# SIMULTANEOUS SPARSE APPROXIMATION VIA GREEDY PURSUIT

*J. A. Tropp, A. C. Gilbert*

Department of Mathematics
The University of Michigan
Ann Arbor, MI 48109

*M. J. Strauss*

Departments of Mathematics and EECS
The University of Michigan
Ann Arbor, MI 48109

## ABSTRACT

A simple sparse approximation problem requests an approximation of a given input signal as a linear combination of $T$ elementary signals drawn from a large, linearly dependent collection. An important generalization is simultaneous sparse approximation. Now one must approximate several input signals at once using different linear combinations of the same $T$ elementary signals. This formulation appears, for example, when analyzing multiple observations of a sparse signal that have been contaminated with noise.

A new approach to this problem is presented here: a greedy pursuit algorithm called Simultaneous Orthogonal Matching Pursuit. The paper proves that the algorithm calculates simultaneous approximations whose error is within a constant factor of the optimal simultaneous approximation error. This result requires that the collection of elementary signals be weakly correlated, a property that is also known as incoherence. Numerical experiments demonstrate that the algorithm often succeeds, even when the inputs do not meet the hypotheses of the proof.

## 1. INTRODUCTION

We work in the complex inner-product space $\mathbb{C}^d$, which is called the *signal space*. We write $\langle \cdot, \cdot \rangle$ for the usual inner product and $\|\cdot\|_2$ for the associated norm. The symbol $\|\cdot\|_p$ indicates the $\ell_p$ vector norm, while $\|\cdot\|_{p,q}$ is the norm on linear operators mapping $\ell_p$ to $\ell_q$. We use $^*$ for the complex conjugate transpose of vectors and matrices.

A *dictionary* $\mathscr{D}$ is a finite collection of unit-norm elementary signals, called *atoms*, that spans the signal space. Each atom is denoted $\boldsymbol{\varphi}_\omega$, where $\omega$ is drawn from an index set $\Omega$. The number of atoms $N$ is typically much larger than the dimension $d$ of the signal space. We also define the $d \times N$ dictionary matrix $\boldsymbol{\Phi}$ whose columns are atoms.

Suppose that $\mathbf{S}$ is a $d \times K$ matrix whose columns are input signals. We wish to approximate all $K$ input signals using different linear combinations of the same $T$ atoms. Typically, $T$ is much smaller than the dimension of the signal space, so the approximation is *sparse*. More precisely, the *simultaneous sparse approximation problem* (SSA) elicits an $N \times K$ coefficient matrix $\mathbf{C}$ that solves the mathematical program

$$\min_{\mathbf{C}} \quad \|\mathbf{S} - \boldsymbol{\Phi}\,\mathbf{C}\|_{\mathrm{F}}^2 \quad \text{subject to}$$
$$\text{the matrix } \mathbf{C} \text{ has at most } T \text{ nonzero rows.} \qquad \text{(SSA)}$$

The squared Frobenius matrix norm $\|\cdot\|_{\mathrm{F}}^2$ returns the sum of the squares of the entries in a matrix.

The (SSA) problem arises if we are given multiple observations of a sparse input signal that are contaminated with noise. For example, the $k$-th input signal might have the form

$$\boldsymbol{s}_k = \boldsymbol{x} + \boldsymbol{\nu}_k$$

where $\boldsymbol{\nu}_k$ is a realization of some random process and where $\boldsymbol{x}$ can be expressed using a linear combination of $T$ atoms. The goal is to identify the atoms that comprise $\boldsymbol{x}$.

To solve (SSA), we propose a greedy pursuit method, *Simultaneous Orthogonal Matching Pursuit* (S-OMP). For general dictionaries, (SSA) cannot be solved without checking every combination of $T$ nonzero rows. This follows from results in [1]. Nevertheless, we have been able to prove that S-OMP correctly solves the simultaneous sparse approximation problem, provided that the atoms are weakly correlated. To quantify this property, we define the *coherence parameter* of the dictionary,

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

When the coherence parameter is small, each pair of atoms is nearly orthogonal.

This paper provides the first proof that any algorithm can obtain provably good solutions to (SSA). A simple version of our result follows[1]. Suppose that the set $\Lambda_{\mathrm{opt}}$ indexes the $T$ atoms that appear in some solution to (SSA). Then we may define the $d \times K$ matrix $\mathbf{A}_{\mathrm{opt}}$ whose $k$-th column is the best approximation of the $k$-th input signal using the $T$ atoms listed in $\Lambda_{\mathrm{opt}}$.

[1]Note that the present result does not yield an optimal bound for the constant, even when the dictionary is orthonormal. A more subtle analysis is necessary to achieve the improvement.

