Greed is Good: Algorithmic Results for Sparse Approximation

Joel A. Tropp, *Student Member, IEEE*

***Abstract*—This article presents new results on using a greedy algorithm, orthogonal matching pursuit (OMP), to solve the sparse approximation problem over redundant dictionaries. It provides a sufficient condition under which both OMP and Donoho's basis pursuit (BP) paradigm can recover the optimal representation of an exactly sparse signal. It leverages this theory to show that both OMP and BP succeed for every sparse input signal from a wide class of dictionaries. These quasi-incoherent dictionaries offer a natural generalization of incoherent dictionaries, and the cumulative coherence function is introduced to quantify the level of incoherence. This analysis unifies all the recent results on BP and extends them to OMP.**

Furthermore, the paper develops a sufficient condition under which OMP can identify atoms from an optimal approximation of a nonsparse signal. From there, it argues that OMP is an approximation algorithm for the sparse problem over a quasi-incoherent dictionary. That is, for *every* input signal, OMP calculates a sparse approximant whose error is only a small factor worse than the minimal error that can be attained with the same number of terms.

***Index Terms*—Algorithms, approximation methods, basis pursuit (BP), iterative methods, linear programming, orthogonal matching pursuit (OMP).**

I. Introduction

SOME signals cannot be represented efficiently in an orthonormal basis. For example, neither impulses nor sinusoids adequately express the behavior of an intermixture of impulses and sinusoids. In this case, two types of structures appear in the signal, but they look so radically different that neither one can effectively mimic the other. Although orthonormal bases have a distinguished service record in approximation theory, examples like this have led researchers to enlist more complicated techniques.

The most basic instrument of approximation projects each signal onto a fixed m-dimensional linear subspace. A familiar example is interpolation by means of fixed-knot polynomial splines. For some types of signals, this elementary procedure works quite well. Later, various nonlinear methods were developed. One fundamental technique is to project a signal onto the *best* linear subspace spanned by m elements of a fixed orthonormal basis. This type of approximation is quite easy to perform due to the rigid structure of an orthonormal system. In comparison with the linear method, it may yield a significant improvement in the approximation error [1], [2]. But, as noted, some signals just do not fit into an orthonormal basis. To deal with this problem, researchers have spent the last 15 years developing redundant systems, called dictionaries, for analyzing and representing complicated signals. A Gabor dictionary, for example, consists of complex exponentials smoothly windowed to short time intervals. It is used for joint time–frequency analysis [3].

The problem of approximating a signal with the best linear combination of m elements from a redundant dictionary is called *sparse approximation* or *highly nonlinear approximation*. The core algorithmic question is the following.

> For a given class of dictionaries, how does one design a fast algorithm that provably calculates a nearly optimal sparse representation of an arbitrary input signal?

Unfortunately, it is quite difficult to answer. At present, there are two major approaches, orthogonal matching pursuit (OMP) and basis pursuit (BP). OMP is an iterative greedy algorithm that selects at each step the dictionary element best correlated with the residual part of the signal. Then it produces a new approximant by projecting the signal onto the disctionary elements that have already been selected. This technique extends the trivial greedy algorithm that succeeds for an orthonormal system. BP is a more sophisticated approach that replaces the original sparse approximation problem by a linear programming problem. Empirical evidence suggests that BP is more powerful than OMP [4]. The major advantage of OMP is that it admits simple, fast implementations [5], [6].

A. Major Results

We begin with a résumé of the major results. Fix a (redundant) dictionary of elementary signals, which are called atoms. A representation of a signal is a linear combination of atoms that equals the signal. Every signal has an infinite number of distinct representations over a redundant dictionary. The EXACT-SPARSE problem is to identify the representation of the input signal that uses the least number of atoms, i.e., the sparsest one.

Our first result is a sufficient condition for OMP and BP to solve EXACT-SPARSE. To state the theorem, we need a little notation. Given an input signal, form a matrix Φ_{opt} whose columns are the atoms that make up the optimal representation of the signal. The pseudoinverse of this matrix is defined as $\Phi_{\text{opt}}^{+} = (\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})^{-1}\Phi_{\text{opt}}$. The notation $\|\cdot\|_1$ indicates the ℓ_1 vector norm, which returns the absolute sum of a vector's components.

Manuscript received March 21, 2003; revised June 6, 2004. This work was supported by a National Science Foundation Graduate Fellowship.

The author was with the Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712 USA. He is now with the Mathematics Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jtropp@umich.edu).

Communicated by G. Battail, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2004.834793

Theorem A: Suppose that

$$\max_{\boldsymbol{\psi}} \left\| \Phi_{\text{opt}}^{+} \boldsymbol{\psi} \right\|_1 < 1 \qquad \text{(ERC)}$$

where the maximization occurs over atoms that do not participate in the optimal representation of the signal. It follows that the sparsest representation of the signal is unique. Moreover, both OMP and BP identify the optimal atoms and their coefficients.

This result encapsulates Theorem 3.1, Theorem 3.3, and discussion from Section III-E. Theorem A is essentially the best possible for OMP (Theorem 3.10), and it is also the best possible for BP in certain cases (Section III-D). It is remarkable that (ERC) is a natural sufficient condition for such disparate techniques to resolve sparse signals. This fact suggests that EXACT-SPARSE has tremendous structure.

Theorem A would not be very useful without a technique for checking when the condition (ERC) holds. To that end, we define the coherence parameter μ, which equals the maximum absolute inner product between two distinct atoms. This quantity reflects how much atoms look alike. A generalization of the coherence parameter is the cumulative coherence function $\mu_1(m)$ which equals the maximum absolute sum of inner products between a fixed atom and m other atoms. If the cumulative coherence function grows slowly, we say that the dictionary is quasi-incoherent.

Theorem B: The condition (ERC) holds for every signal with an m-term representation provided that

$$m < \frac{1}{2}(\mu^{-1} + 1)$$

or, more generally, whenever

$$\mu_1(m-1) + \mu_1(m) < 1.$$

Suppose that the dictionary consists of J concatenated orthonormal bases. The condition (ERC) is in force if

$$m < \left[\sqrt{2} - 1 + \frac{1}{2(J-1)} \right] \mu^{-1}.$$

Theorem B is a restatement of Theorem 3.5, Corollary 3.6, and Corollary 3.9. Note that Theorems A and B unify all of the recent results for BP [7]–[9] and extend them to OMP as well.

Our second problem, SPARSE, requests the best approximation of a general signal using a linear combination of m atoms, where the approximation error is measured with the Euclidean norm $\|\cdot\|_2$. Although EXACT-SPARSE and SPARSE are related, the latter is much harder to solve. Nevertheless, OMP is a provably good approximation algorithm for the sparse problem over a quasi-incoherent dictionary.

Theorem C: Suppose that $\mu_1(m) \leq \frac{1}{3}$. For every input signal $\boldsymbol{s}$, OMP will calculate an m-term approximant $\boldsymbol{a}_m$ that satisfies

$$\|\boldsymbol{s} - \boldsymbol{a}_m\|_2 \leq \sqrt{1+6m}\, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2$$

where $\boldsymbol{a}_{\text{opt}}$ is an optimal m-term approximant of the input signal.

Theorem C is Corollary 4.4 of the sequel. It extends the work of Gilbert, Muthukrishnan, and Strauss [6]. Significantly stronger results for OMP have recently been announced in [10], [11].

II. BACKGROUND

A. Important Definitions

The standard sparse approximation problem is set in the finite-dimensional[1] inner-product space $\mathbb{C}^d$, which is called the *signal space*. We use angle brackets to denote the usual Hermitian inner product: $\langle \boldsymbol{s}, \boldsymbol{x} \rangle \stackrel{\text{def}}{=} \boldsymbol{x}^* \boldsymbol{s}$, where * represents the complex-conjugate transpose. The Euclidean norm is defined via the inner product: $\|\boldsymbol{s}\|_2 \stackrel{\text{def}}{=} \sqrt{\langle \boldsymbol{s}, \boldsymbol{s} \rangle}$.

A *dictionary* for the signal space is a finite collection $\mathcal{D}$ of unit-norm vectors that spans the whole space. The members of the dictionary are called *atoms*, and they are denoted by $\boldsymbol{\varphi}_\omega$, where the parameter ω is drawn from an index set Ω. The indices may have an interpretation such as the time–frequency or time–scale localization of an atom, or they may simply be labels without an underlying meaning. Thus,

$$\mathcal{D} = \{ \boldsymbol{\varphi}_\omega : \omega \in \Omega \}.$$

The letter N will indicate the size of the dictionary. Clearly, $N = |\mathcal{D}| = |\Omega|$, where $|\cdot|$ returns the cardinality of a set.

