ALGORITHMS FOR SIMULTANEOUS SPARSE APPROXIMATION PART I: GREEDY PURSUIT

JOEL A. TROPP, ANNA C. GILBERT, AND MARTIN J. STRAUSS

Abstract. A simultaneous sparse approximation problem requests a good approximation of several input signals at once using different linear combinations of the same elementary signals. At the same time, the problem balances the error in approximation against the total number of elementary signals that participate. These elementary signals typically model coherent structures in the input signals, and they are chosen from a large, linearly dependent collection.

The first part of this paper proposes a greedy pursuit algorithm, called Simultaneous Orthogonal Matching Pursuit, for simultaneous sparse approximation. Then it presents some numerical experiments that demonstrate how a sparse model for the input signals can be identified more reliably given several input signals. Afterward, the paper proves that the S-OMP algorithm can compute provably good solutions to several simultaneous sparse approximation problems.

The second part of the paper develops another algorithmic approach called convex relaxation, and it provides theoretical results on the performance of convex relaxation for simultaneous sparse approximation.

Date: Typeset on March 17, 2005.

Key words and phrases. Greedy algorithms, Orthogonal Matching Pursuit, multiple measurement vectors, simultaneous sparse approximation, subset selection.

The authors may be reached by e-mail at {jtropp,annacg,martinjs}@umich.edu or by post at the Department of Mathematics, The University of Michigan, 2074 East Hall, Ann Arbor, MI 48109-1109.

This research has been supported by NSF DMS 0354600.

1. Introduction

In recent years, the signal processing community has lavished attention on the class of simple sparse approximation problems. These problems have two facets:

(1) A signal vector is approximated using a linear combination of elementary signals, which are drawn from a fixed collection. In modern problems, this collection is often linearly dependent and large.
(2) The problem seeks a compromise between the approximation error (usually measured with Euclidean distance) and the number of elementary signals that participate in the linear combination. The goal is to identify a good approximation involving few elementary signals—a *sparse* approximation.

Simple sparse approximation problems originally arose in the study of linear regression. In this setting, we wish to approximate a data vector using a linear combination of regressors, but we must control the number of regressors to avoid fitting noise in the data. Statisticians developed many of the numerical algorithms that are used for solving simple sparse approximation problems [Mil02].

One striking generalization of simple sparse approximation has garnered little attention in the literature. Consider the following scenario. Suppose that we have several observations of a signal that has a sparse representation. Each view is contaminated with noise, which need not be statistically independent. It seems clear that we should be able to use the additional information to produce a superior estimate of the underlying signal. This intuition suggests that we study *simultaneous sparse approximation*:

> *Given several input signals, approximate all these signals at once using different linear combinations of the same elementary signals, while balancing the error in approximating the data against the total number of elementary signals that are used.*

Simultaneous sparse approximation problems arise in several specific domains. For example, B. D. Rao and his colleagues have considered applications to magnetoencephalography [GGR95] and to the equalization of sparse communications channels [CR02]. R. Gribonval has developed applications to blind source separation [Gri02]. Malioutov et al. have shown that source localization using a linear array of sensors can be posed as a simultaneous sparse approximation problem [Mal03, MÇW03]. It is easy to imagine many other applications in statistics, wireless communications, and machine learning.

1.1. Contributions.

This work examines simultaneous sparse approximation from the practical and the theoretical point of view.

In the first part of the paper, we propose a greedy algorithm that generalizes the familiar Orthogonal Matching Pursuit procedure, which was developed for simple sparse approximation [PRK93, DMA97]. At each iteration, a greedy pursuit makes the best local improvement to the current approximations in hope of obtaining a good overall solution. The same algorithm has been developed independently in [CH04a, CH04b].

Then we summarize some numerical experiments using this greedy algorithm. These experiments confirm our intuition that having multiple observations of a sparse signal can improve our ability to identify the underlying sparse representation. They also give a measure of how the algorithm's performance depends on the number of input signals, the level of sparsity, and the signal-to-noise ratio.

Afterward, we prove that the greedy algorithm can calculate good solutions to simultaneous sparse approximation problems. Moreover, if we have some basic information about the signals, this information can be used to enhance the performance of the algorithm. Our proofs require that the collection of elementary signals possess a geometric property called *incoherence*. Roughly,

incoherence means the elementary signals are weakly correlated with each other. The theoretical arguments build on work in [Tro04b, Tro04e, TGS04].

In the second part of the paper, we develop a more sophisticated numerical method for simultaneous sparse approximation based on convex relaxation. Convex relaxation replaces the difficult simultaneous sparse approximation problem by a convex optimization problem, which can be solved in polynomial time with standard mathematical programming software. Using a variation of the argument in [Tro04c], we prove that convex programming yields good solutions to simultaneous sparse approximation problems, even in the presence of noise.

Our analysis of these two algorithmic methods for simultaneous sparse approximation yields the first rigorous proof that these algorithms can succeed for sparse signals contaminated with noise. The present work also underscores the value of the abstract approach to simple sparse approximation adopted in [Tro04b, Tro04c]. Indeed, it is possible to obtain lovely results for simultaneous sparse approximation just by "capitalizing" proofs from the earlier articles. (That is, signal vectors are replaced by signal matrices.) But this paper offers more than an a slavish repetition of old ideas: our theory for the simultaneous greedy pursuit algorithm also contains strong, qualitatively new results for simple sparse approximation problems. We hope that this work provides a firm practical and theoretical foundation for future research on simultaneous sparse approximation problems.

1.2. **Outline.** Let us offer a brief outline of the paper. Section 2 provides an introduction to the approximation model, as well as the means for measuring approximation error and sparsity. Section 3 states the greedy algorithm and discusses some of its basic properties. The results of some numerical experiments appear in Section 4. Section 5 demonstrates that the greedy pursuit algorithm can calculate provably good solutions to simultaneous sparse approximation problems. In the final Section 6, we make some comparisons with previous work.

2. Background

2.1. **Signal Matrices.** A *signal* is an element of $\mathbb{C}^d$, the linear space of d-dimensional complex vectors. We prefer the complex setting because the real case follows from a transparent adaptation. The usual Euclidean norm on signals will be written as $\|\cdot\|_2$. A *signal matrix* is drawn from $\mathbb{C}^{d\times K}$, the linear space of $d \times K$ complex matrices. The matrix space is equipped with the usual Hermitian inner product:

$$\langle \boldsymbol{S},\ \boldsymbol{X}\rangle \quad \stackrel{\text{def}}{=} \quad \operatorname{trace}(\boldsymbol{X}^*\,\boldsymbol{S}),$$

where the trace of a (square) matrix is the sum of diagonal entries. The Frobenius norm falls from this inner product:

$$\|\boldsymbol{S}\|_{\mathrm{F}}^2 \quad \stackrel{\text{def}}{=} \quad \langle \boldsymbol{S},\ \boldsymbol{S}\rangle\,.$$

The Frobenius norm will be used to measure the error in approximating a signal matrix.

We will frequently adopt the point of view that the columns of a signal matrix can be treated as K independent d-dimensional signals:

$$\boldsymbol{S} \quad = \quad \begin{bmatrix}\boldsymbol{s}_1 & \boldsymbol{s}_2 & \dots & \boldsymbol{s}_K\end{bmatrix}.$$

Note that the Frobenius norm of $\boldsymbol{S}$ can also be written as

$$\|\boldsymbol{S}\|_{\mathrm{F}}^2 \quad = \quad \sum\nolimits_{k=1}^{K} \|\boldsymbol{s}_k\|_2^2\,.$$

2.2. **The Dictionary.** Our goal will be to approximate each column of a signal matrix using a linear model called a *dictionary*. A dictionary is a finite collection $\mathscr{D}$ of unit-norm signals in $\mathbb{C}^d$. The elements of the dictionary are called *atoms*, and each atom is denoted by $\boldsymbol{\varphi}_\omega$, where the parameter ω ranges over an index set Ω. We use a general index set because the atoms may not admit a natural ordering and because it simplifies many of the technical arguments. The letter N will denote the number of atoms in the dictionary. It is clear that $N = |\mathscr{D}| = |\Omega|$, where $|\cdot|$ denotes the cardinality of a set. In summary, the whole dictionary structure can be written

$$\mathscr{D} \quad = \quad \{\boldsymbol{\varphi}_\omega \ : \ \omega \in \Omega\} \quad \subset \quad \mathbb{C}^d.$$

Let us emphasize that the atoms have unit Euclidean norm.

The choice of dictionary depends on the application. For many problems, the atoms are selected to resemble the coherent structures that appear in the input signals. In other applications, it is possible to design a dictionary that has properties favorable for sparse approximation. In this work, we assume that the dictionary has been predetermined.

From the dictionary, we may form a matrix whose ω-th column is the atom $\boldsymbol{\varphi}_\omega$. This matrix is denoted $\boldsymbol{\Phi}$, and it is called the *dictionary synthesis matrix*. Formally, the dictionary synthesis matrix belongs to the linear space[1] $\mathbb{C}^{d \times \Omega}$. We will rarely distinguish between the dictionary and its synthesis matrix. The conjugate transpose $\boldsymbol{\Phi}^*$ of the synthesis matrix is called the *dictionary analysis matrix*.

2.3. **Coherence.** One way of summarizing the behavior of the dictionary is to examine how the atoms are correlated with each other. To that end, we define the *coherence parameter* μ of the dictionary as

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

In words, the coherence is the cosine of the acute angle between the closest pair of atoms. Informally, we say that a dictionary is *incoherent* if we judge that μ is small. A valuable heuristic is that sparse approximation is easy for incoherent dictionaries. Nevertheless, one must recognize that incoherence is not fundamental to sparse approximation; it is only used to provide concrete results.