**Theorem 1** *Assume that $\mu T < \frac{1}{2}$. After $T$ steps, suppose that Simultaneous Orthogonal Matching Pursuit returns the approximation $A_T$. Then the output error is bounded by*

$$\|S - A_T\|_{\mathrm{F}} \leq \sqrt{1 + C(\mathscr{D}, K, T)}\ \|S - A_{\mathrm{opt}}\|_{\mathrm{F}},$$

*and the constant is no worse than*

$$C(\mathscr{D}, K, T) \leq \frac{K\,T\,(1 - \mu\,T)}{(1 - 2\,\mu\,T)^2}.$$

*In particular, if the optimal approximation error is zero, S-OMP returns an approximation that achieves zero error.*

The algorithm S-OMP performs much better in practice than our theory predicts. Not only can we recover the input signals when $T$ is large, the error does not grow as quickly as our bounds would suggest. We have discovered that moderate noise levels create surprising difficulties for our algorithm. To our knowledge, these are the first numerical experiments performed on (SSA).

The rest of the paper expands on the claims of the introduction. Section 2 provides a rigorous statement of our greedy pursuit algorithm. In Section 3, we sketch the proof that the algorithm constructs approximate solutions to (SSA), and we discusses several other factors that affect its performance. The paper concludes with Section 4, which summarizes our numerical experiments.

## 2. THE ALGORITHM

Let us continue with a formal description of the algorithm.

**Algorithm 2 (S-OMP)**
INPUT:

- *A $d \times K$ matrix $S$ of input signals*
- *The number $T$ of atoms in the approximation*

OUTPUT:

- *A set $\Lambda_T$ containing $T$ indices*
- *A $d \times K$ approximation matrix $A_T$*
- *A $d \times K$ residual matrix $R_T$*

PROCEDURE:

1. *Initialize the residual matrix $R_0 = S$, the index set $\Lambda_0 = \emptyset$, and the iteration counter $t = 1$.*
2. *Find an index $\lambda_t$ that solves the easy optimization problem*

$$\max_{\omega \in \Omega} \quad \sum_{k=1}^{K} |\langle R_{t-1}\,\mathbf{e}_k, \varphi_\omega \rangle|\,.$$

   *We use $\mathbf{e}_k$ to denote the $k$-th canonical basis vector.*
3. *Set $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$.*
4. *Determine the orthogonal projector $P_t$ onto the span of the atoms indexed in $\Lambda_t$.*
5. *Calculate the new approximation and residual:*

$$A_t = P_t\,S$$
$$R_t = S - A_t.$$

6. *Increment $t$, and return to Step 2 if $t \leq T$.*

This procedure reduces to standard Orthogonal Matching Pursuit [1] when $K = 1$.

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 contributes the most energy to as many of the input signals as possible. Note that this absolute sum can also be written as $\|{R_t}^* \varphi_\omega\|_1$. In contrast, Leviatan and Temlyakov [2] have studied a greedy algorithm for (SSA) that picks an atom by maximizing $\|{R_t}^* \varphi_\omega\|_\infty$.

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 [3, Ch. 5] for extensive details. It is important to note that each column of the residual $R_t$ is orthogonal to the atoms indexed in $\Lambda_t$. Therefore, no atom is ever chosen twice.

## 3. PROOF OF CORRECTNESS

We will develop a condition which guarantees that S-OMP selects an optimal atom at iteration $t$. From this condition, it is easy to prove that Simultaneous Orthogonal Matching Pursuit can compute approximate solutions to (SSA).

**Theorem 3** *Assume that $\mu T < \frac{1}{2}$, and fix a signal matrix $S$. At iteration $t$, suppose that S-OMP has chosen $t$ optimal atoms, and let $A_t$ be the current approximation of the signal matrix. At iteration $(t+1)$, greedy selection will identify another optimal atom provided that*

$$\|S - A_t\|_{\mathrm{F}}^2 > \|S - A_{\mathrm{opt}}\|_{\mathrm{F}}^2 + \frac{T\,(1 - \mu\,T)}{(1 - 2\,\mu\,T)^2}\ \|\Phi^*(S - A_{\mathrm{opt}})\|_{\infty,\infty}^2 \quad (1)$$

*where $A_{\mathrm{opt}}$ denotes an optimal approximation of the signal matrix using $T$ atoms.*

In words, the algorithm selects another optimal atom whenever the current approximation is somewhat worse than an optimal approximation. We interpret $\|\Phi^*(S - A_{\mathrm{opt}})\|_{\infty,\infty}$ as the maximum total correlation between a fixed atom and the residuals left over from the optimal approximation.

