf{Z})$ satisfies

$$\mathbb{P} \left\{ \sigma_m(\mathbf{Z}) \geq 1 - \sqrt{m/N} - \varepsilon \right\} \geq 1 - e^{-\varepsilon^2 N/2}$$

for any positive ε .

2.1. Joint Correlation. The joint correlation property (G2) is essentially a large deviation bound for sums of random variables. For the Gaussian measurement ensemble, we can leverage classical techniques to establish this property.

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

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

Proof. Let \mathbf{z} be a random vector whose entries are i.i.d. $\text{NORMAL}(0, 1)$. Define the event

$$E \stackrel{\text{def}}{=} \{\mathbf{z} : \max_t |\langle \mathbf{z}, \mathbf{u}_t \rangle| \leq \varepsilon \sqrt{N}\},$$

which is identical with the event that interests us. We will develop a lower bound on $\mathbb{P}(E)$. Observe that this probability decreases if we replace each vector \mathbf{u}_t by a unit vector pointing in the same direction. Therefore, we may assume that $\|\mathbf{u}_t\|_2 = 1$ for each t .

Geometrically, we can view $\mathbb{P}(E)$ as the Gaussian measure of m intersecting symmetric slabs. Sidak's Lemma [Bal02, Lemma 2] shows that the Gaussian measure of this intersection is no smaller than the product of the measures of the slabs. In probabilistic language,

$$\mathbb{P}(E) \geq \prod_{t=1}^m \mathbb{P}\{|\langle \mathbf{z}, \mathbf{u}_t \rangle| \leq \varepsilon \sqrt{N}\}.$$

Since each \mathbf{u}_t is a unit vector, each of the random variables $\langle \mathbf{z}, \mathbf{u}_t \rangle$ has a $\text{NORMAL}(0, 1)$ distribution on the real line. It follows that each of the m probabilities can be calculated as

$$\begin{aligned} \mathbb{P}\{|\langle \mathbf{z}, \mathbf{u}_t \rangle| \leq \varepsilon \sqrt{N}\} &= \frac{1}{\sqrt{2\pi}} \int_{-\varepsilon}^{\varepsilon} e^{-x^2/2} dx \\ &\leq 1 - e^{-\varepsilon^2 N/2}. \end{aligned}$$

The final estimate is a well-known Gaussian tail bound. See [Bal02, p. 118], for example. \square

2.2. Smallest Singular Value. The singular value property (G3) follows directly from a theorem of Davidson and Szarek [DS02].

Proposition 5 (Davidson–Szarek). *Suppose that \mathbf{Z} is a tall $N \times m$ matrix whose entries are i.i.d. $\text{NORMAL}(0, N^{-1})$. Then its smallest singular value σ_m satisfies*

$$\mathbb{P}\left\{\sigma_m(\mathbf{Z}) \geq 1 - \sqrt{m/N} - \varepsilon\right\} \geq 1 - e^{-\varepsilon^2 N/2}.$$

It is a standard consequence of measure concentration that the minimum singular value of a Gaussian matrix clusters around its expected value (see [Led01], for example). Calculating the expectation, however, involves much more ingenuity. Davidson and Szarek produce their result with a clever application of the Slepian–Gordon lemma.

3. SIGNAL RECOVERY WITH ORTHOGONAL MATCHING PURSUIT

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

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

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

The success probability here is best calculated numerically. Some analysis yields a slightly weaker but more serviceable corollary.

Corollary 7. Fix $\delta \in (0, 0.36)$, and choose $N \geq Km \log(d/\delta)$ where K is an absolute constant. Suppose that \mathbf{s} is an arbitrary m -sparse signal in \mathbb{R}^d , and draw a random $N \times d$ Gaussian measurement matrix Φ independent from the signal. Given the data $\mathbf{v} = \Phi \mathbf{s}$, Orthogonal Matching Pursuit can reconstruct the signal with probability exceeding $1 - 2\delta$.