A generalization of the coherence parameter is the *cumulative coherence function* $\mu_1(\cdot)$. It is defined for each natural number t by the formula

$$\mu_1(t) \quad \stackrel{\text{def}}{=} \quad \max_{|\Lambda| \leq t} \ \max_{\omega \notin \Lambda} \ \sum\nolimits_{\lambda \in \Lambda} |\langle \boldsymbol{\varphi}_\omega, \ \boldsymbol{\varphi}_\lambda \rangle|$$

where the index set $\Lambda \subset \Omega$. We place the convention that $\mu_1(0) = 0$. Roughly, the cumulative coherence measures the maximum total correlation between a fixed atom and t distinct atoms. We have the trivial bound $\mu_1(t) \leq t \, \mu$ for each natural number t.

The coherence parameter was first introduced in [DMA97]. The cumulative coherence function was independently developed in [DE03, Tro04b].

2.4. **Coefficient Matrices.** Next we will develop a formal mechanism for synthesizing signal matrices using atoms from the dictionary. A *coefficient matrix* is an element of the linear space $\mathbb{C}^{\Omega \times K}$. The k-th column of a coefficient matrix $\boldsymbol{C}$ will be denoted by $\boldsymbol{c}_k$. The (ω, k) entry of the coefficient matrix is written as $c_{\omega k}$ or in functional notation as $\boldsymbol{C}(\omega, k)$. We will use whichever notation is typographically felicitous.

Given a coefficient matrix $\boldsymbol{C}$, observe that the matrix product $\boldsymbol{S} = \boldsymbol{\Phi}\,\boldsymbol{C}$ yields a signal matrix. It should be obvious that

$$\boldsymbol{s}_k \quad = \quad \boldsymbol{\Phi}\,\boldsymbol{c}_k \quad = \quad \sum\nolimits_{\omega \in \Omega} c_{\omega k}\, \boldsymbol{\varphi}_\omega \tag{2.1}$$

[1] In case this notation is unfamiliar, $\mathbb{C}^{d \times \Omega}$ is the set of functions from $\{1, \dots, d\} \times \Omega$ to $\mathbb{C}$. We equip this set with the usual addition of functions multiplication by complex scalars to form a linear space.

for each k. In other words, the k-th column of the signal matrix is synthesized with a linear combination of atoms whose coefficients are listed in the k-th column of the coefficient matrix.

Suppose that Λ is a subset of Ω. We will often consider coefficient matrices in $\mathbb{C}^{\Lambda \times K}$. Without notice, these small coefficient matrices may be treated as elements of $\mathbb{C}^{\Omega \times K}$ by extending them with zeros. Likewise, we may restrict a coefficient matrix in $\mathbb{C}^{\Omega \times K}$ to its nonzero rows. These transformations will be clear in context.

2.5. **Cost of Approximation.** A sparse approximation problem seeks an approximation of a signal matrix that can be expressed with low cost. In this work, the cost is measured as the total number of atoms that participate in the approximation. In light of (2.1), we see that a coefficient matrix uses the atom $\boldsymbol{\varphi}_\omega$ to synthesize a signal matrix if and only if the ω-th row of the coefficient matrix is nonzero. Therefore, the *row support* of a coefficient matrix is defined as the set of indices for its nonzero rows. More precisely,

$$\operatorname{rowsupp}(\boldsymbol{C}) \stackrel{\text{def}}{=} \{\omega \in \Omega \ : \ c_{\omega k} \neq 0 \text{ for some } k\}. \tag{2.2}$$

In particular, the support of a coefficient vector is the set of indices at which it is nonzero. Note that the definition (2.2) is not standard.

We define the row-ℓ_0 quasi-norm of a coefficient matrix to be the number of nonzero rows.

$$\|\boldsymbol{C}\|_{\text{row-0}} \stackrel{\text{def}}{=} |\operatorname{rowsupp}(\boldsymbol{C})|. \tag{2.3}$$

In particular, $\|\boldsymbol{c}\|_{\text{row-0}}$ is the number of nonzero entries in the vector $\boldsymbol{c}$. Therefore, the matrix ℓ_0 quasi-norm can also be calculated as

$$\|\boldsymbol{C}\|_{\text{row-0}} = \left| \bigcup\nolimits_{k=1}^{K} \operatorname{supp}(\boldsymbol{c}_k) \right|$$

where the $\boldsymbol{c}_k$ ranges over the columns of $\boldsymbol{C}$ and $\operatorname{supp}(\cdot)$ denotes the support of a vector. If we judge that a coefficient matrix has few nonzero rows, we may refer to it as *row-sparse*.

2.6. **Vector and Matrix Norms.** This work relies heavily on the use of matrix norms, some of which are probably unfamiliar. This subsection provides an overview of the tools that we will need, and it assumes a basic working knowledge of functional analysis. The casual reader may prefer to skip this material.

We will be working with vectors and matrices from finite-dimensional complex linear spaces. We equip these spaces with the usual Hermitian inner product, which generates the Euclidean or Frobenius norm. In addition, we will impose one or more norm structures. It is implicit that vectors and matrices are expressed with respect to the canonical coordinate basis.

If $\boldsymbol{x}$ is a vector, its ℓ_p norms are defined as

$$\begin{aligned} \|\boldsymbol{x}\|_p &\stackrel{\text{def}}{=} \left[\sum\nolimits_j |x_j|^p\right]^{1/p} \qquad \text{for } 1 \leq p < \infty, \text{ and} \\ \|\boldsymbol{x}\|_\infty &\stackrel{\text{def}}{=} \max\nolimits_j |x_j|. \end{aligned}$$

The dual of the normed linear space $(\mathbb{C}^m, \|\cdot\|_p)$ is the normed linear space $(\mathbb{C}^m, \|\cdot\|_{p'})$ with the conjugacy relation $1/p + 1/p' = 1$.

Suppose that X and Y are normed linear spaces of vectors or matrices. If $\boldsymbol{A}$ is a matrix with appropriate dimensions, we may view it as a linear operator acting on X via left matrix multiplication to produce elements of Y. Formally, the adjoint $\boldsymbol{A}^*$ is treated as a map between the dual spaces $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$. In the current setting, $\boldsymbol{A}^*$ is simply the conjugate transpose of $\boldsymbol{A}$, and it also acts by left matrix multiplication.

If the matrix $\boldsymbol{A}$ maps X to Y, its *operator norm* is defined as

$$\|\boldsymbol{A}\|_{X,Y} \stackrel{\text{def}}{=} \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|_Y}{\|\boldsymbol{x}\|_X}. \tag{2.4}$$

We will sometimes write this operator norm in the form $\|\boldsymbol{A}\|_{X \to Y}$. Meanwhile, the operator norm of the adjoint satisfies the identity

$$\|\boldsymbol{A}^*\|_{Y^*, X^*} \quad = \quad \|\boldsymbol{A}\|_{X, Y} \,.$$

In consequence of (2.4), we always have the following upper norm bound:

$$\|\boldsymbol{A}\,\boldsymbol{x}\|_Y \quad \leq \quad \|\boldsymbol{A}\|_{X, Y}\, \|\boldsymbol{x}\|_X \,.$$

It is also possible to develop a lower norm bound. The notation $\mathscr{R}(\cdot)$ indicates the range (i.e., column span) of a matrix, and the dagger marks a pseudo-inverse.

Proposition 2.1. *Every matrix $\boldsymbol{A}$ satisfies the following estimate.*

$$\min_{\substack{\boldsymbol{x} \in \mathscr{R}(\boldsymbol{A}^*) \\ \boldsymbol{x} \neq \boldsymbol{0}}} \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|_Y}{\|\boldsymbol{x}\|_X} \quad \geq \quad \left\|\boldsymbol{A}^\dagger\right\|_{Y, X}^{-1}. \tag{2.5}$$

If $\mathscr{R}(\boldsymbol{A}^*) = X$, then equality holds in (2.5)*. When $\boldsymbol{A}$ is invertible,*

$$\min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\,\boldsymbol{x}\|_Y}{\|\boldsymbol{x}\|_X} \quad = \quad \left\|\boldsymbol{A}^{-1}\right\|_{Y, X}^{-1} .$$

The proof follows the same lines as Proposition 3.2 from [Tro04d].

The norm on operators mapping from ℓ_p to ℓ_q will be written $\|\cdot\|_{p,q}$. Several of the (p, q) operator norms can be computed easily.

- The $(1, q)$ norm is the maximum ℓ_q norm of any column of $\boldsymbol{A}$.
- The $(2, 2)$ norm yields the maximum singular value of $\boldsymbol{A}$.
- The (p, ∞) norm is the maximum ℓ_p norm of any row of $\boldsymbol{A}$.

Note that the dual of the (p, q) operator norm is *not* generally the (q', p') operator norm. In particular, the dual of the (∞, ∞) norm is *not* the $(1, 1)$ norm. For more discussion of this point, see Section 3.2 of the second part of this paper [Tro04a].

We often view matrices as maps from one matrix space to another, and the attendant norms can become quite complicated. In several places, we will encounter the operator norm for maps between matrices equipped with the Frobenius norm and the (∞, ∞) norm. The following two results will be useful.

- If the matrix $\boldsymbol{A}$ has K rows, then $\|\boldsymbol{A}\|_{\mathrm{F} \to (\infty,\infty)} \leq K\, \|\boldsymbol{A}\|_{2,\infty}$.
- On the other hand, $\|\boldsymbol{A}\|_{(\infty,\infty) \to \mathrm{F}} = \|\boldsymbol{A}\|_{\infty,2}$.

Note that it is not trivial to establish the identity in the second bullet. Incidentally, it is generally NP-hard to calculate the $(\infty, 2)$ norm of a matrix in consequence of results from [Roh00].

3. Simultaneous Orthogonal Matching Pursuit

In this section, we present a greedy pursuit algorithm that can be used to solve several different simultaneous sparse approximation problems. To tune the algorithm for different problems, one simply changes the criterion for halting the algorithm.

3.1. Statement of Algorithm.

First, we give a formal description of the algorithm, and then we discuss some of its basic properties.