**Proof.** Suppose that some solution of (SSA) involves the $T$ atoms indexed in $\Lambda_{\text{opt}}$. Define the $d \times T$ matrix $\Phi_{\text{opt}}$ whose columns are the atoms listed by $\Lambda_{\text{opt}}$. Let the $d \times (N-T)$ matrix $\Psi_{\text{opt}}$ contain the remaining atoms. Recall the definition $R_t = S - A_t$.

First, observe that each row of the matrix $(\Phi_{\text{opt}}{}^* R_t)$ lists the inner products between a fixed atom in $\Lambda_{\text{opt}}$ and the columns of $R_t$. The rows of the matrix $(\Psi_{\text{opt}}{}^* R_t)$ have an analogous interpretation. The $(\infty, \infty)$ matrix norm returns the maximum absolute row sum of its argument, and so the algorithm chooses another optimal atom if and only if the ratio

$$\rho \stackrel{\text{def}}{=} \frac{\|\Psi_{\text{opt}}{}^* R_t\|_{\infty,\infty}}{\|\Phi_{\text{opt}}{}^* R_t\|_{\infty,\infty}} \tag{2}$$

is strictly less than one. We must ensure that $\rho < 1$.

Rewrite $R_t = (S - A_{\text{opt}}) + (A_{\text{opt}} - A_t)$. Substitute this expression into (2). The term $\Phi_{\text{opt}}{}^*(S - A_{\text{opt}})$ vanishes from the denominator because of orthogonality. Apply the triangle inequality to the numerator to see that $\rho$ is no greater than

$$\frac{\|\Psi_{\text{opt}}{}^*(A_{\text{opt}} - A_t)\|_{\infty,\infty}}{\|\Phi_{\text{opt}}{}^*(A_{\text{opt}} - A_t)\|_{\infty,\infty}} + \frac{\|\Psi_{\text{opt}}{}^*(S - A_{\text{opt}})\|_{\infty,\infty}}{\|\Phi_{\text{opt}}{}^*(A_{\text{opt}} - A_t)\|_{\infty,\infty}}. \tag{3}$$

Now we bound the first fraction in (3). Let $\Phi_{\text{opt}}{}^+$ denote the generalized inverse of $\Phi_{\text{opt}}$. Using an argument analogous to that in [4, Thm. 3.1], we discover that

$$\frac{\|\Psi_{\text{opt}}{}^*(A_{\text{opt}} - A_t)\|_{\infty,\infty}}{\|\Phi_{\text{opt}}{}^*(A_{\text{opt}} - A_t)\|_{\infty,\infty}} \le \left\|\Phi_{\text{opt}}{}^+ \Psi_{\text{opt}}\right\|_{1,1}. \tag{4}$$

It can be shown that the second fraction in (3) is no greater than

$$\frac{\|\Psi_{\text{opt}}{}^*(S - A_{\text{opt}})\|_{\infty,\infty}}{\left\|\Phi_{\text{opt}}{}^+\right\|_{2,1}^{-1} \|A_{\text{opt}} - A_t\|_{\text{F}}}. \tag{5}$$

In the numerator, we replace $\Psi_{\text{opt}}$ with $\Phi = [\Phi_{\text{opt}}\ \Psi_{\text{opt}}]$, using the fact that the columns of $(S - A_{\text{opt}})$ are orthogonal to the columns of $\Phi_{\text{opt}}$.

We substitute the bounds (4) and (5) into (3) and perform some algebraic manipulations to find a condition that $\rho < 1$. To complete the argument, we introduce the coherence estimates developed in [5, Sec. 3].

□

**Corollary 4** *Assume that $\mu\, T < \frac{1}{2}$. Given any input matrix $S$, Simultaneous Orthogonal Matching Pursuit will always construct a $T$-term approximation $A_T$ that satisfies the error bound*

$$\|S - A_T\|_{\text{F}}^2 \le \|S - A_{\text{opt}}\|_{\text{F}}^2 + \frac{T\,(1 - \mu\, T)}{(1 - 2\,\mu\, T)^2}\, \|\Phi^*(S - A_{\text{opt}})\|_{\infty,\infty}^2$$

*where $A_{\text{opt}}$ is an optimal $T$-term approximation of $S$.*

From here, one reaches Theorem 1 by noting that

$$\|\Phi^*(S - A_{\text{opt}})\|_{\infty,\infty}^2 \le K\, \|S - A_{\text{opt}}\|_{\text{F}}^2$$

because the columns of $\Phi$ all have unit norm.

## 4. NUMERICAL EXPERIMENTS

To evaluate the performance of our algorithm, we have tested it with three types of input signal. Each type is a variant on 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 (possibly the same for each $k$ as given in the introductory example). In each dimension $d$, our dictionary is the collection of $d$ complex exponentials and $d$ impulses. That is, $\boldsymbol{\varphi}_\omega[t] = \mathrm{e}^{2\pi \mathrm{i}\, t\omega/d}$ for $\omega = 1, \dots, d$ and $\boldsymbol{\varphi}_\lambda[t] = \boldsymbol{\delta}_\lambda[t]$ for $\lambda = 1, \dots, d$. Note that this dictionary has coherence $\mu = 1/\sqrt{d}$.