The preceding theorem holds with $K \leq 20$ for any $m \geq 1$. When the number m of nonzero signal components approaches infinity, it is possible to take $K \leq 4 + \eta$ for any positive number η . If $\delta \geq d^{-1}$, the success probability can also be improved to $1 - 2\delta^2$. These facts will emerge during the proof.

3.1. Proof of Theorem 6. Most of the argument follows the approach developed in [Tro04]. The main difficulty here is to deal with the nasty independence issues that arise in the 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 \mathbf{s} are nonzero, while the remaining $d - m$ entries equal zero. Therefore, the data vector \mathbf{v} is a linear combination of the first m columns from the matrix Φ . Partition the matrix as $\Phi = [\Phi_{\text{opt}} \mid \Psi]$ so that Φ_{opt} has m columns and Ψ has $d - m$ columns. Note that the vector $\mathbf{v} = \Phi \mathbf{s}$ is statistically independent from the random matrix Ψ .

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

$$\Sigma \stackrel{\text{def}}{=} \{\sigma_m(\Phi_{\text{opt}}) \geq \sigma\}.$$

Applying the definition of conditional probability, we reach

$$\mathbb{P}(E_{\text{succ}}) \geq \mathbb{P}(E_{\text{succ}} \cap \Sigma) = \mathbb{P}(E_{\text{succ}} \mid \Sigma) \cdot \mathbb{P}(\Sigma). \quad (3.1)$$

Property (G3) 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 Φ_{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 Φ_{opt} has full column rank. In other words, there is just one explanation for the signal \mathbf{s} using the columns in Φ_{opt} .

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

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

where the maximization takes place over the columns of Ψ . If \mathbf{r} is the residual vector that arises in Step 2 of OMP, the algorithm picks a column from Φ_{opt} whenever $\rho(\mathbf{r}) < 1$. In case $\rho(\mathbf{r}) = 1$, 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 [Tro04].

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

Execute OMP with the input signal \mathbf{s} and the full matrix Φ to obtain the actual sequence of residuals $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{m-1}$ and the actual sequence of column indices $\lambda_1, \lambda_2, \dots, \lambda_m$. Conditional on Σ , OMP succeeds in reconstructing \mathbf{s} after m iterations if and only if the algorithm selects the m

columns of Φ_{opt} in some order. We use induction to prove that this situation occurs when $\rho(\mathbf{q}_t) < 1$ for each $t = 0, 1, \dots, m-1$.

The statement of the algorithm ensures that the initial residuals satisfy $\mathbf{q}_0 = \mathbf{r}_0$. Clearly, the condition $\rho(\mathbf{q}_0) < 1$ ensures $\rho(\mathbf{r}_0) < 1$. It follows that the actual invocation chooses the column λ_1 from Φ_{opt} whose inner product with \mathbf{r}_0 has the largest magnitude (ties broken deterministically). Meanwhile, the imaginary invocation chooses the column ω_1 from Φ_{opt} whose inner product with \mathbf{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 $j = 1, 2, \dots, k$. Since the algorithm calculates the new residual as the (unique) best approximation of the signal \mathbf{s} from the span of the chosen columns, the actual and imaginary residuals must be identical at the beginning of iteration k . In symbols, $\mathbf{r}_k = \mathbf{q}_k$. An obvious consequence is that $\rho(\mathbf{q}_k) < 1$ implies $\rho(\mathbf{r}_k) < 1$. Repeat the argument of the last paragraph to establish that $\lambda_{k+1} = \omega_{k+1}$.