Algorithm 3.1 (S-OMP)**.**
Input:

- A $d \times K$ signal matrix $\boldsymbol{S}$
- A stopping criterion

Output:

- A set Λ_T containing T indices, where T is the number of iterations completed

- A $d \times K$ approximation matrix $\boldsymbol{A}_T$
- A $d \times K$ residual matrix $\boldsymbol{R}_T$

PROCEDURE:

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

$$\max_{\omega \in \Omega} \quad \sum\nolimits_{k=1}^{K} |\langle \boldsymbol{R}_{t-1}\, \mathbf{e}_k,\ \boldsymbol{\varphi}_\omega \rangle| \,.$$

We use $\mathbf{e}_k$ to denote the k-th canonical basis vector in $\mathbb{C}^K$.
(3) Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$.
(4) Determine the orthogonal projector $\boldsymbol{P}_t$ onto the span of the atoms indexed in Λ_t.
(5) Calculate the new approximation and residual:

$$\boldsymbol{A}_t = \boldsymbol{P}_t\, \boldsymbol{S}$$
$$\boldsymbol{R}_t = \boldsymbol{S} - \boldsymbol{A}_t.$$

(6) Increment t, and return to Step 2 unless the stopping criterion is met.

This procedure reduces to the usual Orthogonal Matching Pursuit [DMA97] when $K = 1$. Note that Chen and Huo have independently introduced an identical algorithm [CH04a, CH04b]. This algorithm seems to be distinct from the vector greedy algorithms studied by Temlyakov et al. in [LT03a, LT03b]. Please turn to Section 6 for comparison between S-OMP and other algorithms for simultaneous sparse approximation.

Step 2 of the algorithm is referred to as the *greedy selection*. The intuition behind maximizing the sum of absolute correlations is that we wish to find an atom that can contribute a lot of energy to every column of the signal matrix. This approach is likely to be most effective when all the input signals are well approximated by the same set of atoms. If the signals involve disparate sets of atoms, another greedy selection may be preferable. We will require the facts that the absolute sum in Step 2 equals $\|\boldsymbol{R}_{t-1}{}^*\, \boldsymbol{\varphi}_\omega\|_1$ and that its the maximum has the equivalent expressions

$$\max_{\omega \in \Omega} \quad \sum\nolimits_{k=1}^{K} |\langle \boldsymbol{R}_{t-1}\, \mathbf{e}_k,\ \boldsymbol{\varphi}_\omega \rangle| \quad = \quad \|\boldsymbol{R}_{t-1}{}^*\, \boldsymbol{\Phi}\|_{1,1} \quad = \quad \|\boldsymbol{\Phi}^*\, \boldsymbol{R}_{t-1}\|_{\infty,\infty} \,.$$

Steps 4 and 5 have been written to emphasize the conceptual structure of the algorithm. It is possible to implement them much more efficiently using standard techniques for least-squares problems. See [GVL96, Ch. 5] for extensive details. It is important to recognize that each column of the residual $\boldsymbol{R}_t$ is orthogonal to the atoms indexed in Λ_t. Therefore, no atom is ever chosen twice.

Remark 3.2. Note that the greedy selection can also be weakened. Instead of searching for an atom that maximizes the absolute sum, we could also choose an atom that comes within a constant factor of the maximum. This type of weak greedy step may admit a more efficient implementation [GMS03]. Weak greedy algorithms for simultaneous sparse approximation are discussed in [LT03a, LT03b].

3.2. **Stopping Criteria.** Since Simultaneous Orthogonal Matching Pursuit is an iterative algorithm, we must supply a method for deciding when to halt the iteration. There are three obvious possibilities:

(1) Stop the algorithm after a fixed number T of iterations, i.e., when $t = T$.
(2) Wait until the Frobenius norm of the residual declines to a level ε. That is, $\|\boldsymbol{R}_t\|_{\mathrm{F}} \leq \varepsilon$.
(3) Halt the algorithm when the maximum total correlation between an atom and the residual drops below a threshold τ. In symbols, $\|\boldsymbol{\Phi}^*\, \boldsymbol{R}_t\|_{\infty,\infty} \leq \tau$.

In the sequel, we will see how these stopping rules apply to different flavors of simultaneous sparse approximation.

4. Numerical Experiments

We have performed some numerical experiments to demonstrate the simultaneous sparse approximation problems are a valuable extension of simple sparse approximation. These experiments also provide confirmation that the Simultaneous Orthogonal Matching Pursuit algorithm can solve these problems in practice.

In this section, we will be working with the Dirac–Fourier dictionary, which consists of impulses and complex exponentials. In a d-dimensional signal space, the dictionary contains $2d$ atoms: $\boldsymbol{\varphi}_\omega[t] = \boldsymbol{\delta}_\omega[t]$ for $\omega = 1, \dots, d$ and also $\boldsymbol{\varphi}_\omega[t] = \mathrm{e}^{-2\pi \mathrm{i}\, t\omega/d}$ for $\omega = d+1, \dots, 2d$. Note that this dictionary has coherence $\mu = 1/\sqrt{d}$.

Our test signals will be formed using three different models. In each case, the signals take the form $\boldsymbol{s}_k = \boldsymbol{x}_k + \boldsymbol{\nu}_k$, where $\boldsymbol{\nu}_k$ is random noise and where $\boldsymbol{x}_k$ can be expressed using a linear combination of T atoms. The models differ in the way the atoms and their coefficients are drawn. In each case, we will seek a sparse approximation of the K input signals using a total of T atoms, so we halt S-OMP after precisely T iterations.

For our first experiment, we constructed signals of the form

$$\boldsymbol{s}_k \quad = \quad \sum_{j=1}^{T} \alpha_{jk}\, \boldsymbol{\varphi}_{\omega_{jk}}. \tag{I}$$

For each signal $\boldsymbol{s}_k$, we select T distinct atoms independently and uniformly from the dictionary. The coefficients α_{jk} are drawn from i.i.d. normal distributions with zero mean and unit variance. Our goal is to identify the best T atoms with which to represent all K signals, each of which is a linear combination of T atoms. We have observed that the S-OMP algorithm always recovers T atoms from the collection of approximately KT distinct atoms that participate in the K input signals. Indeed, all of the error in the residual is due to the fact that the input signals involve more atoms than we are allowed to use. We omit the figures since they are not very illuminating.

The second type of input signal has the form

$$\boldsymbol{s}_k \quad = \quad \sum_{j=1}^{T} \alpha_{jk}\, \boldsymbol{\varphi}_{\omega_j} \tag{II}$$

For all K signals, we use the same core of T atoms, but the coefficients α_{jk} are chosen from i.i.d. normal distributions. For these experiments, we fix the number of signals at $K = 2$ and the dimension of the signal space at $d = 128$. We vary the value of T to explore how many core atoms we can successfully recover with our algorithm. For each set of parameters, we performed 1000 independent trials. We computed the Hamming distance between the set of recovered atoms and the core set. (Hamming distance zero means that we recover the entire core set, while distance one means that we fail to recover any of the core atoms.) In Figure 1, we plot the average Hamming distance as a function of T. The error bars mark one standard deviation from the mean. Even when $T = 90$ (in dimension 128!), we typically recover most of the core set. We can see from this figure that our theoretical bounds are far too pessimistic in this case.

We have performed a more detailed version of the same experiment with $K = 1$. The results are displayed in Figure 2. In this case, the problem is no longer *simultaneous* sparse approximation since we have only one input signal Although the performance of the algorithm when $K = 1$ is similar to the performance when $K = 2$, an additional input vector does provide a slight improvement in the success rate. That is, the average Hamming distance between the core set and the recovered set is greater when $K = 1$ than when $K = 2$. For a theoretical explanation of a closely related phenomenon, see the forthcoming paper [TG05].

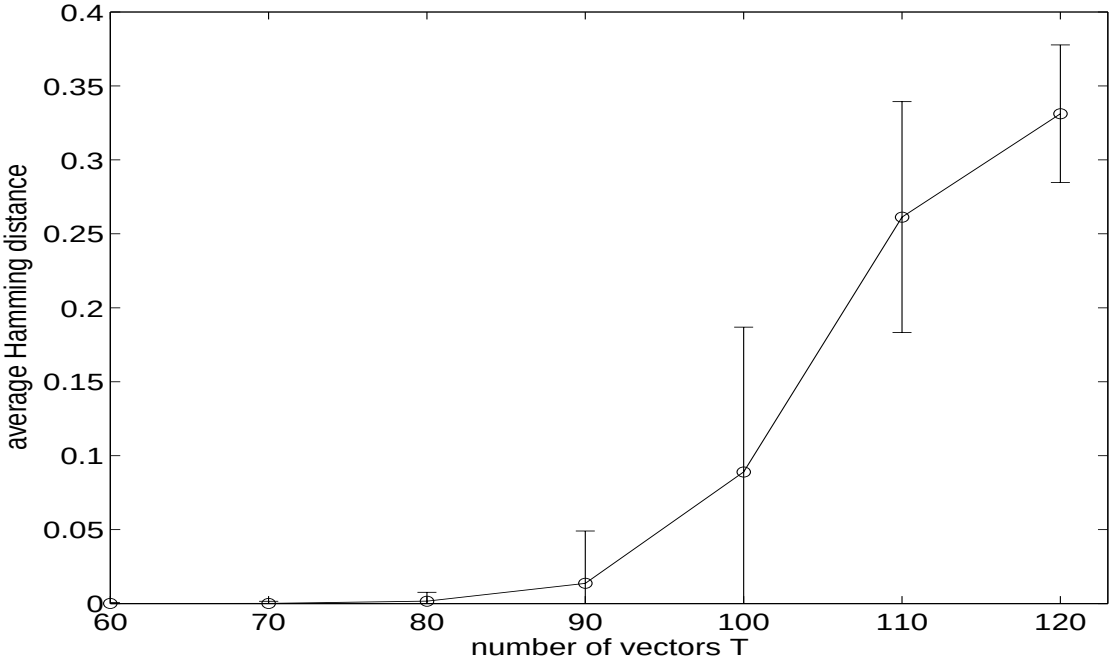

FIGURE 1. (Input type II) The average Hamming distance between the core set of the vectors and the recovered set as a function of the number of vectors T in the core set. In this experiment, $d = 128$ and $K = 2$. Compare with Figure 2.