A *representation* of a signal is a linear combination of atoms that equals the signal. Without loss of generality, we assume that all the coefficients in a representation are nonzero. Naturally, an m-term representation is a representation that involves m atoms. Identifying the atoms and coefficients that appear in the sparsest representation of a signal will be referred to as *recovering* the sparsest representation or, equivalently, recovering the signal.

B. Sparse Approximation Problems

The fundamental problem is to approximate a given signal $\boldsymbol{s}$ using a linear combination of m atoms. Since m is taken to be much smaller than the dimension d of the signal space, the approximant is *sparse*. Specifically, we seek a solution to the minimization problem

$$\min_{|\Lambda| = m} \quad \min_{\{b_\lambda\}} \quad \left\| \boldsymbol{s} - \sum\nolimits_{\lambda \in \Lambda} b_\lambda \boldsymbol{\varphi}_\lambda \right\|_2 \qquad (1)$$

where the index set $\Lambda \subset \Omega$ and $\{b_\lambda\}$ is a collection of complex coefficients. For a fixed Λ, the inner minimization of (1) can be accomplished with the usual least-squares techniques. The real difficulty lies in the optimal selection of Λ, since the naïve strategy would involve sifting through all $\binom{N}{m}$ possibilities.

The computational problem (1) will be called $(\mathcal{D}, m)$-SPARSE. Note that it is posed for an *arbitrary* input signal with respect to a *fixed* dictionary and sparsity level. One reason for posing the problem with respect to a specific dictionary is to reduce the time complexity of the problem. If the dictionary were an input parameter, then an algorithm would have to process the entire dictionary as one of its computational duties. It is better

[1]We work in a finite-dimensional space because infinite-dimensional vectors do not fit inside a computer. Nonetheless, the theory carries over with appropriate modifications to an infinite-dimensional setting.

to transfer this burden to a preprocessing stage because we are likely to use the same dictionary for many approximations. A second reason is that solving or even approximating the solution of (1) is NP-hard if the dictionary is unrestricted [5], [12]. Nevertheless, it is not quixotic to seek algorithms for the sparse problem over a *particular* dictionary.

We will also consider a second problem called $(\mathcal{D}, m)$-EXACT-SPARSE, where the input signal is required to have a representation using m atoms or fewer from $\mathcal{D}$. There are several motivations. Although natural signals are not perfectly sparse (Proposition 4.1), one might imagine applications in which a sparse signal is constructed and transmitted without error. EXACT-SPARSE models just this situation. Second, analysis of the simpler problem can provide lower bounds on the computational complexity of SPARSE; if the first problem is NP-hard, the second one is too. Finally, we might hope that understanding EXACT-SPARSE will lead to insights on the more general case.

C. Algorithms

In this subsection, we will describe some of the basic algorithms for sparse approximation. The methods come in two flavors. Greedy methods make a sequence locally optimal choices in an effort to determine a globally optimal solution. Convex relaxation methods replace the combinatorial sparse approximation problem with a related convex program. We begin with the greedy techniques.

1) Matching Pursuit (MP): If the dictionary is orthonormal, the sparse approximation problem admits a straightforward algorithm. It is possible to build a solution one term at a time by selecting at each step the atom that correlates most strongly with the residual signal. Matching pursuit (MP) extends this idea to other types of dictionaries.

MP begins by setting the initial residual equal to the input signal $\boldsymbol{s}$ and making a trivial initial approximation. That is,

$$\boldsymbol{r}_0 = \boldsymbol{s} \quad \text{and} \quad \boldsymbol{a}_0 = \mathbf{0}.$$

At step k, MP chooses another index λ_k by solving an easy optimization problem

$$\lambda_k \in \arg\max_{\omega \in \Omega} |\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_\omega \rangle| . \tag{2}$$

Then it calculates a new approximation and a new residual

$$\boldsymbol{a}_k = \boldsymbol{a}_{k-1} + \langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k} \rangle \boldsymbol{\varphi}_{\lambda_k}$$

and

$$\boldsymbol{r}_k = \boldsymbol{r}_{k-1} - \langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k} \rangle \boldsymbol{\varphi}_{\lambda_k}. \tag{3}$$

The residual can also be expressed as $\boldsymbol{r}_k = \boldsymbol{s} - \boldsymbol{a}_k$.

When the dictionary is an orthonormal basis, the approximant $\boldsymbol{a}_m$ is always an optimal m-term representation of the signal. For general dictionaries, Jones has shown that the norm of the residual converges to zero [13]. In fact, it converges exponentially when the signal space is finite dimensional [5].

MP was developed in the statistics community under the cognomen Projection Pursuit Regression [14]. It was introduced to the signal processing community by [15] and independently by [16]. In the approximation communitity, MP is known as the Pure Greedy Algorithm [2]. For more history, theory, and a list of references, see Temlyakov's monograph [2].

2) OMP: OMP adds a least-squares minimization to each step of MP to obtain the best approximation over the atoms that have already been chosen. This revision significantly improves the behavior of the algorithm.

OMP is initialized the same way as MP, and at each step, a new atom is selected according to the same rule as MP, via (2). But the approximants are calculated differently. Let $\Lambda_k = \{\lambda_1, \ldots, \lambda_k\}$ list the atoms that have been chosen at step k. Then the kth approximant is

$$\boldsymbol{a}_k \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{a}} \|\boldsymbol{s} - \boldsymbol{a}\|_2 \quad \text{subject to } \boldsymbol{a} \in \operatorname{span}\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_k\}. \tag{4}$$

This minimization can be performed incrementally with standard least-squares techniques. As before, the residual is calculated as $\boldsymbol{r}_k = \boldsymbol{s} - \boldsymbol{a}_k$.

Note that OMP never selects the same atom twice because the residual is orthogonal to the atoms that have already been chosen. In consequence, the residual must equal zero after d steps.

OMP was developed independently by many researchers. The earliest reference appears to be a 1989 paper of Chen, Billings, and Luo [17]. The first signal processing papers on OMP arrived around 1993 [18], [19].

3) Weak Greedy Algorithms: OMP has a cousin called weak OMP (WOMP) that makes a brief appearance in this paper. Instead of selecting the optimal atom at each step, WOMP settles for one that is nearly optimal. Specifically, it finds an index λ_k so that

$$|\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k} \rangle| \geq \alpha \max_{\omega} |\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_\omega \rangle| \tag{5}$$

where $\alpha \in (0, 1]$ is a fixed *weakness parameter*. Once the new atom is chosen, the approximation is calculated as before, via (4).

4) Basis Pursuit (BP): Convex relaxation offers another approach to sparse approximation. The fundamental idea is that the number of terms in a representation can be approximated by the absolute sum of the coefficients. This absolute sum is a convex function, and so it can be minimized in polynomial time.

BP is a convex relaxation method designed for $(\mathcal{D}, m)$-EXACT-SPARSE [4]. Given an input signal $\boldsymbol{s}$, the BP problem is

$$\min_{\{b_\omega\}} \sum\nolimits_{\omega \in \Omega} |b_\omega| \quad \text{subject to} \quad \sum\nolimits_{\omega \in \Omega} b_\omega \boldsymbol{\varphi}_\omega = \boldsymbol{s}$$

where $\{b_\omega\}$ is a collection of complex coefficients. One hopes that the nonzero coefficients in the solution of the BP problem will identify the atoms in the optimal representation of the input signal and their coefficients.

Strictly speaking, BP is not an algorithm but a principle. At least two algorithms have been proposed for solving the BP problem. The original paper advocates interior-point methods of linear programming [4]. Sardy, Bruce, and Tseng have suggested another procedure called Block Coordinate Relaxation [20]. Both techniques are computationally intensive.

D. Dictionary Analysis

To prove some of our major results, we need a way to summarize the behavior of the dictionary. The coherence parameter and the cumulative coherence function perform this duty.

1) Coherence: The most fundamental quantity associated with a dictionary is the *coherence parameter* μ. It equals the maximum absolute inner product between two distinct atoms

$$\mu \stackrel{\text{def}}{=} \max_{j \neq k} \left| \langle \boldsymbol{\varphi}_{\omega_j}, \boldsymbol{\varphi}_{\omega_k} \rangle \right|.$$

Roughly speaking, this number measures how much two atoms can look alike. Coherence is a blunt instrument since it only reflects the most extreme correlations in the dictionary. Nevertheless, it is easy to calculate, and it captures well the behavior of uniform dictionaries. Informally, we say that a dictionary is *incoherent* when we judge that μ is small.