We conclude that the conditional probability satisfies

$$\mathbb{P}(E_{\text{succ}} \mid \Sigma) \geq \mathbb{P}\{\max_t \rho(\mathbf{q}_t) < 1 \mid \Sigma\} \quad (3.2)$$

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

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

$$\rho(\mathbf{q}_t) = \frac{\max_{\psi} |\langle \psi, \mathbf{q}_t \rangle|}{\|\Phi_{\text{opt}}^T \mathbf{q}_t\|_{\infty}}.$$

Since $\Phi_{\text{opt}}^T \mathbf{q}_t$ is an m -dimensional vector,

$$\rho(\mathbf{q}_t) \leq \frac{\sqrt{m} \max_{\psi} |\langle \psi, \mathbf{q}_t \rangle|}{\|\Phi_{\text{opt}}^T \mathbf{q}_t\|_2}.$$

To simplify this expression, define the vector

$$\mathbf{u}_t \stackrel{\text{def}}{=} \frac{\sigma \mathbf{q}_t}{\|\Phi_{\text{opt}}^T \mathbf{q}_t\|_2}.$$

The basic properties of singular values furnish the inequality

$$\frac{\|\Phi_{\text{opt}}^T \mathbf{q}\|_2}{\|\mathbf{q}\|_2} \geq \sigma_m(\Phi_{\text{opt}}) \geq \sigma$$

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

$$\rho(\mathbf{q}_t) \leq \frac{\sqrt{m}}{\sigma} \max_{\psi} |\langle \psi, \mathbf{u}_t \rangle|$$

for each index t . On account of this fact,

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

Exchange the two maxima and use the independence of the columns of Ψ to obtain

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

Since every column of Ψ is independent from $\{\mathbf{u}_t\}$ and from Σ , Property (G2) of the measurement matrix yields a lower bound on each of the $d-m$ terms appearing in the product. It emerges that

$$\mathbb{P}\{\max_t \rho(\mathbf{q}_t) < 1 \mid \Sigma\} \geq \left(1 - e^{-\sigma^2 N/2m}\right)^{m(d-m)}.$$

We may choose the parameter

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

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

$$\mathbb{P} \{ \max_t \rho(\mathbf{q}_t) < 1 \mid \Sigma \} \geq \left(1 - e^{-(\sqrt{N/m} - 1 - \varepsilon)^2/2} \right)^{m(d-m)}.$$

With the foregoing choice of σ , Property (G3) furnishes

$$\mathbb{P} \{ \sigma_m(\Phi_{\text{opt}}) \geq \sigma \} \geq 1 - e^{-\varepsilon^2 m/2}.$$

Introduce the last two facts into (3.2) and substitute the result into (3.1). This action yields

$$\mathbb{P}(E_{\text{succ}}) \geq \left[1 - e^{-(\sqrt{N/m} - 1 - \varepsilon)^2/2} \right]^{m(d-m)} \left[1 - e^{-\varepsilon^2 m/2} \right] \quad (3.3)$$

for $\varepsilon \in (0, \sqrt{N/m} - 1)$. The optimal value of this probability estimate is best determined numerically.

3.2. Proof of Corollary 7. We need to show that it possible to choose the number of measurements on the order of $m \ln d$ while maintaining an error as small as we like. We begin with (3.3), in which we apply the inequality $(1 - x)^k \geq 1 - kx$, valid for $k \geq 1$ and $x \leq 1$. Then invoke the bound $m(d - m) \leq d^2/4$ to reach

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - \frac{d^2}{4} \exp \left\{ -(\sqrt{N/m} - 1 - \varepsilon)^2/2 \right\} - \exp \{ -\varepsilon^2 m/2 \}. \quad (3.4)$$

where we have also discarded a positive term of higher order. We will bound the two terms on the right-hand side separately.

Fix a number $\delta \in (0, 0.36)$. Select $N \geq Km \ln(d/\delta)$, where the constant $K = K(m)$ will be determined in a moment. Now, we set

$$\varepsilon = \left(\frac{2 \ln(d/\delta)}{m} \right)^{1/2}.$$

Clearly, ε is positive; our choice of K will also ensure that ε is not too large. Substituting the value of ε into the last term on the right-hand side of (3.4), we find that

$$\exp \{ -\varepsilon^2 m/2 \} = \delta/d.$$

Evidently, this term does not exceed δ . In fact, when $\delta \geq d^{-1}$, it is smaller than δ^2 .