Our third set of experiments involves input signals of the form

$$\boldsymbol{s}_k = \sum_{j=1}^{T} \alpha_j \, \boldsymbol{\varphi}_{\omega_j} + \boldsymbol{\nu}_k. \tag{III}$$

In this case, we choose T atoms at random and form a linear combination with random coefficients $\alpha_j \in \{\pm 1\}$. Here, the coefficients have unit magnitude to ensure that the noise does not obliterate any of the atoms. Then we construct K input signals by corrupting the original signal with i.i.d. additive white Gaussian noise $\boldsymbol{\nu}_k$. For these experiments, we fix the dimension $d = 256$; we vary T from 2 to 4; we vary K from 2 to 6; and we examine SNR values of 10, 13, 16, and 20 dB. For each parameter set, we perform 1000 trials. Figure 3 displays the average Hamming distance as a function of the number of signals. For each value of T, we use a distinct line type (e.g., dashed), so the four dashed lines correspond to four different SNR values. Naturally, the Hamming distance increases as the SNR decreases. Observe that, independent of the number of core atoms T and the SNR, we recover the core signal better when we have more observations. Furthermore, the presence of noise has a significant effect on the performance of the algorithm. The previous example showed that we can often recover core sets of atoms that are almost as large as the dimension of the signal space. Yet for moderate SNR (e.g., 13 dB), we cannot reliably recover three atoms in a 256-dimensional signal space. With the parameter settings we have chosen, Theorem 5.1 of the sequel predicts that

$$\frac{\|\boldsymbol{S} - \boldsymbol{A}_T\|_{\mathrm{F}}^2}{\sum_{k=1}^{K} \|\boldsymbol{\nu}_k\|_2^2} \quad \leq \quad 1 + 3\,KT.$$

To see if this bound accurately identifies the dependence of the error on K and T, we plot in Figure 4 the total relative error as a function of the number of signals K. For each T, we use a

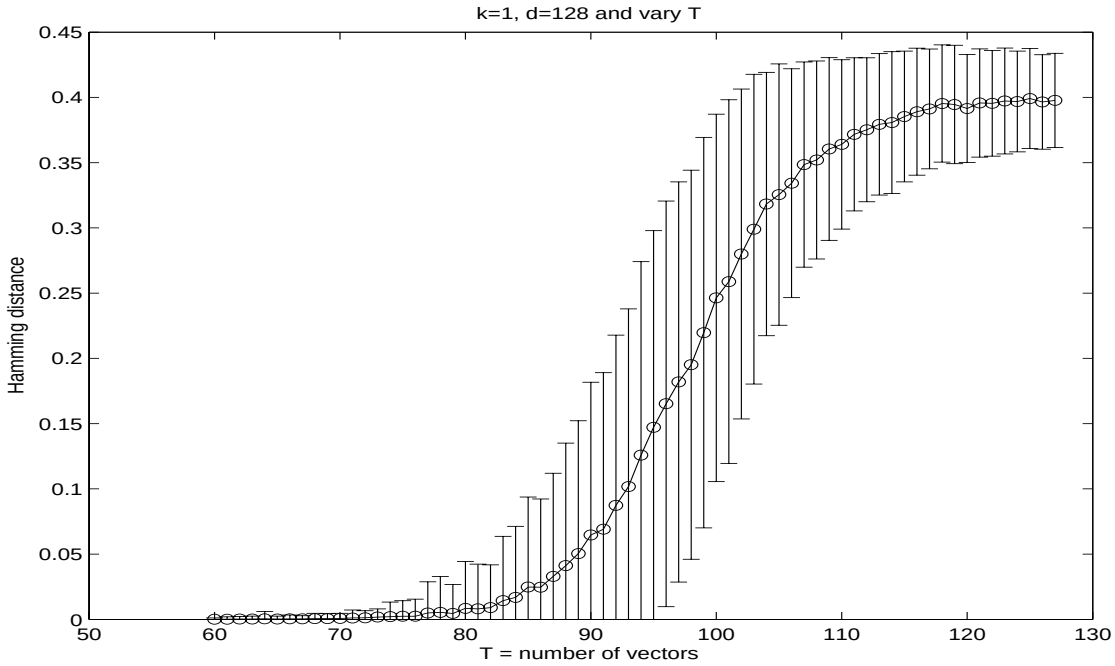

FIGURE 2. (Input type II) The average Hamming distance between the core set of the vectors and the recovered set as a function of the number of vectors T in the core set. In this experiment, $d = 128$ and $K = 1$. Compare with Figure 1.

different line type. The two groups of lines represent the extreme SNR values (10 and 20 dB). The plot shows that the size of T has a negligible effect on the error. That is, the theoretical bounds reflect a dependence on T that is absent in the empirical evidence. Again, our algorithm performs better than the theoretical results might lead us to believe.

In the foregoing, we have observed that for fixed values of T and SNR, we recover the signal better as the number of observations K increases. We cannot, however, effectively recover a small number of atoms in the presence of noise. To explore more thoroughly the interaction between noise and the number of atom T in our sparse representation, we fix the number of observations $K = 1$ and examine the average Hamming distance between the core and the recovered sets of vectors as a function of T and the SNR. While we cannot hope to recover many of the atoms in the core set when the noise level is significant, Figure 5 shows that we do recover a higher fraction of them as the size of the core set grows. We also observe that the error grows as a function of T vectors in the core set until it hits a maximum value and then decreases as we add more to the core set. A more thorough analysis of this intriguing and surprising behavior is beyond the scope of this paper. It is possible that some of the effect is due to the noise overwhelming atoms with small coefficients.

5. Performance Guarantees for S-OMP

In this section, we will develop theoretical results for the performance of S-OMP for different types of simultaneous sparse approximation problems. These problems reflect different types of *a priori* information that we might have about the desired approximation input signal. For example, we might know that the approximation involves only T atoms or that the error in the residual is

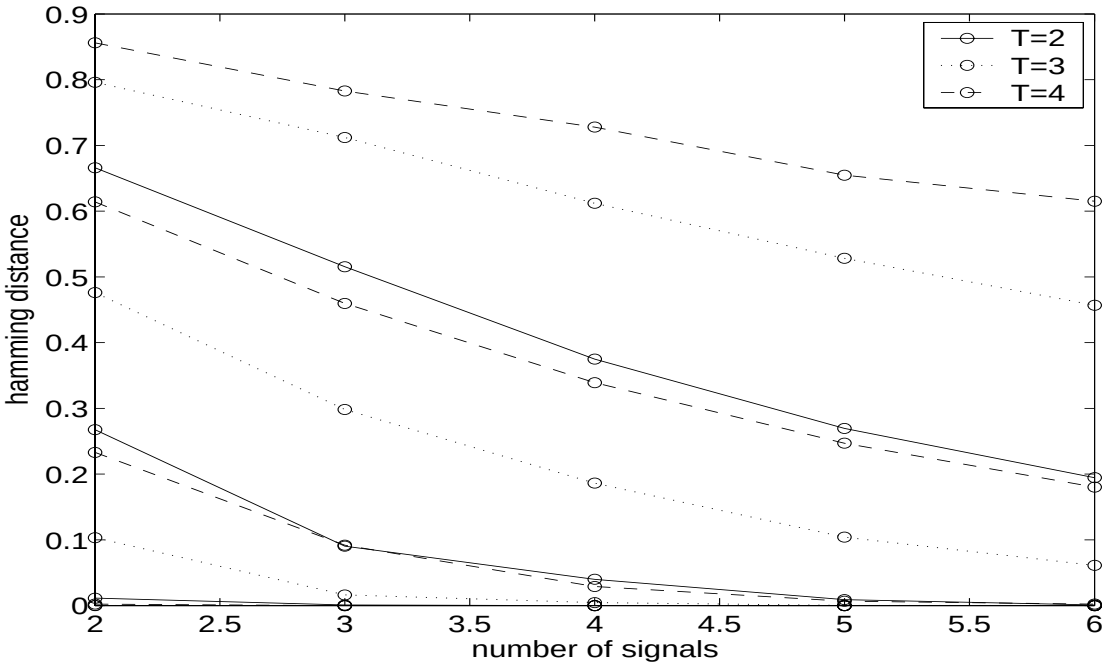

FIGURE 3. (Input type III) The average Hamming distance between the core set of vectors and the recovered set as a function of the number of signals and the SNR. Each line type corresponds to a different number of atoms. For each line type, the bottom line reflects the highest SNR (20 dB) and the top line the lowest (10 dB).

bounded. By varying the stopping criterion, we can tune S-OMP to obtain the best performance for each problem.

5.1. **Approximation with a Sparsity Bound.** In the numerical experiments, we considered how to approximate the columns of a signal matrix $\boldsymbol{S}$ using different linear combinations of the same T atoms. We may state this problem more rigorously as

$$\min_{\boldsymbol{C} \in \mathbb{C}^{\Omega \times K}} \|\boldsymbol{S} - \boldsymbol{\Phi}\,\boldsymbol{C}\|_{\mathrm{F}} \qquad \text{subject to} \qquad \|\boldsymbol{C}\|_{\text{row-0}} \leq T. \tag{SPARSE}$$

Note that a solution to (SPARSE) is a coefficient matrix, not a signal matrix. If $\boldsymbol{C}_{\mathrm{opt}}$ solves the optimization problem, the corresponding approximation of the signal matrix is $\boldsymbol{A}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{C}_{\mathrm{opt}}$.

To solve the sparsity-constrained approximation problem, we may apply the S-OMP algorithm, stopping at the end of T iterations. At this stage, the algorithm produces an approximation of $\boldsymbol{S}$ using at most T atoms. We have the following theoretical guarantee.