It is obvious that every orthonormal basis has coherence $\mu = 0$. A union of two orthonormal bases has coherence $\mu \geq d^{-1/2}$. This bound is attained, for example, by the Dirac–Fourier dictionary, which consists of impulses and complex exponentials. A dictionary of concatenated orthonormal bases (ONB) is called a *multi-ONB*. For some d, it is possible to build a multi-ONB that contains d or even $(d+1)$ bases yet retains the minimal possible coherence $\mu = d^{-1/2}$ [21]. For general dictionaries, a lower bound on the coherence is

$$\mu \geq \sqrt{\frac{N-d}{d(N-1)}}.$$

If each atomic inner product meets this bound, the dictionary is called an equiangular tight frame. See [22] for more details.

The idea of using the coherence parameter to summarize a dictionary has a distinguished pedigree. Mallat and Zhang introduced it as a quantity of heuristic interest for MP [15]. The first theoretical developments appeared in Donoho and Huo's paper [23]. Stronger results for BP, phrased in terms of coherence, were provided in [7]–[9]. Gilbert, Muthukrishnan, and Strauss have recently exhibited an approximation algorithm for sparse problems over suitably incoherent dictionaries [6].

2) The Cumulative Coherence: The coherence parameter does not characterize a dictionary very well since it only reflects the most extreme correlations between atoms. When most of the inner products are tiny, the coherence can be downright misleading. A wavelet packet dictionary exhibits this type of behavior. To remedy this shortcoming, we introduce the cumulative coherence function, which measures the maximum total coherence between a fixed atom and a collection of other atoms. In a sense, the cumulative coherence indicates how much the atoms are "speaking the same language." It is much simpler to distinguish Russian from English than it is to distinguish Russian from Ukrainian. Likewise, if the vectors in the dictionary are foreign to each other, they are much easier to tell apart. The cumulative coherence function will arise naturally in the analysis. Although it is more difficult to compute than the coherence, it is a sharper scalpel. Donoho and Elad have defined a similar notion of generalized incoherence, but they did not develop it sufficiently for present purposes [8].

For a positive integer m, the cumulative coherence function is defined as

$$\mu_1(m) \stackrel{\text{def}}{=} \max_{|\Lambda|=m} \max_{\boldsymbol{\psi}} \sum_{\Lambda} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle| \tag{6}$$

where the vector $\boldsymbol{\psi}$ ranges over the atoms indexed by $\Omega \setminus \Lambda$. We place the convention that $\mu_1(0) = 0$. The subscript in the notation serves to distinguish the cumulative coherence function from the coherence and to remind us that it is an absolute sum. When the cumulative coherence of a dictionary grows slowly, we say informally that the dictionary is *quasi-incoherent*.

Inspection of the definition (6) shows that $\mu_1(1) = \mu$ and that μ_1 is a nondecreasing function of m. The next proposition provides more evidence that the cumulative coherence generalizes the coherence parameter.

Proposition 2.1: If a dictionary has coherence μ, then $\mu_1(m) \leq m\mu$ for every natural number m.

Proof: Calculate that

$$\begin{aligned} \mu_1(m) &= \max_{|\Lambda|=m} \max_{\boldsymbol{\psi}} \sum_{\Lambda} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle| \\ &\leq \max_{|\Lambda|=m} \sum_{\Lambda} \mu \\ &= m\mu. \end{aligned}$$

□

3) An Example: For a realistic dictionary where the atoms have analytic definitions, the cumulative coherence function is not too difficult to compute. As a simple example, let us study a dictionary of decaying atoms. To streamline the calculations, we work in the infinite-dimensional Hilbert space ℓ_2 of square-summable complex-valued sequences.

Fix a parameter $\beta < 1$. For each index $k \geq 0$, define an atom by

$$\boldsymbol{\varphi}_k(t) = \begin{cases} 0, & 0 \leq t < k \\ \beta^{t-k}\sqrt{1-\beta^2}, & k \leq t. \end{cases}$$

A specimen appears in Fig. 1. It can be shown that the atoms span ℓ_2, so they form a dictionary. The absolute inner product between two atoms is

$$|\langle \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_j \rangle| = \beta^{|k-j|}.$$

In particular, each atom has unit norm. It also follows that the coherence of the dictionary equals β.

Here is the calculation of the cumulative coherence function in detail:

$$\begin{aligned} \mu_1(m) &= \max_{|\Lambda|=m} \max_{\boldsymbol{\psi}} \sum_{\Lambda} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle| \\ &= \max_{|\Lambda|=m} \max_{k \notin \Lambda} \sum_{j \in \Lambda} |\langle \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_j \rangle| \\ &= \max_{|\Lambda|=m} \max_{k \notin \Lambda} \sum_{j \in \Lambda} \beta^{|k-j|}. \end{aligned}$$

The maximum occurs, for example, when $k = \lfloor \frac{m}{2} \rfloor$ and

$$\Lambda = \left\{0, 1, 2, \ldots, \left\lfloor \frac{m}{2} \right\rfloor - 1, \left\lfloor \frac{m}{2} \right\rfloor + 1, \ldots, m-1, m\right\}.$$

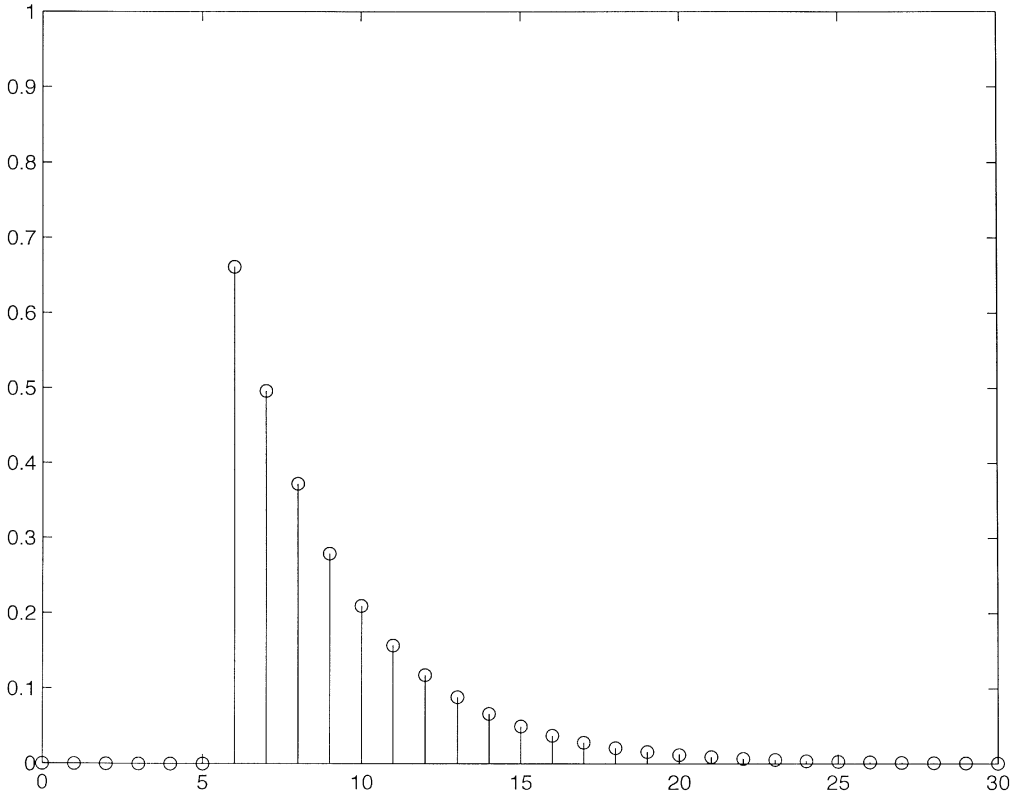

Fig. 1. The atom $\boldsymbol{\varphi}_6$ with parameter $\beta = 0.75$.

The exact form of the cumulative coherence function depends on the parity of m. For m even

$$\mu_1(m) = \frac{2\beta(1-\beta^{m/2})}{1-\beta}$$

while for m odd

$$\mu_1(m) = \frac{2\beta(1-\beta^{(m-1)/2})}{1-\beta} + \beta^{(m+1)/2}.$$

Notice that $\mu_1(m) < 2\beta/(1-\beta)$ for all m. On the other hand, the quantity $m\mu$ grows without bound. Later, we will return to this example to demonstrate how much the cumulative coherence function improves on the coherence parameter.