Next, we consider the second term on the right-hand side of (3.4). By construction of N and ε , we have

$$\left(\sqrt{N/m} - 1 - \varepsilon \right)^2 \geq \left(\sqrt{K} - \sqrt{2/m} - u \right)^2 \ln(d/\delta)$$

where $u^{-2} = \ln(d/\delta)$. The last displayed equation implies

$$\frac{d^2}{4} \exp \left\{ -(\sqrt{N/m} - 1 - \varepsilon)^2/2 \right\} \leq \frac{1}{4} d^{2 - (\sqrt{K} - \sqrt{2/m} - u)^2/2} \cdot \delta^{(\sqrt{K} - \sqrt{2/m} - u)^2/2}.$$

We set $K = (2 + u + \sqrt{2/m})^2$ to zero the exponent on d . It follows that the second term on the right-hand side of (3.4) is no larger than $\delta^2/4$.

In view of these bounds,

$$\mathbb{P}(E_{\text{succ}}) \geq 1 - (0.25 + 1)\delta > 1 - 2\delta.$$

When $\delta \geq d^{-1}$, we can improve the success probability to $1 - 2\delta^2$. Moreover, the argument provides a sufficient choice for the constant:

$$K \leq \left(2 + \frac{1}{\sqrt{\ln(d/\delta)}} + \sqrt{\frac{2}{m}} \right)^2$$

For the worst-case values $m = 1$, $d = 1$, and $\delta = 0.36$, it suffices to take $K \leq 20$. On the other hand, as m tends to infinity (hence also $d \rightarrow \infty$), we may select $K \leq 4 + \eta$ for any positive η .

REFERENCES

- [Bal02] K. Ball. Convex geometry and functional analysis. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 161–193. Elsevier, 2002.
- [Bjö96] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.
- [CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM Rev.*, 43(1):129–159, 2001.
- [CRT06] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Trans. Info. Theory*, 52(2):489–509, Feb. 2006.
- [CT05] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Info. Theory*, 51(12):4203–4215, Dec. 2005.
- [CT06] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Info. Theory*, 52(12):5406–5425, Dec. 2006.
- [DDM04] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413–1457, 2004.
- [DeV98] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.
- [DH93] D.-Z. Du and F. K. Hwang. *Combinatorial group testing and its applications*. World Scientific, 1993.
- [DMA97] G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. *Constr. Approx.*, 13:57–98, 1997.
- [Don06a] D. L. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, Apr. 2006.
- [Don06b] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.
- [DS02] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of Banach Space Geometry*, pages 317–366. Elsevier, 2002.
- [DT96] R. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5:173–187, 1996.
- [EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004.
- [FNW07] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to Compressed Sensing and other inverse problems. Submitted for publication, 2007.
- [KKL⁺07] 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. Submitted for publication, 2007.
- [KR07] S. Kunis and H. Rauhut. Random sampling of sparse trigonometric polynomials II: Orthogonal Matching Pursuit versus Basis Pursuit. To appear, *Foundations Comp. Math.*, 2007.
- [Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, Providence, 2001.
- [MCW05] D. Malioutov, M. Çetin, and A. Willsky. Homotopy continuation for sparse signal representation. In *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, volume 5, pages 733–736, Philadelphia, 2005.
- [Mil02] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 2nd edition, 2002.
- [MZ93] S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.
- [NN94] Y. E. Nesterov and A. S. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.
- [PRK93] Y. C. Pati, R. Rezaifar, 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.
- [RKD99] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Trans. Signal Processing*, 47(1):187–200, 1999.
- [RV05] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. *Int. Math. Res. Not.*, 64:4019–4041, 2005.
- [Tem02] V. Temlyakov. Nonlinear methods of approximation. *Foundations of Comput. Math.*, July 2002.
- [Tro04] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, Oct. 2004.