Theorem 5.1 (S-OMP with a Sparsity Bound)**.** *Assume that $\mu_1(T) < \frac{1}{2}$. Given an input matrix $\boldsymbol{S}$, suppose that $\boldsymbol{C}_{\mathrm{opt}}$ solves* (SPARSE) *and that $\boldsymbol{A}_{\mathrm{opt}} = \boldsymbol{\Phi}\,\boldsymbol{C}_{\mathrm{opt}}$. After T iterations, S-OMP will produce an approximation $\boldsymbol{A}_T$ that satisfies the error bound*

$$\|\boldsymbol{S} - \boldsymbol{A}_T\|_{\mathrm{F}} \quad \leq \quad \left[1 \;+\; KT\,\frac{1-\mu_1(T)}{[1-2\,\mu_1(T)]^2}\right]^{1/2} \|\boldsymbol{S} - \boldsymbol{A}_{\mathrm{opt}}\|_{\mathrm{F}}\,. \tag{5.1}$$

In words, S-OMP is an approximation algorithm for (SPARSE)*.*

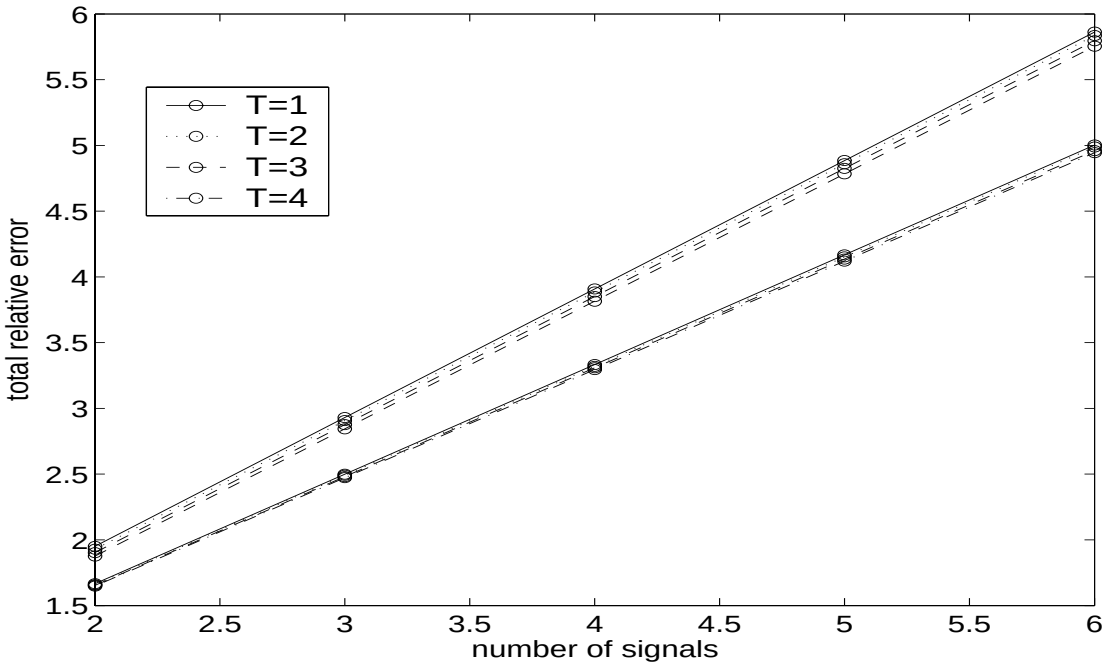

FIGURE 4. (Input type III) The total relative error as a function of the number of signals and the number of core vectors for two values of SNR. The bottom group of lines corresponds to an SNR of 20 dB and the top group corresponds to 10 dB.

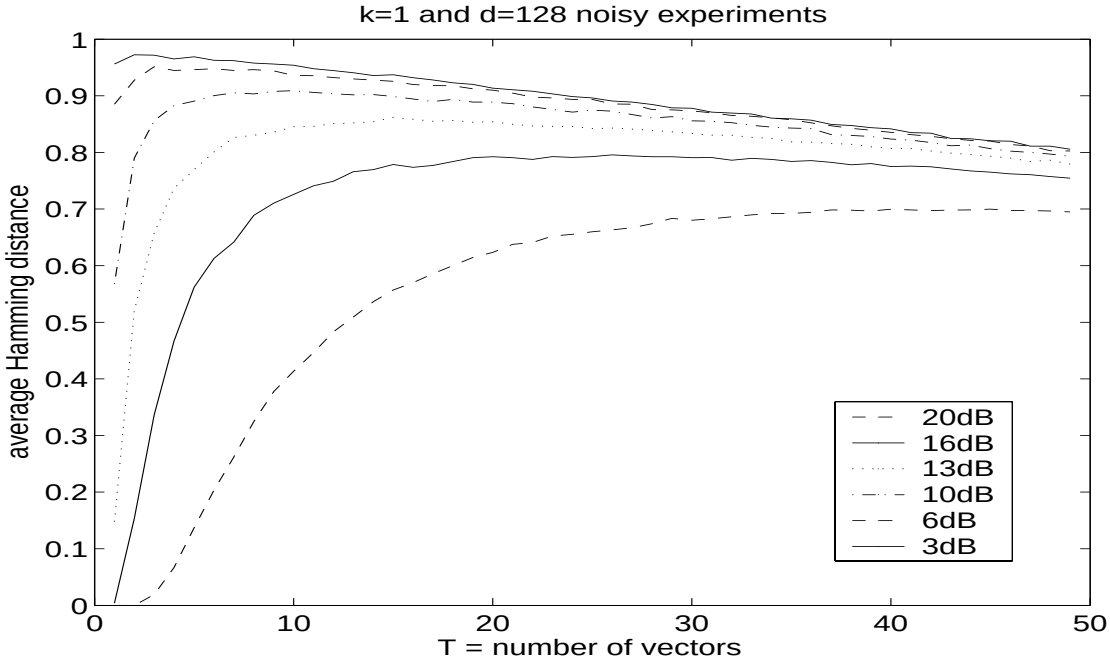

FIGURE 5. (Input type III) The average Hamming distance between the core set of vectors and the recovered set as a function of the number of signals and the SNR.

A sketch of the proof appears in [TGS04]. It follows the basic outline of an analogous result for simple sparse approximation from [Tro04b]. In fact, when $K = 1$, we retrieve Corollary 4.3 from [Tro04b].

Some comments may clarify the meaning of this theorem and its limitations. First, observe that the error in the computed approximation is never more than a constant factor greater than the optimal approximation error. In particular, if the signal matrix can be expressed exactly using T atoms or fewer, then S-OMP will also produce an exact representation of the signal matrix. Indeed, a slightly weaker version of Theorem 4.3 from [CH04a] can be obtained as a special case of this result.

Suppose that the dictionary is orthonormal, so the cumulative coherence function is identically zero. We see that the bracket in (5.1) simplifies to $\sqrt{1 + KT}$. Unfortunately, the theorem offers an overly pessimistic assessment of the algorithm's performance in this case. Indeed, the factor of T in the constant seems to be an artifact of the proof method, and it could probably be eliminated with a more subtle approach. It is also possible that the S-OMP algorithm is more appropriate for an error measure slightly different from the Frobenius norm. These points deserve further attention, but they lie beyond the scope of the present work.

Finally, we remark that properties of the cumulative coherence function can be used to simplify the theorem. For example, one may apply the bound $\mu_1(T) \leq T\,\mu$. In addition, we can obtain a more quantitative result by placing a sharper restriction on $\mu_1(T)$. If $\mu_1(T) \leq \frac{1}{3}$, then

$$\frac{1 - \mu_1(T)}{(1 - 2\,\mu_1(T))^2} \quad \leq \quad 6.$$

Of course, tighter bounds on the cumulative coherence will lead to better results.

5.2. **Approximation with an Error Bound.** Suppose that our signal matrix consists of multiple views of an artificially generated sparse signal contaminated by bounded noise. Then we may have information about the ideal set of atoms for approximating the signal matrix as well as a bound on the approximation error. In this setting, we can prove the following theorem on the performance of S-OMP.

Theorem 5.2 (S-OMP with an Error Bound)**.** *Let Λ_{opt} be an index set containing T atoms or fewer, where $\mu_1(T) < \frac{1}{2}$. Suppose that the best approximation $\boldsymbol{A}_{\text{opt}}$ of the signal matrix $\boldsymbol{S}$ over Λ_{opt} satisfies the error bound*

$$\|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{\text{F}} \quad \leq \quad \varepsilon.$$

Let us halt S-OMP at the end of iteration t if the norm of the residual satisfies

$$\|\boldsymbol{R}_t\|_{\text{F}} \quad \leq \quad \left[1 \;+\; KT\,\frac{1 - \mu_1(T)}{(1 - 2\,\mu_1(T))^2}\right]^{1/2} \varepsilon.$$

It follows that each atom chosen is optimal, i.e., $\Lambda_t \subset \Lambda_{\text{opt}}$.

The proof follows the same lines as the analogous theorem for simple sparse approximation [Tro04e, Thm. 5.9], so we do not reproduce it. In fact, it is possible to develop a rather more sophisticated result using the same techniques, but we have opted to present a more comprehensible version.

In words, the theorem says that the algorithm can calculate an approximation that achieves an error within a constant factor of ε. Meanwhile, it guarantees that every atom participating in the computed approximation is drawn from the ideal set of atoms.

If the dictionary is orthonormal, the bracketed constant equals $\sqrt{1 + KT}$. Unlike the case for (SPARSE), the factor of T is essential in this result. See Section 5.3 of [Tro04e] for some discussion of this fact in the case of simple sparse approximation. We may also apply the cumulative coherence function to develop simpler versions of the result. See the previous subsection for details.

5.3. **Approximation with a Correlation Bound.** Suppose now that our signal matrix consists of multiple observations of a sparse signal contaminated with additive noise. It is often possible to develop estimates on the correlation between the dictionary and the residual left over in approximation. For an example, see [Tro04c, Sec. IV-D]. If bounds on this correlation are available, it is possible to develop strong results on the performance of S-OMP. To make the statement of the theorem more transparent, define the quantity

$$M(t) \quad = \quad M(t;T) \quad \stackrel{\text{def}}{=} \quad \frac{\mu_1(T-t)}{1 \; - \; \mu_1(t)} \qquad \text{for } t = 0, \dots, T.$$

We will discuss this function more in a moment.