4) Uniqueness: The cumulative coherence can be used to develop conditions under which m-term representations are unique. The material in this subsection is not essential to understand most of the paper.

The *spark* of a dictionary is the least number of atoms that form a linearly dependent set [8]. The following theorem from [8], [9] is fundamental.

Theorem 2.2 (Donoho–Elad, Gribonval–Nielsen): A necessary and sufficient condition for every linear combination of m atoms to have a unique m-term representation is that $m < \frac{1}{2}\mathrm{spark}(\mathcal{D})$.

We can use the cumulative coherence function and the coherence parameter to develop lower bounds on the spark of a dictionary. Let Φ_m be a matrix whose columns are m distinct atoms, indexed $\lambda_1, \ldots, \lambda_m$. The following lemma and its proof are essentially due to Donoho and Elad [8].

Lemma 2.3: The squared singular values of Φ_m exceed $(1 - \mu_1(m-1))$.

Proof: Consider the Gram matrix $\mathsf{G} \stackrel{\text{def}}{=} (\Phi_m^* \Phi_m)$. The Gershgorin Disc Theorem [24] states that every eigenvalue of G lies in one of the m discs

$$\Delta_k = \left\{ z : |\mathsf{G}_{kk} - z| \leq \sum_{j \neq k} |\mathsf{G}_{jk}| \right\}.$$

The normalization of the atoms implies that $\mathsf{G}_{kk} \equiv 1$. The sum is bounded above by

$$\sum_{j \neq k} |\mathsf{G}_{jk}| = \sum_{j \neq k} \left| \langle \boldsymbol{\varphi}_{\lambda_k}, \boldsymbol{\varphi}_{\lambda_j} \rangle \right| \leq \mu_1(m-1).$$

The result follows since the eigenvalues of G equal the squared singular values of Φ_m. □

If the singular values of Φ_m are nonzero, then the m atoms that comprise the matrix are linearly independent. Lower bounds on the spark follow instantly.

Theorem 2.4 (Donoho–Elad [8]): The spark of a dictionary satisfies the bounds

1) $\mathrm{spark}\,(\mathcal{D}) \geq \min\{m : \mu_1(m-1) \geq 1\}$ and
2) $\mathrm{spark}\,(\mathcal{D}) \geq \mu^{-1} + 1$.

The second bound also appears in [9].

If the dictionary has additional structure, it may be possible to refine these estimates.

Theorem 2.5 (Gribonval–Nielsen [9]): If $\mathcal{D}$ is a μ-coherent dictionary consisting of L orthonormal bases

$$\operatorname{spark}(\mathcal{D}) \geq \left[1 + \frac{1}{L-1}\right] \mu^{-1}.$$

E. Related Work

This subsection contains a brief survey of other major results on sparse approximation, but it makes no pretense of being comprehensive. We will pay close attention to theory about whether or not each algorithm is provably correct.

1) Structured Dictionaries: Early computational techniques for sparse approximation concentrated on specific dictionaries. For example, Coifman and Wickerhauser designed the best orthogonal basis (BOB) algorithm to calculate sparse approximations over wavelet packet and cosine packet dictionaries, which have a natural tree structure. BOB minimizes an entropy function over a subclass of the orthogonal bases contained in the dictionary. Then it returns the best m-term approximation with respect to the distinguished basis [25]. Although BOB frequently produces good results, it does not offer any guarantees on the quality of approximation. Later, Villemoes developed an algorithm that produces provably good approximations over the Haar wavelet packet dictionary [26].

2) OMP and the Sparse Problem: Gilbert, Muthukrishnan, and Strauss have shown that OMP is an approximation algorithm for $(\mathcal{D}, m)$-SPARSE, provided that the dictionary is suitably incoherent [6]. One version of their result is the following.

Theorem 2.6 (Gilbert–Muthukrishnan–Strauss [6]): Let $\mathcal{D}$ have coherence μ, and assume that $m < \frac{1}{8\sqrt{2}}\mu^{-1} - 1$. For an arbitrary signal $\boldsymbol{s}$, OMP generates an m-term approximant $\boldsymbol{a}_m$ that satisfies

$$\|\boldsymbol{s} - \boldsymbol{a}_m\|_2 \leq 8\sqrt{m}\,\|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2$$

where $\boldsymbol{a}_{\text{opt}}$ is an optimal m-term approximation of $\boldsymbol{s}$.

This theorem is a progenitor of the results in the current paper, although the techniques differ significantly.

3) Basis Pursuit (BP): For BP, there is a sequence of attractive results on $(\mathcal{D}, m)$- EXACT-SPARSE. In their seminal paper [23], Donoho and Huo established a connection between uncertainty principles and sparse approximation. Using this link, they proved a recovery theorem for BP.

Theorem 2.7 (Donoho–Huo [23]): Let $\mathcal{D}$ be a union of two orthonormal bases with coherence μ. If $m < \frac{1}{2}(\mu^{-1} + 1)$, then BP recovers every signal that has an m-term representation.

In [7], Elad and Bruckstein made some improvements to the bounds on m, which turn out to be sharp [27]. More recently, the theorem of Donoho and Huo has been extended to multi-ONBs and arbitrary incoherent dictionaries [8], [9]. Donoho and Elad have also developed a generalized notion of incoherence that is equivalent to the cumulative coherence function defined in this paper. We will discuss these results in more detail later.

Very recently, BP has been modified to solve sparse approximation problems for general input signals. These results appear in [10], [11].

III. Recovering Sparse Signals

In this section, we consider the restricted problem $(\mathcal{D}, m)$-EXACT-SPARSE. The major result is a single sufficient condition under which both OMP and BP recover a linear combination of m atoms from the dictionary. We also develop a method for checking when this condition is in force for an arbitrary m-term superposition. Together, these results prove that OMP and BP are both correct algorithms for EXACT-SPARSE over quasi-incoherent dictionaries.

A. The Exact Recovery Condition

Suppose that a given signal $\boldsymbol{s}$ can be written as a linear combination of m atoms and no fewer. Thus,

$$\boldsymbol{s} = \sum_{\lambda \in \Lambda_{\text{opt}}} b_\lambda \boldsymbol{\varphi}_\lambda$$

where Λ_{opt} is a subset of Ω with cardinality m. Without loss of generality, assume that the atoms in Λ_{opt} are linearly independent and that the coefficients b_λ are nonzero. Otherwise, the signal has a representation using fewer than m atoms.

Let Φ_{opt} be the $d \times m$ matrix whose columns are the atoms listed in Λ_{opt}

$$\Phi_{\text{opt}} \stackrel{\text{def}}{=} [\,\boldsymbol{\varphi}_{\lambda_1} \quad \boldsymbol{\varphi}_{\lambda_2} \quad \cdots \quad \boldsymbol{\varphi}_{\lambda_m}\,]$$

where $\Lambda_{\text{opt}} = \{\lambda_1, \ldots, \lambda_m\}$. (The order of the indices is unimportant, so long as it is fixed.) Then the signal can also be expressed as

$$\boldsymbol{s} = \Phi_{\text{opt}} \boldsymbol{b}_{\text{opt}}$$

where $\boldsymbol{b}_{\text{opt}}$ is a vector of m complex coefficients. Since the optimal atoms are linearly independent, Φ_{opt} has full column-rank. Define a second matrix Ψ_{opt} whose columns are the $(N - m)$ atoms indexed by $\Omega \setminus \Lambda_{\text{opt}}$. Thus, Ψ_{opt} contains the atoms that *do not* participate in the optimal representation.

Theorem 3.1 (Exact Recovery for OMP): A sufficient condition for OMP to recover the sparsest representation of the input signal is that

$$\max_{\boldsymbol{\psi}} \left\| \Phi_{\text{opt}}^{+} \boldsymbol{\psi} \right\|_1 < 1 \qquad \text{(ERC)}$$

where $\boldsymbol{\psi}$ ranges over the columns of Ψ_{opt}.

A fortiori, OMP is a correct algorithm for $(\mathcal{D}, m)$-EXACT-SPARSE so long as the condition (ERC) holds for every signal with an m-term representation.

The tag (ERC) abbreviates the phrase "Exact Recovery Condition." It guarantees that no spurious atom can masquerade as part of the signal well enough to fool OMP. Theorem 3.10 of

the sequel shows that (ERC) is essentially the best possible for OMP. Incredibly, (ERC) also provides a natural sufficient condition for BP to recover a sparse signal, which we will discover in Section III-B.