We begin with input signals of the form

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

For each signal $\boldsymbol{s}_k$, we select $T$ atoms independently and uniformly from the dictionary. The coefficients $\alpha_{jk}$ are chosen from iid normal distributions. Our algorithm is, therefore, searching for the best $T$ atoms with which to represent $K$ signals, each of which is a linear combination of $T$ atoms. We have observed that our algorithm always recovers $T$ atoms from the collection of approximately $KT$ distinct atoms that participate in the $K$ input signals. All of the error in the residual is due to the fact that the input signals involve more atoms that we are allowed to use.

The second type of input signal has the form

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

For all $K$ signals, we use the same core of $T$ atoms, but the coefficients $\alpha_{jk}$ are chosen from iid normal distributions. For these experiments, we fixed the dimension of the signal space at $d = 128$ and the number of signals at $K = 2$. 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. We can see from this figure that our theoretical bounds are far too pessimistic. Even for $T = 90$ (out of a possible 128 atoms), we typically recover most of the core set.

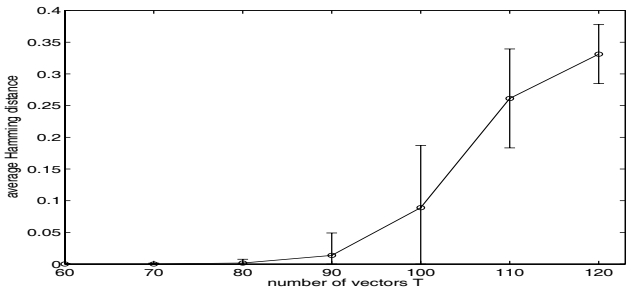


**Fig. 1**. The average Hamming distance between the core set of the vectors and the recovered set as a function of the number of vectors $T$ in the core set. (Input type two)

The third input type has the form

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

That is, we choose $T$ atoms at random and form a linear combination with random coefficients $\alpha_j \in \{\pm 1\}$. Then we construct $K$ input signals by corrupting the original signal with iid 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 2 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 the four SNR values. Naturally, the Hamming distance increases as 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., 13dB), we cannot reliably recover three atoms in a 256-dimensional signal space. With the parameter settings we have chosen, our theoretical results predict that

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

To see if this bound accurately predicts the dependence on $K$ and $T$, we plot in Figure 3 the total relative error as a function of the number of signals $K$. For each $T$, we use a 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.

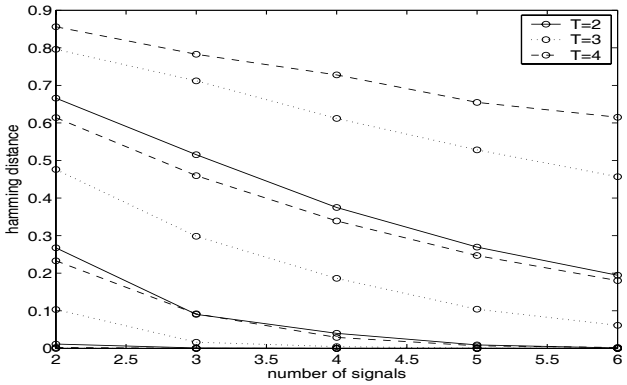


**Fig. 2**. The average Hamming distance between the core set of vectors and the recovered set as a function of the number of signals and the SNR. (Input type three)

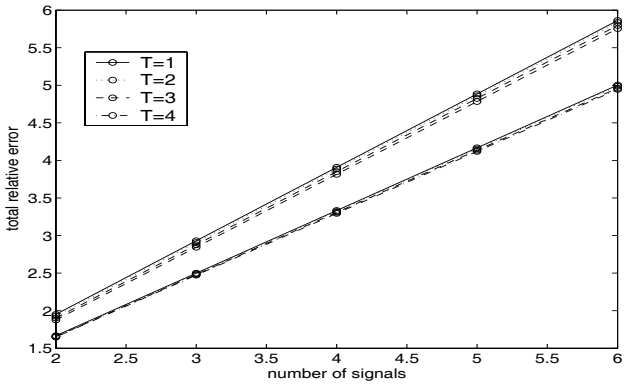


**Fig. 3**. The total relative error as a function of the number of signals and the number of core vectors for two values of SNR. (Input type three)

## 5. REFERENCES

[1] G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *J. Constr. Approx.*, vol. 13, pp. 57–98, 1997.

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

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

[4] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, October 2004, To appear.

[5] 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.