Theorem 5.3 (S-OMP with a Correlation Bound)**.** *Suppose that Λ_{opt} lists at most T atoms, where $\mu_1(T) < \frac{1}{2}$ and $M(T) < \frac{1}{2}$. Let $\boldsymbol{S}$ be a signal matrix, $\boldsymbol{A}_{\text{opt}}$ its best approximation over Λ_{opt}, and $\boldsymbol{C}_{\text{opt}}$ the coefficient matrix that synthesizes $\boldsymbol{A}_{\text{opt}}$. Finally, assume we have a bound*

$$\|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_{\text{opt}})\|_{\infty,\infty} \quad \leq \quad \tau.$$

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

$$\|\boldsymbol{\Phi}^* \boldsymbol{R}_t\|_{\infty,\infty} \quad \leq \quad \frac{1 - M(t)}{1 - 2\,M(t)}\,\tau.$$

If the algorithm terminates at the end of iteration t, we may conclude that

- *the algorithm has chosen t indices from Λ_{opt}, and*
- *it has identified every index λ from Λ_{opt} for which*

$$\sum\nolimits_{k=1}^{K} |\boldsymbol{C}_{\text{opt}}(\lambda, k)| \quad > \quad \frac{\tau}{1 - 2\,M(t)}.$$

- *The absolute error in the computed approximation satisfies*

$$\|\boldsymbol{S} - \boldsymbol{A}_t\|_{\text{F}}^2 \quad \leq \quad \|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{\text{F}}^2 \; + \; \tau^2 \left[\frac{1 - M(t)}{1 - 2\,M(t)}\right]^2 \frac{T - t}{1 - \mu_1(T-t)}.$$

- *In particular, if*

$$\min_{\lambda \in \Lambda_{\text{opt}}} \sum\nolimits_{k=1}^{K} |\boldsymbol{C}_{\text{opt}}(\lambda, k)| \quad > \quad \frac{\tau}{1 - 2\,M(t)}$$

then $t = T$, $\Lambda_t = \Lambda_{\text{opt}}$, and $\boldsymbol{A}_t = \boldsymbol{A}_{\text{opt}}$.

This result is qualitatively new, although an analogous theorem for simple sparse approximation has been announced without proof in [Tro04d]. Donoho, Elad, and Temlyakov have announced a related theorem for simple sparse approximation [DET04, Thm. 4.1]. Their result is contained in the last bullet of Theorem 5.3. We will provide a complete demonstration of the theorem in the next subsection.

In words, Theorem 5.3 says that S-OMP can be used to recover all the atoms whose coefficients are sufficiently large provided that the maximum total correlation between the residual and the remaining atoms is small. If a bound on the norm of the residual is available, it is possible to develop simple bounds on the maximum correlation. As examples,

$$\begin{aligned} \|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_{\text{opt}})\|_{\infty,\infty} \quad &\leq \quad \sqrt{K}\, \|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{\text{F}} \\ \|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_{\text{opt}})\|_{\infty,\infty} \quad &\leq \quad K\, \|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{1,2}\,. \end{aligned}$$

It follows that, if the signal matrix can be represented perfectly using the atoms in Λ_{opt}, then we can choose $\tau = 0$. Therefore, S-OMP will recover the signal matrix perfectly (provided the other hypotheses of the theorem). It is also worth noting that this theorem is essentially as strong as the analogous result for convex relaxation [Tro04a, Cor. 5.1].

We may use properties of the cumulative coherence function to simplify the theorem. Provided that $T\,\mu < 1$, we have the bound

$$M(t) \quad \leq \quad \frac{(T-t)\,\mu}{1-t\,\mu} \qquad \text{for each } t = 0, \dots, T.$$

Maximizing the right-hand side over t, we discover that the extreme value occurs when $t = 0$. Therefore,

$$M(t) \quad \leq \quad T\,\mu.$$

If we place numerical restrictions, we can obtain a more quantitative result. For example, if $T\,\mu \leq \frac{1}{3}$, then we discover that

$$\frac{1-M(t)}{1-2\,M(t)} \quad \leq \quad 2 \qquad \text{and} \qquad \left[\frac{1-M(t)}{1-2\,M(t)}\right]^2 \frac{1}{1-\mu_1(T-t)} \quad \leq \quad 6.$$

As usual, tighter bounds lead to sharper versions of the theorem.

5.4. **Proof of Theorem 5.3.** Now we will prove the theorem on the performance of the algorithm. The major technical challenge is to develop bounds on the maximum total correlation between an optimal atom and the residual matrix at iteration t. We retain the notation from the statement of Theorem 5.3 and from the statement of the S-OMP algorithm. Finally, given an index set Λ_{opt}, we will partition the dictionary matrix as

$$\boldsymbol{\Phi} \quad = \quad [\,\boldsymbol{\Phi}_{\text{opt}} \mid \boldsymbol{\Psi}_{\text{opt}}\,],$$

where $\boldsymbol{\Phi}_{\text{opt}}$ contains the columns listed in Λ_{opt} and $\boldsymbol{\Psi}_{\text{opt}}$ contains the remaining columns.

Lemma 5.4. *Assume that $\Lambda_t \subset \Lambda_{\text{opt}}$, and let $\boldsymbol{C}_{\sim t}$ be the $(T-t) \times K$ matrix formed from the rows of $\boldsymbol{C}_{\text{opt}}$ listed in $\Lambda_{\text{opt}} \setminus \Lambda_t$. Then the following bounds are in force.*

$$\begin{aligned} \|\boldsymbol{\Phi}_{\text{opt}}{}^*\,\boldsymbol{R}_t\|_{\infty,\infty} \quad &\geq \quad (1 - M(t))\,\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \\ \|\boldsymbol{\Psi}_{\text{opt}}{}^*\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty} \quad &\leq \quad M(t)\,\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \end{aligned}$$

Proof. This lemma contains the most difficult part of the proof, and it is further complicated by a heavy notational burden. Since $\Lambda_t \subset \Lambda_{\text{opt}}$, we may split the optimal synthesis matrix $\boldsymbol{\Phi}_{\text{opt}}$ into two pieces. The matrix $\boldsymbol{\Phi}_t$ contains the t columns that are indexed by Λ_t and the matrix $\boldsymbol{\Phi}_{\sim t}$ contains the remaining $(T-t)$ columns. Then, we write $\boldsymbol{C}_{\sim t}$ for the matrix formed from the $(T-t)$ rows of $\boldsymbol{C}_{\text{opt}}$ indexed by $\Lambda_{\text{opt}} \setminus \Lambda_t$. As usual, $\boldsymbol{P}_t$ indicates the orthogonal projector onto the range of $\boldsymbol{\Phi}_t$.

First, observe that

$$\boldsymbol{\Phi}_{\text{opt}}{}^*\,\boldsymbol{R}_t \quad = \quad \boldsymbol{\Phi}_{\text{opt}}{}^*\,(\boldsymbol{S} - \boldsymbol{A}_t) \quad = \quad \boldsymbol{\Phi}_{\text{opt}}{}^*\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)$$

since $(\boldsymbol{S} - \boldsymbol{A}_{\text{opt}})$ is orthogonal to the range of $\boldsymbol{\Phi}_{\text{opt}}$. Then rewrite the matrix $(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)$ in the following manner:

$$\begin{aligned} \boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t \quad &= \quad (\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{A}_{\text{opt}} \\ &= \quad (\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\text{opt}}\,\boldsymbol{C}_{\text{opt}} \\ &= \quad (\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}\,\boldsymbol{C}_{\sim t}. \end{aligned}$$

The last equality holds because $(\mathbf{I} - \boldsymbol{P}_t)$ annihilates the atoms indexed in Λ_t. Therefore, our goal is to produce a lower bound for

$$\|\boldsymbol{\Phi}_{\text{opt}}{}^*\,\boldsymbol{R}_t\|_{\infty,\infty} \quad = \quad \|\boldsymbol{\Phi}_{\sim t}{}^*\,(\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \tag{5.2}$$

and to produce an upper bound for

$$\|\boldsymbol{\Psi}_{\text{opt}}{}^*\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty} \quad = \quad \|\boldsymbol{\Psi}_{\text{opt}}{}^*\,(\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,. \tag{5.3}$$

These last two equations are the key to the proof. We will use the cumulative coherence function to develop the estimates.

We begin with a bound on (5.3), since we will reuse our observations when we study (5.2). First, apply the standard norm bound to (5.3) and invoke the triangle inequality.

$$\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty} \quad \leq \quad \left[\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty} + \|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{P}_t\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty}\right] \|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,. \tag{5.4}$$

We must find an upper bound for the bracketed term in (5.4). This process takes five steps.

(1) Each row of the matrix $(\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{\Phi}_{\sim t})$ lists the inner products between an atom and $(T-t)$ distinct atoms. It follows that $\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty} \leq \mu_1(T-t)$.

(2) Rewrite the orthogonal projector as $\boldsymbol{P}_t = \boldsymbol{\Phi}_t\,(\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t{}^{*}$, and use the fact that $\|\cdot\|_{\infty,\infty}$ is submultiplicative to see that

$$\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{P}_t\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty} \quad \leq \quad \|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{\Phi}_t\|_{\infty,\infty}\,\left\|(\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_t)^{-1}\right\|_{\infty,\infty}\,\|\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty}\,.$$

(3) Using the same rationale as in Step (1), we see that $\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,\boldsymbol{\Phi}_t\|_{\infty,\infty} \leq \mu_1(t)$. Similarly, $\|\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty} \leq \mu_1(T-t)$.

(4) Note that the Gram matrix $(\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_t)$ lists the inner products between t different atoms. Since the atoms are normalized, the matrix has a unit diagonal. Therefore, we may write

$$\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_t \quad = \quad \mathbf{I} - \boldsymbol{X}$$

where $\|\boldsymbol{X}\|_{\infty,\infty} \leq \mu_1(t)$. We expand the inverse in a Neumann series and make the standard estimate to obtain

$$\left\|(\boldsymbol{\Phi}_t{}^{*}\,\boldsymbol{\Phi}_t)^{-1}\right\|_{\infty,\infty} \quad \leq \quad \frac{1}{1-\mu_1(t)}.$$

For details, refer to the proof of Proposition 3.6 in [Tro04d].