Proof: Suppose that, after the first k steps, OMP has computed an approximant $\boldsymbol{a}_k$ that is a linear combination of k atoms listed in Λ_{opt}. Recall that the residual is defined as $\boldsymbol{r}_k = \boldsymbol{s} - \boldsymbol{a}_k$. We would like to develop a condition to guarantee that the next atom is also optimal.

Observe that the vector $\Phi_{\text{opt}}^* \boldsymbol{r}_k$ lists the inner products between the residual and the optimal atoms. So the expression $\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty$ gives the largest magnitude attained among the inner products, where $\|\cdot\|_\infty$ denotes the ℓ_∞ vector norm. Similarly, $\left\|\Psi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty$ expresses the largest inner product between the residual and any nonoptimal atom. In consequence, to see whether the largest inner product occurs at an optimal atom, we just need to examine the quotient

$$\rho(\boldsymbol{r}_k) \stackrel{\text{def}}{=} \frac{\left\|\Psi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}{\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}. \tag{7}$$

On account of the selection criterion (2), we see that a greedy choice[2] will recover another one of the optimal atoms if and only if $\rho(\boldsymbol{r}_k) < 1$.

Notice that the ratio (7) bears a suspicious resemblance to an induced matrix norm. Before we can apply the usual norm bound, the term $\Phi_{\text{opt}}^* \boldsymbol{r}_k$ must appear in the numerator. Since $\boldsymbol{s}$ and $\boldsymbol{a}_k$ both lie in the column span of Φ_{opt}, so does the residual $\boldsymbol{r}_k$. The matrix $\left(\Phi_{\text{opt}}^+\right)^* \Phi_{\text{opt}}^*$ is a projector onto the column span of Φ_{opt}, and so we may calculate that

$$\begin{aligned}\rho(\boldsymbol{r}_k) &= \frac{\left\|\Psi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}{\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty} \\ &= \frac{\left\|\Psi_{\text{opt}}^* (\Phi_{\text{opt}}^+)^* \Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}{\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty} \\ &\le \left\|\Psi_{\text{opt}}^* (\Phi_{\text{opt}}^+)^*\right\|_{\infty,\infty}.\end{aligned}$$

We use $\|\cdot\|_{p,p}$ to denote the induced norm for linear operators mapping $(\mathbb{C}^d, \|\cdot\|_p)$ into itself. Since $\|\cdot\|_{\infty,\infty}$ equals the maximum absolute *row* sum of its argument and $\|\cdot\|_{1,1}$ equals the maximum absolute *column* sum of its argument, we take a conjugate transpose and switch norms. Continuing the calculation

$$\begin{aligned}\rho(\boldsymbol{r}_k) &\le \left\|\Psi_{\text{opt}}^* (\Phi_{\text{opt}}^+)^*\right\|_{\infty,\infty} \\ &= \left\|\Phi_{\text{opt}}^+ \Psi_{\text{opt}}\right\|_{1,1} \\ &= \max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^+ \boldsymbol{\psi}\right\|_1\end{aligned}$$

where the maximimation occurs over the columns of Ψ_{opt}, the nonoptimal atoms.

[2]In case that $\rho(\boldsymbol{r}_k) = 1$, an optimal atom and a nonoptimal atom both attain the maximal inner product. The algorithm has no provision for determining which one to select. In the sequel, we make the pessimistic assumption that a greedy procedure never chooses an optimal atom when a nonoptimal atom also satisfies the selection criterion. This convention forces greedy techniques to fail for borderline cases, which is appropriate for analyzing algorithmic correctness.

In summary, assuming that $\boldsymbol{r}_k$ lies in the column span of Φ_{opt}, the relation $\rho(\boldsymbol{r}_k) < 1$ will obtain whenever

$$\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^+ \boldsymbol{\psi}\right\|_1 < 1. \tag{ERC}$$

Suppose that (ERC) holds. Since the initial residual $\boldsymbol{r}_0$ lies in the column span of Φ_{opt}, a greedy selection recovers an optimal atom at each step. Each residual is orthogonal to the atoms that have already been selected, so OMP will never choose the same atom twice. It follows that m steps of OMP will identify all m atoms that make up the optimal representation of $\boldsymbol{s}$. Therefore, $\boldsymbol{a}_m = \boldsymbol{s}$. □

An immediate consequence of the proof technique is a result for WOMP.

Corollary 3.2: A sufficient condition for WOMP (α) to recover the sparsest representation of the input signal is that

$$\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^+ \boldsymbol{\psi}\right\|_1 < \alpha \tag{8}$$

where $\boldsymbol{\psi}$ ranges over the columns of Ψ_{opt}.

Gribonval and Nielsen have pointed out that the proofs here also apply to MP [28].

B. Recovery Via BP

It is even easier to prove that the Exact Recovery Condition is sufficient for BP to recover a sparse signal. This theorem will allow us to unify all the recent results about BP. We retain the same notation as before.

Theorem 3.3 (Exact Recovery for BP): A sufficient condition for BP to recover the sparsest representation of the input signal is that

$$\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^+ \boldsymbol{\psi}\right\|_1 < 1 \tag{ERC}$$

where $\boldsymbol{\psi}$ ranges over the atoms that do not participate in Φ_{opt}.

A fortiori, BP is a correct algorithm for $(\mathcal{D}, m)$-EXACT-SPARSE whenever (ERC) holds for every signal that has an m-term representation.

We require a simple lemma about ℓ_1 norms.

Lemma 3.4: Suppose that $\boldsymbol{v}$ is a vector with nonzero components and that A is a matrix whose columns do not have identical ℓ_1 norms. Then $\|\mathsf{A}\boldsymbol{v}\|_1 < \|\mathsf{A}\|_{1,1} \|\boldsymbol{v}\|_1$.

We omit the easy proof and move on to the demonstration of the theorem.

Proof: Suppose that $\boldsymbol{s}$ is an input signal whose sparsest representation can be written as $\boldsymbol{s} = \Phi_{\text{opt}} \boldsymbol{b}_{\text{opt}}$. Assume that the Exact Recovery Condition holds for the input signal.

Let $\boldsymbol{s} = \Phi_{\text{alt}} \boldsymbol{b}_{\text{alt}}$ be a different representation with nonzero coefficients. It follows that Φ_{alt} has at least one column $\boldsymbol{\psi}_0$ that does not appear in Φ_{opt}. According to (ERC), we have $\left\|\Phi_{\text{opt}}^+ \boldsymbol{\psi}_0\right\|_1 < 1$. Meanwhile, $\left\|\Phi_{\text{opt}}^+ \boldsymbol{\varphi}\right\|_1 \le 1$ for every other atom $\boldsymbol{\varphi}$, optimal or nonoptimal.

Assume that the columns of $\Phi_{\text{opt}}^{+}\Phi_{\text{alt}}$ do not have identical ℓ_1 norms. We may use the lemma to calculate that

$$\begin{aligned}\|\boldsymbol{b}_{\text{opt}}\|_1 &= \left\|\Phi_{\text{opt}}^{+}\Phi_{\text{opt}}\boldsymbol{b}_{\text{opt}}\right\|_1 \\ &= \left\|\Phi_{\text{opt}}^{+}\boldsymbol{s}\right\|_1 \\ &= \left\|\Phi_{\text{opt}}^{+}\Phi_{\text{alt}}\boldsymbol{b}_{\text{alt}}\right\|_1 \\ &< \left\|\Phi_{\text{opt}}^{+}\Phi_{\text{alt}}\right\|_{1,1}\|\boldsymbol{b}_{\text{alt}}\|_1 \\ &\leq \|\boldsymbol{b}_{\text{alt}}\|_1 .\end{aligned}$$

If perchance the columns of $\Phi_{\text{opt}}^{+}\Phi_{\text{alt}}$ all have the same ℓ_1 norm, that norm must equal $\left\|\Phi_{\text{opt}}^{+}\boldsymbol{\psi}_0\right\|_1$, which is strictly less than one. Repeat the calculation. Although the first inequality is no longer strict, the second inequality becomes strict in compensation. We reach the same conclusion.

In words, any set of nonoptimal coefficients for representing the signal has strictly larger ℓ_1 norm than the optimal coefficients. Therefore, BP will recover the optimal representation. □

C. Cumulative Coherence Estimates

Since we are unlikely to know the optimal atoms *a priori*, Theorems 3.1 and 3.3 may initially seem useless. But for many dictionaries, the Exact Recovery Condition holds for every m-term signal, so long as m is not too large.