(5) Introduce the bounds from Steps (1)–(4) into (5.4). We reach

$$\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty} \quad \leq \quad \left[\mu_1(T-t) \;+\; \frac{\mu_1(t)\,\mu_1(T-t)}{1-\mu_1(t)}\right] \|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,.$$

Finally, simplify the bracketed expression to conclude that

$$\|\boldsymbol{\Psi}_{\text{opt}}{}^{*}\,(\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty} \quad \leq \quad \frac{\mu_1(T-t)}{1-\mu_1(t)}\,\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,.$$

Identify $M(t)$ to complete the first bound.

It remains to provide a lower bound for $\|\boldsymbol{\Phi}_{\sim t}{}^{*}\,(\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty}$. For the moment, let us abbreviate $\boldsymbol{Y} = \boldsymbol{\Phi}_{\sim t}{}^{*}\,(\mathbf{I} - \boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}$. The lower norm bound of Proposition 2.1 gives

$$\|\boldsymbol{Y}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \quad \geq \quad \left\|\boldsymbol{Y}^{-1}\right\|_{\infty,\infty}^{-1}\,\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}$$

Write $\boldsymbol{Y} = \mathbf{I} - (\mathbf{I} - \boldsymbol{Y})$, and expand its inverse in a Neumann series. We make the usual estimates to obtain a geometric series, and we sum this series to reach

$$\|\boldsymbol{Y}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \quad \geq \quad \Big(1 \;-\; \|\mathbf{I} - \boldsymbol{Y}\|_{\infty,\infty}\Big)\,\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,. \tag{5.5}$$

Recall the definition of $\boldsymbol{Y}$ and apply the triangle inequality to obtain

$$\|\mathbf{I} - \boldsymbol{Y}\|_{\infty,\infty} \quad \leq \quad \|(\mathbf{I} - \boldsymbol{\Phi}_{\sim t}{}^{*}\,\boldsymbol{\Phi}_{\sim t})\|_{\infty,\infty} \;+\; \|\boldsymbol{\Phi}_{\sim t}^{*}\,\boldsymbol{P}_t\,\boldsymbol{\Phi}_{\sim t}\|_{\infty,\infty}\,.$$

This bound on the right-hand side of this inequality is completely analogous with the bound we made in Steps (1)–(5). Repeating these steps, *mutatis mutandis*, we have

$$\|\mathbf{I} - \boldsymbol{Y}\|_{\infty,\infty} \quad \leq \quad \frac{\mu_1(T-t)}{1-\mu_1(t)}. \tag{5.6}$$

Introduce (5.6) into (5.5) and write out $\boldsymbol{Y}$ in full to conclude

$$\|\boldsymbol{\Phi}_{\sim t}{}^{*}\,(\mathbf{I}-\boldsymbol{P}_t)\,\boldsymbol{\Phi}_{\sim t}\,\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \quad\geq\quad \left[1\;-\;\frac{\mu_1(T-t)}{1-\mu_1(t)}\right]\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty}\,.$$

Identify $M(t)$ to complete the argument. □

With Lemma 5.4 at hand, it is easy to develop a condition which ensures that S-OMP chooses an optimal atom in the $(t+1)$-st iteration.

Lemma 5.5 (Optimal Atom Selection)**.** *Assume that $\Lambda_t \subset \Lambda_{\mathrm{opt}}$. Provided that*

$$\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad>\quad \frac{1-M(t)}{1-2\,M(t)}\,\|\boldsymbol{\Phi}^{*}\,(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\|_{\infty,\infty}\,,$$

S-OMP will identify another optimal atom in the $(t+1)$-st iteration.

Proof. Assume that the greedy selection identifies a nonoptimal atom in the $(t+1)$-st iteration. We will use this hypothesis to develop an upper bound on $\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty}$. The condition in the statement of the theorem is the logical negation of this upper bound. Therefore, the condition ensures that the algorithm must identify an optimal atom in the $(t+1)$-st iteration.

The assumption that greedy selection picks a nonoptimal atom is equivalent with the relations

$$\|\boldsymbol{\Phi}_{\mathrm{opt}}{}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad\leq\quad \|\boldsymbol{\Psi}_{\mathrm{opt}}{}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \tag{5.7}$$

$$\|\boldsymbol{\Psi}_{\mathrm{opt}}{}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad=\quad \|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty}\,. \tag{5.8}$$

In words, the maximum correlation between an optimal atom and the residual is smaller than the maximum correlation between a nonoptimal atom and the residual, which equals the maximum correlation between any atom and the residual. We begin our calculation by rewriting

$$\boldsymbol{R}_t \quad=\quad (\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\;+\;(\boldsymbol{A}_{\mathrm{opt}}-\boldsymbol{A}_t). \tag{5.9}$$

Invoke (5.8) and (5.9). Then apply the triangle inequality to reach

$$\begin{aligned}\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad&=\quad \|\boldsymbol{\Psi}_{\mathrm{opt}}{}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty}\\ &\leq\quad \|\boldsymbol{\Psi}_{\mathrm{opt}}{}^{*}\,(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\|_{\infty,\infty}\;+\;\|\boldsymbol{\Psi}_{\mathrm{opt}}{}^{*}\,(\boldsymbol{A}_{\mathrm{opt}}-\boldsymbol{A}_t)\|_{\infty,\infty}\end{aligned}$$

Since the columns of $\boldsymbol{\Phi}_{\mathrm{opt}}$ are orthogonal to the columns of $(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})$, one may replace $\boldsymbol{\Psi}_{\mathrm{opt}}$ by $\boldsymbol{\Phi}$ in the first term without changing its value. Then apply Lemma 5.4 to the second term (twice!). It follows that

$$\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad\leq\quad \|\boldsymbol{\Phi}^{*}\,(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\|_{\infty,\infty}\;+\;\frac{M(t)}{1-M(t)}\,\|\boldsymbol{\Phi}_{\mathrm{opt}}{}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty}\,.$$

In consequence of (5.7) and (5.8), we reach

$$\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad\leq\quad \|\boldsymbol{\Phi}^{*}\,(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\|_{\infty,\infty}\;+\;\frac{M(t)}{1-M(t)}\,\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty}\,.$$

Rearrange this relation to obtain

$$\|\boldsymbol{\Phi}^{*}\,\boldsymbol{R}_t\|_{\infty,\infty} \quad\leq\quad \frac{1-M(t)}{1-2\,M(t)}\,\|\boldsymbol{\Phi}^{*}\,(\boldsymbol{S}-\boldsymbol{A}_{\mathrm{opt}})\|_{\infty,\infty}\,.$$

If this inequality fails, then S-OMP must identify another optimal atom in iteration $(t+1)$. □

We are now prepared to complete the proof of the theorem.

Proof of Theorem 5.3. Observe that $\Lambda_0 = \emptyset \subset \Lambda_{\text{opt}}$, and make the inductive hypothesis that $\Lambda_t \subset \Lambda_{\text{opt}}$. By assumption, the maximum correlation between an atom and the residual satisfies

$$\|\boldsymbol{\Phi}^* (\boldsymbol{S} - \boldsymbol{A}_{\text{opt}})\|_{\infty,\infty} \quad \leq \quad \tau.$$

Therefore, Lemma 5.5 ensures that the greedy selection will identify an optimal atom in iteration $(t+1)$ provided that

$$\|\boldsymbol{\Phi}^* \boldsymbol{R}_t\|_{\infty,\infty} \quad > \quad \frac{1 - M(t)}{1 - 2\,M(t)}\,\tau. \tag{5.10}$$

If this inequality fails, we halt the algorithm. By induction, the algorithm identifies an optimal atom in each iteration.

Next, we will argue that the algorithm cannot halt unless all the atoms that have been chosen are associated with small coefficients. To that end, we assume that the algorithm has halted, i.e.,

$$\|\boldsymbol{\Phi}^* \boldsymbol{R}_t\|_{\infty,\infty} \quad \leq \quad \frac{1 - M(t)}{1 - 2\,M(t)}\,\tau, \tag{5.11}$$

and we will derive an upper bound on the coefficients associated with the remaining atoms. It is always true that $\|\boldsymbol{\Phi}_{\text{opt}}{}^* \boldsymbol{R}_t\|_{\infty,\infty} \leq \|\boldsymbol{\Phi}^* \boldsymbol{R}_t\|_{\infty,\infty}$ so

$$\|\boldsymbol{\Phi}_{\text{opt}}{}^* \boldsymbol{R}_t\|_{\infty,\infty} \quad \leq \quad \frac{1 - M(t)}{1 - 2\,M(t)}\,\tau. \tag{5.12}$$

Bound the left-hand side below using Lemma 5.4, and cancel the terms $(1 - M(t))$ from each side to obtain

$$\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \quad \leq \quad \frac{\tau}{1 - 2\,M(t)}$$

where $\boldsymbol{C}_{\sim t}$ contains the $(T - t)$ rows of $\boldsymbol{C}_{\text{opt}}$ listed in $\Lambda_{\text{opt}} \setminus \Lambda_t$. In consequence,

$$\|\boldsymbol{C}_{\sim t}\|_{\infty,\infty} \quad = \quad \max_{\lambda \in \Lambda_{\text{opt}} \setminus \Lambda_t} \sum\nolimits_{k=1}^{K} |\boldsymbol{C}_{\text{opt}}(\lambda, k)|\,.$$

Therefore, each row of $\boldsymbol{C}_{\sim t}$ has ℓ_1 norm less than $\tau/(1 - 2\,M(t))$ when the algorithm halts.

Finally, we develop the absolute error bound. If the algorithm has halted, then (5.12) is in force. Recall from the proof of Lemma 5.4 that

$$\boldsymbol{\Phi}_{\text{opt}}{}^* \boldsymbol{R}_t \quad = \quad \boldsymbol{\Phi}_{\text{opt}}{}^* (\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t) \quad = \quad \boldsymbol{\Phi}_{\sim t}{}^* (\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t).$$

Proposition 2.1 and the succeeding remarks yield the lower norm bound

$$\left\|(\boldsymbol{\Phi}_{\sim t})^\dagger\right\|_{2,1}^{-1} \|\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t\|_{\text{F}} \quad \leq \quad \|\boldsymbol{\Phi}_{\sim t}{}^* (\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t)\|_{\infty,\infty}\,.$$

Combining these results, we reach

$$\|\boldsymbol{A}_{\text{opt}} - \boldsymbol{A}_t\|_{\text{F}} \quad \leq \quad \frac{1 - M(t)}{1 - 2\,M(t)} \left\|(\boldsymbol{\Phi}_{\sim t})^\dagger\right\|_{2,1} \tau.$$

Square this inequality and add $\|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{\text{F}}^2$ to both sides. Then apply the Pythagorean Theorem to the left-hand side to obtain

$$\|\boldsymbol{S} - \boldsymbol{A}_t\|_{\text{F}}^2 \quad \leq \quad \|\boldsymbol{S} - \boldsymbol{A}_{\text{opt}}\|_{\text{F}}^2 \;+\; \left[\frac{1 - M(t)}{1 - 2\,M(t)} \left\|(\boldsymbol{\Phi}_{\sim t})^\dagger\right\|_{2,1} \tau\right]^2.$$

Proposition 3.9 of [Tro04d] shows that

$$\left\|(\boldsymbol{\Phi}_{\sim t})^\dagger\right\|_{2,1}^2 \quad \leq \quad \frac{T - t}{1 - \mu_1(T - t)}.$$

Combine the last two bounds to complete the argument. □

6. Comparison with Previous Work

In this section, we will try to address the relationship between this paper and other work on simultaneous sparse approximation. The second part of the paper contains a separate section that compares our theory about convex relaxation with results from the literature.

First, let us discuss the numerical investigation of simultaneous sparse approximation. In this work, we have only performed experiments using Simultaneous Orthogonal Matching Pursuit. Other papers have performed experiments with ℓ_1 minimization [CH04b] and with M-FOCUSS [CREKD04]. Aside from the actual algorithm used, the methodology in our experiments is quite different from other numerical simulations that have appeared in the literature. First of all, we change each parameter in the problem separately, and we perform 1000 trials for each choice of parameters so that we may draw statistical inferences with some degree of confidence. Moreover, we have explored a broad range of parameters to identify distinct regimes in the algorithm's performance.

The greedy pursuit algorithm that we have proposed is also distinct from some other greedy algorithms that have been suggested for simultaneous sparse approximation. One major difference among these algorithms is the greedy selection that occurs during each iteration. Several variations on the greedy selection rule take the form

$$\max_{\omega \in \Omega} \ \| \boldsymbol{R}_{t-1}{}^* \, \boldsymbol{\varphi}_\omega \|_p^p \,. \tag{6.1}$$

Intuitively, small values of p promote the selection of atoms that contribute to many different signals at once, even if the contributions are relatively small. Larger values of p favor atoms that contribute a lot to a single signal. If each column of the signal matrix can be approximated well with the same set of atoms, then it is preferable to set $p = 1$. On the other hand, signal matrices whose columns require somewhat different collections of atoms may be approximated more successfully when p is larger. An interesting possibility that has not been explored is to select $0 \leq p < 1$, although the lack of convexity may complicate the analysis.

Different versions of the selection rule (6.1) have been considered in the literature. The S-OMP algorithm, described here and in [CH04b], uses $p = 1$. Cotter et al. [CREKD04] have proposed to set $p = 2$. Leviatan and Temlyakov [LT03a] have also proposed $p = 2$ (their WSOGA2 algorithm). In their paper, Chen and Huo have compared several of these algorithms for the problem of recovering a sparse signal matrix [CH04b, Sec. IV-B]. They prove that S-OMP can recover any signal matrix that has a representation using $(\mu^{-1} + 1)/2$ atoms or fewer. They also show that, when $p = 2$, the corresponding version of S-OMP can recover $(\mu^{-1}+1)/(1+\sqrt{K})$ atoms or fewer. These facts suggest (but do not prove) that S-OMP is more effective for recovering sparse signals than algorithms that make p larger.

Another way to alter the greedy algorithm is to change the way that new approximations are formed. Cotter et al. have considered greedy algorithms, called M-BMP and M-ORMP, that use different methods for constructing the new approximation [CREKD04].

The literature contains several different types of theoretical results for simultaneous greedy pursuit. The results of Chen and Huo mentioned above demonstrate that S-OMP and a variant can recover a signal matrix that has a sufficiently sparse representation [CH04b]. At present, their work does not address approximation of sparse signal matrices that have been corrupted with noise. The conference paper [TGS04] contains a result for the noisy case, which we quoted as Theorem 5.1. These results all depend on the coherence parameter.

Results of a very different type have been established by Temlyakov and his colleagues. See for example [LT03c, LT03a, LT03b, Tem04]. Temlyakov et al. have developed many different types of greedy algorithms for simultaneous sparse approximation in Hilbert spaces and Banach spaces. These methods include the (Orthogonal) Vector Weak Greedy Algorithm, the Weak Simultaneous (Orthogonal) Greedy Algorithms 1 and 2, the Vector Weak Relaxed Greedy Algorithm, the Vector

Weak Chebyshev Greedy Algorithm, and the Chebyshev Vector Weak Greedy Algorithm. (Each one is abbreviated by its acronym.) These algorithms use different types of selection rules and different techniques for forming the approximation at each iteration. In particular, we remark that WSOGA2 is essentially S-OMP with a (weak) selection rule of the form (6.1) for $p = 2$.

A typical result for this family of algorithms shows that the norm of the residual decays at a certain rate as the number of atoms increases. For example, the residual matrix $\boldsymbol{R}_t$ generated by OVWGA, WSOGA1, or WSOGA2 (with weakness parameters identically one) satisfies

$$\|\boldsymbol{R}_t\|_{\mathrm{F}}^2 \quad \leq \quad K^2 \left[1 \; + \; \frac{t}{K}\right]^{-1} \qquad \text{for } t = 1, 2, 3, \ldots.$$

Note that this result holds for an arbitrary dictionary; it does not involve the coherence parameter. On the other hand, it does not compare the residual against the optimal t-term residual. Nor does it guarantee that any specific choice of atoms will participate in the approximation. The scope of this paper does not allow us to discuss the theory of Temlyakov et al. in more detail. It is unfortunate that our brief remarks cannot do justice to this body of work.

There are still many open questions in the study of greedy pursuit algorithms for simultaneous sparse approximation. For example, our result for sparsity-constrained approximation can certainly be improved. It would also be valuable to study the performance of the algorithm when the approximation error is measured with other norms or divergences. Moreover, an average-case analysis of greedy pursuit is still lacking. Of course, it is also essential to investigate the performance of greedy algorithms in the context of particular applications.

References

[CH04a] J. Chen and X. Huo. Sparse representations for multiple measurement vectors (MMV) in an overcomplete dictionary. Submitted to ICASSP 2005, 2004.

[CH04b] J. Chen and X. Huo. Theoretical results about finding the sparsest representations of multiple measurement vectors (MMV) in an overcomplete dictionary using ℓ_1 minimization and greedy algorithms. Manuscript, 2004.

[CR02] S. Cotter and B. D. Rao. Sparse channel estimation via matching pursuit with application to equialization. *IEEE Trans. Communications*, 50(3):374–377, Mar. 2002.

[CREKD04] S. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions of linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 2004. Accepted.

[DE03] D. L. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100:2197–2202, March 2003.

[DET04] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. Working draft, February 2004.

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

[GGR95] I. F. Gorodnitsky, J. S. George, and B. D. Rao. Neuromagnetic source imaging with focuss: a recursive weighted minimum norm algorithm. *J. Electroencephalography and Clinical Neurophysiology*, 95(4):231–251, Oct. 1995.

[GMS03] 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.

[Gri02] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind source separation of more than two sources from a stereo mixture. In *Proc. of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

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

[LT03a] D. Leviatan and V. N. Temlyakov. Simultaneous approximation by greedy algorithms. IMI Report 2003:02, Univ. of South Carolina at Columbia, 2003.

[LT03b] D. Leviatan and V. N. Temlyakov. Simultaneous greedy approximation in Banach spaces. IMI Report 03:26, Univ. of South Carolina at Columbia, 2003.

[LT03c] A. Lutoborski and V. N. Temlyakov. Vector greedy algorithms. *J. Complexity*, 19:458–473, 2003.

[Mal03] D. Malioutov. A sparse signal reconstruction perspective for source localization with sensor arrays. Ms.C. thesis, MIT, July 2003.

[MÇW03] D. Malioutov, M. Çetin, and A. Willsky. Source localization by enforcing sparsity through a Laplacian prior: an SVD-based approach. In *IEEE Statistical Signal Processing Workshop*, pages 553–556, Oct. 2003.

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

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

[Roh00] J. Rohn. Computing the norm $\|A\|_{\infty,1}$ is NP-hard. *Linear and Multilinear Algebra*, 47:195–204, 2000.

[Tem04] V. N. Temlyakov. A remark on simultaneous greedy approximation. *East J. Approx.*, 100:17–25, 2004.

[TG05] J. A. Tropp and A. C. Gilbert. Signal recovery from partial information via Orthogonal Matching Pursuit. In preparation, 2005.

[TGS04] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Simultaneous sparse approximation by greedy pursuit. Submitted to ICASSP 2005, October 2004.

[Tro04a] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. Submitted for publication, Nov. 2004.

[Tro04b] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10), Oct. 2004.

[Tro04c] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. Submitted for publication, June 2004.

[Tro04d] J. A. Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. ICES Report 04-04, The University of Texas at Austin, 2004.

[Tro04e] J. A. Tropp. *Topics in Sparse Approximation*. Ph.D. dissertation, Computational and Applied Mathematics, The University of Texas at Austin, August 2004.