Theorem 3.5: Suppose that μ_1 is the cumulative coherence function of $\mathcal{D}$. The Exact Recovery Condition holds whenever

$$\mu_1(m-1) + \mu_1(m) < 1. \tag{9}$$

Thus, OMP and BP are correct algorithms for $(\mathcal{D}, m)$-SPARSE whenever (9) is in force. In other words, this condition guarantees that either procedure will recover every signal with an m-term representation.

One interpretation of this theorem is that the Exact Recovery Condition holds for sparse signals over quasi-incoherent dictionaries. The present result for BP is slightly stronger than the most general theorem in [8], which is equivalent to Corollary 3.6 of the sequel.

Proof: Begin the calculation by expanding the pseudo-inverse

$$\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^{+}\boldsymbol{\psi}\right\|_1 = \max_{\boldsymbol{\psi}} \left\|(\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})^{-1}\Phi_{\text{opt}}^{*}\boldsymbol{\psi}\right\|_1 .$$

Then apply the usual norm bound

$$\begin{aligned}\max_{\boldsymbol{\psi}} &\left\|(\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})^{-1}\Phi_{\text{opt}}^{*}\boldsymbol{\psi}\right\|_1 \\ &\leq \left\|(\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})^{-1}\right\|_{1,1} \max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^{*}\boldsymbol{\psi}\right\|_1 .\end{aligned} \tag{10}$$

The cumulative coherence function offers a tailor-made estimate of the second factor on the right-hand side of (10)

$$\begin{aligned}\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^{*}\boldsymbol{\psi}\right\|_1 &= \max_{\boldsymbol{\psi}} \sum_{\Lambda_{\text{opt}}} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle| \\ &\leq \mu_1(m).\end{aligned} \tag{11}$$

Bounding the first factor on the right-hand side of (10) requires more sophistication. We develop the inverse as a Neumann series and use Banach algebra methods to estimate its norm. First, notice that $(\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})$ has a unit diagonal because all atoms are normalized. So the off-diagonal part A satisfies

$$\Phi_{\text{opt}}^{*}\Phi_{\text{opt}} = \mathsf{I}_m + \mathsf{A}.$$

Each column of A lists the inner products between one atom of Φ_{opt} and the remaining $(m-1)$ atoms. By definition of the cumulative coherence function

$$\begin{aligned}\|\mathsf{A}\|_{1,1} &= \max_{k} \sum_{j \neq k} |\langle \boldsymbol{\varphi}_{\lambda_k}, \boldsymbol{\varphi}_{\lambda_j} \rangle| \\ &\leq \mu_1(m-1).\end{aligned}$$

Whenever $\|\mathsf{A}\|_{1,1} < 1$, the Neumann series $\sum(-\mathsf{A})^k$ converges to the inverse $(\mathsf{I}_m + \mathsf{A})^{-1}$ [29]. In this case, we may compute

$$\begin{aligned}\left\|(\Phi_{\text{opt}}^{*}\Phi_{\text{opt}})^{-1}\right\|_{1,1} &= \left\|(\mathsf{I}_m + \mathsf{A})^{-1}\right\|_{1,1} \\ &= \left\|\sum_{k=0}^{\infty} (-\mathsf{A})^k\right\|_{1,1} \\ &\leq \sum_{k=0}^{\infty} \|\mathsf{A}\|_{1,1}^{k} \\ &= \frac{1}{1 - \|\mathsf{A}\|_{1,1}} \\ &\leq \frac{1}{1 - \mu_1(m-1)}.\end{aligned} \tag{12}$$

Introduce the bounds (11) and (12) into inequality (10) to obtain

$$\max_{\boldsymbol{\psi}} \left\|\Phi_{\text{opt}}^{+}\boldsymbol{\psi}\right\|_1 \leq \frac{\mu_1(m)}{1 - \mu_1(m-1)}.$$

We reach the result by applying Theorems 3.1 and 3.3. □

A weaker corollary follows directly from basic facts about the cumulative coherence function.

Corollary 3.6: OMP and BP both recover every superposition of m atoms from $\mathcal{D}$ whenever one of the following conditions is satisfied:

$$m < \frac{1}{2}(\mu^{-1} + 1) \tag{13}$$

or

$$\mu_1(m) < \frac{1}{2}. \tag{14}$$

The incoherence condition is the best possible. It would fail for any $\left\lceil \frac{1}{2}(\mu^{-1}+1) \right\rceil$ atoms chosen from an equiangular tight frame with $N = d + 1$ vectors. The bound (8) appears in both [8] and [9] with reference to BP. The bound (14) also appears in [8].

To see the difference between the two conditions in Corollary 3.6, let us return to the dictionary of decaying atoms from Section II-D3. Recall that

$$\mu = \beta \quad \text{and} \quad \mu_1(m) < \frac{2\beta}{1-\beta}.$$

Set $\beta = \frac{1}{5}$. Then the incoherence condition (13) requires that $m < 3$. On the other hand, $\mu_1(m) < \frac{1}{2}$ for every m. Therefore, (14) shows that OMP or BP can recover any (finite) linear combination of decaying atoms!

D. Structured Dictionaries

If the dictionary has special form, better estimates are possible.

Theorem 3.7: Suppose that $\mathcal{D}$ consists of J concatenated orthonormal bases with overall coherence μ. Suppose that the input signal can be written as a superposition of p_j atoms from the jth basis, $j = 1, \dots, J$. Without loss of generality, assume that $0 < p_1 \leq p_2 \leq \cdots \leq p_J$. The Exact Recovery Condition holds whenever

$$\sum_{j=2}^{J} \frac{\mu p_j}{1 + \mu p_j} < \frac{1}{2(1 + \mu p_1)} \tag{15}$$

in which case, both OMP and BP recover the optimal representation of the signal.

The proof of Theorem 3.7 is quite delicate. We refer the interested reader to the technical report [31].

The major theorem of Gribonval and Nielsen's paper [9] is that (15) is a sufficient condition for BP to succeed in this setting. When $J = 2$, we retrieve the major theorem of Elad and Bruckstein's paper [7].

Corollary 3.8: Suppose that $\mathcal{D}$ consists of two orthonormal bases with overall coherence μ, and suppose that the input signal has a representation using p atoms from the first basis and q atoms from the second basis, where $p \leq q$. The Exact Recovery Condition holds whenever

$$2\mu^2 pq + \mu q < 1. \tag{16}$$

Feuer and Nemirovsky have shown that the bound (16) is the best possible for BP [27]. It follows by contraposition that Corollary 3.8 is the best possible result on the Exact Recovery Condition for a two-ONB.

For an arbitrary m-term superposition from a multi-ONB, revisit the calculations of Gribonval and Nielsen [9] to discover the following corollary.

Corollary 3.9: If $\mathcal{D}$ is a μ-coherent dictionary comprised of J orthonormal bases, the condition

$$m < \left[\sqrt{2} - 1 + \frac{1}{2(J-1)}\right] \mu^{-1}$$

is sufficient to ensure that the Exact Recovery Condition holds for every signal with an m-term representation.

The bound in Corollary 3.9 is the best possible when $J = 2$ on account of [27], but Donoho and Elad have pointed out that the result can be improved when $J > 2$.

E. Uniqueness and Recovery

Theorem 3.1 has another important consequence. If the Exact Recovery Condition holds for every linear combination of m atoms, then all m-term superpositions are unique. Otherwise, the Exact Recovery Theorem states that OMP would simultaneously recover two distinct m-term representations of the same signal, a *reductio ad absurdum*. Therefore, the conditions of Theorem 3.5, Corollary 3.6, and Corollary 3.9 ensure that all m-term representations are unique. On the other hand, Theorem 2.2 shows that the Exact Recovery Condition must fail for some linear combination of m atoms whenever $m \geq \frac{1}{2} \operatorname{spark}(\mathcal{D})$.

That a signal has a unique m-term representation does not guarantee the Exact Recovery Condition holds. For a union of two orthonormal bases, Theorem 2.5 implies that all m-term representations are unique whenever $m < \mu^{-1}$. But the discussion in the last section demonstrates that the Exact Recovery Condition may fail for $m \geq \left(\sqrt{2} - \frac{1}{2}\right) \mu^{-1}$. Within this pocket[3] lie uniquely determined signals that cannot be recovered by OMP, as this partial converse of Theorem 3.1 shows.

Theorem 3.10 (Exact Recovery Converse for OMP): Assume that all m-term representations are unique but that the Exact Recovery Condition fails for a signal with optimal synthesis matrix Φ_{opt}. Then there are signals in the column span of Φ_{opt} that OMP cannot recover.

Proof: If the Exact Recovery Condition fails, then

$$\max_{\boldsymbol{\psi}} \left\| \Phi_{\text{opt}}^{+} \boldsymbol{\psi} \right\|_1 \geq 1. \tag{17}$$

By the uniqueness of m-term representations, every signal that has a representation using the atoms in Φ_{opt} yields the same two matrices Φ_{opt} and Ψ_{opt}. Next, choose $\boldsymbol{c}_{\text{bad}} \in \mathbb{C}^m$ to be a vector $\boldsymbol{c}$ for which equality holds in the bound

$$\frac{\left\| \Psi_{\text{opt}}^{*} (\Phi_{\text{opt}}^{+})^{*} \boldsymbol{c} \right\|_\infty}{\|\boldsymbol{c}\|_\infty} \leq \left\| \Psi_{\text{opt}}^{*} (\Phi_{\text{opt}}^{+})^{*} \right\|_{\infty,\infty}.$$

Optimal synthesis matrices have full column rank, so Φ_{opt}^{*} maps the column span of Φ_{opt} onto $\mathbb{C}^m$. Therefore, the column span of Φ_{opt} contains a signal $\boldsymbol{s}_{\text{bad}}$ for which $\Phi_{\text{opt}}^{*} \boldsymbol{s}_{\text{bad}} = \boldsymbol{c}_{\text{bad}}$. Working backward from (17) through the proof of the Exact Recovery Theorem, we discover that $\rho(\boldsymbol{s}_{\text{bad}}) \geq 1$. In conclusion, if we run OMP with $\boldsymbol{s}_{\text{bad}}$ as input, it chooses a nonoptimal atom in the first step. Since Φ_{opt} provides the unique m-term representation of $\boldsymbol{s}_{\text{bad}}$, the initial incorrect selection damns OMP from obtaining an m-term representation of $\boldsymbol{s}_{\text{bad}}$. □

IV. Recovering General Signals

The usual goal of sparse approximation is the analysis or compression of natural signals. But the assumption that a signal has a sparse representation is completely academic on account of the following result.

Proposition 4.1: If $m < d$, the collection of signals that have an exact representation as a linear combination of m atoms forms a set of Lebesgue measure zero in $\mathbb{C}^d$.

Proof: The signals that lie in the span of m distinct atoms form an m-dimensional subspace, which has measure zero. There are $\binom{N}{m}$ ways to choose m atoms, so the collection of signals that have a representation over m atoms is a finite

[3]See the paper of Elad and Bruckstein [7] for a very enlightening graph that delineates the regions of uniqueness and recovery for two-ONB dictionaries.

union of m-dimensional subspaces. This union has measure zero in $\mathbb{C}^d$. □

It follows that a generic signal does not have a sparse representation. Even worse, the optimal m-term approximant is a discontinuous, multivalent function of the input signal. In consequence, proving that an algorithm succeeds for $(\mathcal{D}, m)$-EXACT-SPARSE is very different from proving that it succeeds for $(\mathcal{D}, m)$-SPARSE. Nevertheless, the analysis in Section III-A suggests that OMP may be able to recover atoms from the optimal representation even when the signal is not perfectly sparse.

A. OMP as an Approximation Algorithm

Let $\boldsymbol{s}$ be an arbitrary signal, and suppose that $\boldsymbol{a}_{\text{opt}}$ is an optimal m-term approximation of $\boldsymbol{s}$. That is, $\boldsymbol{a}_{\text{opt}}$ is a solution to the minimization problem (1). Note that $\boldsymbol{a}_{\text{opt}}$ may not be unique. We write

$$\boldsymbol{a}_{\text{opt}} = \sum_{\lambda \in \Lambda_{\text{opt}}} b_\lambda \boldsymbol{\varphi}_\lambda$$

for an index set Λ_{opt} of size m. Once again, denote by Φ_{opt} the $d \times m$ matrix whose columns are the atoms listed in Λ_{opt}. We may assume that the atoms in Λ_{opt} form a linearly independent set because any atom that is linearly dependent on the others could be replaced by a linearly independent atom to improve the quality of the approximation. Let Ψ_{opt} be the matrix whose columns are the $(N - m)$ remaining atoms.

Now we may formulate a condition under which OMP recovers optimal atoms.

Theorem 4.2 (General Recovery): Assume that $\mu_1(m) < \frac{1}{2}$, and suppose that $\boldsymbol{a}_k$ is a linear combination of atoms from Λ_{opt}. At step $(k+1)$, OMP will recover another atom from Λ_{opt} provided that

$$\|\boldsymbol{s} - \boldsymbol{a}_k\|_2 > \sqrt{1 + \frac{m(1-\mu_1(m))}{(1-2\mu_1(m))^2}}\, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 . \qquad (18)$$

We will call (18) the General Recovery Condition. It says that a greedy algorithm makes absolute progress whenever the current k-term approximant compares unfavorably with an optimal m-term approximant. Theorem 4.2 has an important structural implication: *every* optimal representation of a signal contains the same kernel of atoms. This fact follows from the observation that OMP selects the same atoms irrespective of the optimal approximation that appears in the calculation. But the principal corollary of Theorem 4.2 is that OMP is an approximation algorithm for $(\mathcal{D}, m)$-SPARSE.

Corollary 4.3: Assume that $\mu_1(m) < \frac{1}{2}$, and let $\boldsymbol{s}$ be a completely arbitrary signal. Then OMP produces an m-term approximant $\boldsymbol{a}_m$ that satisfies

$$\|\boldsymbol{s} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + C(\mathcal{D}, m)}\, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 \qquad (19)$$

where $\boldsymbol{a}_{\text{opt}}$ is an optimal m-term approximant. We may estimate the constant as

$$C(\mathcal{D}, m) \leq \frac{m(1-\mu_1(m))}{(1-2\mu_1(m))^2}.$$

Proof: Imagine that (18) fails at step $(K+1)$. Then, we have an upper bound on the K-term approximation error as a function of the optimal m-term approximation error. If we continue to apply OMP even after k exceeds K, the approximation error will only continue to decrease. □

Although OMP may not recover an optimal approximant $\boldsymbol{a}_{\text{opt}}$, it always constructs an approximant whose error lies within a constant factor of optimal. One might argue that an approximation algorithm has the potential to inflate a moderate error into a large error. But a moderate error indicates that the signal does not have a good sparse representation over the dictionary, and so sparse approximation may not be an appropriate tool. In practice, if it is easy to find a nearly optimal solution, there is no reason to waste a lot of time and resources to reach the *ne plus ultra*. As the French say, "The best is the enemy of the good."

Placing a restriction on the cumulative coherence function leads to a simpler statement of the result, which generalizes and improves the work in [6].

Corollary 4.4: Assume that $m \leq \frac{1}{3}\mu^{-1}$ or, more generally, that $\mu_1(m) \leq \frac{1}{3}$. Then OMP generates m-term approximants that satisfy

$$\|\boldsymbol{s} - \boldsymbol{a}_m\|_2 \leq \sqrt{1+6m}\, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 . \qquad (20)$$

The constant here is not small, so it is better to regard this as a qualitative theorem on the performance of OMP. See [32] for another greedy algorithm with a much better constant of approximation. Significantly better results for OMP have also been announced in [10], [11].

Let us return again to the example of Section II-D3. This time, set $\beta = \frac{1}{7}$. The coherence condition of Corollary 4.4 suggests that we can achieve the approximation constant $\sqrt{1+6m}$ only if $m = 1, 2$. But the cumulative coherence condition demonstrates that, in fact, the approximation constant is never more than $\sqrt{1+6m}$.

Another consequence of the analysis is a corollary for WOMP.

Corollary 4.5: Weak orthogonal matching pursuit with parameter α calculates m-term approximants that satisfy

$$\frac{\|\boldsymbol{s} - \boldsymbol{a}_m\|_2}{\|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2} \leq \sqrt{1 + \frac{m(1-\mu_1(m))}{(\alpha - (1+\alpha)\mu_1(m))^2}}.$$

As an example, assume that $\mu_1(m) \leq \frac{1}{3}$. Then WOMP $\left(\frac{3}{4}\right)$ has an approximation constant that does not exceed $\sqrt{1+24m}$.

B. Proof of the General Recovery Theorem

Proof: Suppose that, after k steps, OMP has produced an approximant $\boldsymbol{a}_k$ that is a linear combination of k atoms listed

in Λ_{opt}. The residual is $\boldsymbol{r}_k = \boldsymbol{s} - \boldsymbol{a}_k$, and the condition for recovering another optimal atom is

$$\rho(\boldsymbol{r}_k) \stackrel{\text{def}}{=} \frac{\left\|\Psi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}{\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty} < 1.$$

We may divide the ratio into two pieces, which we bound separately

$$\begin{aligned}
\rho(\boldsymbol{r}_k) &= \frac{\left\|\Psi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty}{\left\|\Phi_{\text{opt}}^* \boldsymbol{r}_k\right\|_\infty} \\
&= \frac{\left\|\Psi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_k)\right\|_\infty}{\left\|\Phi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_k)\right\|_\infty} \\
&= \frac{\left\|\Psi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}) + \Psi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty}{\left\|\Phi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}) + \Phi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty} \\
&\le \frac{\left\|\Psi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})\right\|_\infty + \left\|\Psi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty}{\left\|\Phi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty} \\
&\stackrel{\text{def}}{=} \rho_{\text{err}} + \rho_{\text{opt}}. \qquad (21)
\end{aligned}$$

The term $\Phi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})$ has vanished from the denominator since $(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})$ is orthogonal to the column span of Φ_{opt}.

To bound ρ_{opt}, repeat the arguments of Section III-C, *mutatis mutandis*. This yields

$$\begin{aligned}
\rho_{\text{opt}} &\le \frac{\mu_1(m)}{1 - \mu_1(m-1)} \\
&\le \frac{\mu_1(m)}{1 - \mu_1(m)}. \qquad (22)
\end{aligned}$$

Meanwhile, ρ_{err} has the following simple estimate:

$$\begin{aligned}
\rho_{\text{err}} &= \frac{\left\|\Psi_{\text{opt}}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})\right\|_\infty}{\left\|\Phi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty} \\
&= \frac{\max_{\boldsymbol{\psi}} |\boldsymbol{\psi}^* (\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})|}{\left\|\Phi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_\infty} \\
&\le \frac{\max_{\boldsymbol{\psi}} \|\boldsymbol{\psi}\|_2 \, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2}{m^{-1/2} \left\|\Phi_{\text{opt}}^* (\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)\right\|_2} \\
&\le \frac{\sqrt{m} \, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2}{\sigma_{\min}(\Phi_{\text{opt}}) \, \|\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k\|_2}. \qquad (23)
\end{aligned}$$

Since Φ_{opt} has full column rank, $\sigma_{\min}(\Phi_{\text{opt}})$ is nonzero.

Now we can develop a concrete condition under which OMP retrieves optimal atoms. In the following calculation, assume that $\mu_1(m) < \frac{1}{2}$. Combine inequalities (21)–(23). Then estimate the singular value with Lemma 2.3. We discover that $\rho(\boldsymbol{r}_k) < 1$ whenever

$$\frac{\sqrt{m} \, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2}{\sqrt{1 - \mu_1(m)} \, \|\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k\|_2} + \frac{\mu_1(m)}{1 - \mu_1(m)} < 1.$$

Some algebraic manipulations yield the inequality

$$\|\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k\|_2 > \frac{\sqrt{m(1 - \mu_1(m))}}{1 - 2\mu_1(m)} \, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 .$$

Since the vectors $(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})$ and $(\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_k)$ are orthogonal, we may apply the Pythagorean Theorem to reach

$$\|\boldsymbol{s} - \boldsymbol{a}_k\|_2 > \sqrt{1 + \frac{m(1 - \mu_1(m))}{(1 - 2\mu_1(m))^2}} \, \|\boldsymbol{s} - \boldsymbol{a}_{\text{opt}}\|_2 .$$

If this relation is in force, then a step of OMP will retrieve another optimal atom. □

Remark 4.6: The term $\sqrt{m}$ is an unpleasant aspect of (23), but it cannot be avoided without a more subtle approach. When the atoms in our optimal representation have approximately equal correlations with the signal, the estimate of the infinity norm is reasonably accurate. An assumption on the relative size of the coefficients in $\boldsymbol{b}_{\text{opt}}$ might improve the estimate, but this is a severe restriction. An astute reader could whittle the factor down to $\sqrt{m - k}$, but the subsequent analysis would not realize any benefit. It is also possible to strengthen the bound if one postulates a model for the deficit $(\boldsymbol{s} - \boldsymbol{a}_{\text{opt}})$. If, for example, the nonsparse part of the signal were distributed "uniformly" across the dictionary vectors, a single atom would be unlikely to carry the entire error. But we will retreat from a battle that should be fought on behalf of a particular application.

ACKNOWLEDGMENT

This paper would never have been possible without the encouragement and patience of Anna Gilbert, Martin Strauss, and Muthu Muthukrishnan.

REFERENCES

[1] R. A. DeVore, "Nonlinear approximation," *Acta Num.*, pp. 51–150, 1998.

[2] V. Temlyakov, "Nonlinear methods of approximation," *Foundations of Comp. Math.*, vol. 3, no. 1, pp. 33–107, July 2003.

[3] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Boston, MA: Birkhäuser, 2001.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

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

[6] A. C. Gilbert, M. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proc. 14th Annu. ACM-SIAM Symp. Discrete Algorithms*, Baltimore, MD, Jan. 2003, pp. 243–252.

[7] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2558–2567, Sept. 2002.

[8] D. L. Donoho and M. Elad, "Maximal sparsity representation via ℓ_1 minimization," *Proc. Natl. Acad. Sci.*, vol. 100, pp. 2197–2202, Mar. 2003.

[9] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, pp. 3320–3325, Dec. 2003.

[10] J. A. Tropp. (2004, Feb.) Just Relax: Convex Programming Methods for Subset Selection and Sparse Approximation, Univ. Texas at Austin, ICES Rep. 04-04, submitted for publication. [Online]. Available: http://www.ices.utexas.edu/reports/2004.html

[11] D. L. Donoho, M. Elad, and V. N. Temlyakov. (2004, Feb.) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise, submitted for publication. [Online]. Available: http://www.cs.technion.ac.il/~elad/

[12] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.

[13] L. K. Jones, "On a conjecture of Huber concerning the convergence of projection pursuit regression," *Ann. Statist.*, vol. 15, no. 2, pp. 880–882, 1987.

[14] J. H. Friedman and W. Stuetzle, "Projection pursuit regressions," *J. Amer. Statist. Soc.*, vol. 76, pp. 817–823, 1981.
[15] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.
[16] S. Qian and D. Chen, "Signal representation using adaptive normalized Gaussian functions," *Signal Processing*, vol. 36, pp. 329–355, 1994.
[17] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to nonlinear system identification," *Intl. J. Contr.*, vol. 50, no. 5, pp. 1873–1896, 1989.
[18] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signals, Systems and Computers*, Nov. 1993.
[19] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *Opt. Eng.*, vol. 33, no. 7, pp. 2183–2191, July 1994.
[20] S. Sardy, A. G. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *Comp. and Graph. Statist.*, vol. 9, no. 2, 2000.
[21] R. Heath, T. Strohmer, and A. J. Paulraj, "On quasi-orthogonal signatures for CDMA systems," in *Proc. 2002 Allerton Conf. Communication, Control and Computers*.
[22] T. Strohmer and R. W. Heath, "Grassmannian frames with applications to coding and communication," *Appl. Comp. Harmonic Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.
[23] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2845–2862, Nov. 2001.
[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[25] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best-basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 713–718, Mar. 1992.
[26] L. F. Villemoes, "Best approximation with Walsh atoms," *Constr. Approx.*, vol. 13, pp. 329–355, 1997.
[27] A. Feuer and A. Nemirovsky, "On sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1579–1581, June 2003.
[28] R. Gribonval and M. Nielsen, "On the Exponential Convergence of Matching Pursuits in Quasi-Incoherent Dictionaries," Univ. Rennes I, Rennes, France, IRISA Rep. 1619, 2004.
[29] E. Kreyszig, *Introductory Functional Analysis With Applications*. New York: Wiley, 1989.
[30] D. L. Donoho and M. Elad, "On the stability of Basis Pursuit in the presence of noise," working draft.
[31] J. A. Tropp. (2003, Feb.) Greed is Good: Algorithmic Results for Sparse Approximation, Univ. of Texas at Austin, ICES Rep. 03-04. [Online]. Available: http://www.ices.utexas.edu/reports/2003.html
[32] J. A. Tropp, A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, "Improved sparse approximation over quasi-incoherent dictionaries," in *Proc. 2003 IEEE Int. Conf. Image Processing*, vol. 1, Barcelona, Spain, Sept. 2003, pp. I-37–I-40.