GREED IS GOOD:
ALGORITHMIC RESULTS FOR SPARSE APPROXIMATION

JOEL A. TROPP

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 contains a single sufficient condition under which both OMP and Donoho's Basis Pursuit paradigm (BP) can recover an exactly sparse signal. It leverages this theory to show that both OMP and BP can recover all exactly sparse signals from a wide class of dictionaries. These *quasi-incoherent dictionaries* offer a natural generalization of incoherent dictionaries, and the *Babel function* is introduced to quantify the level of incoherence. Indeed, 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 retrieve the common atoms from all optimal representations of a nonsparse signal. From there, it argues that Orthogonal Matching Pursuit is an approximation algorithm for the sparse problem over a quasi-incoherent dictionary. That is, for *every* input signal, OMP can calculate a sparse approximant whose error is only a small factor worse than the optimal error which can be attained with the same number of terms.

1. INTRODUCTION

They were never meant to be together. Some signals just 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 and orthogonal transformations have a distinguished service record, examples like this have led researchers to enlist more complicated techniques.

The most basic instrument of approximation is to project each signal onto a fixed m-dimensional linear subspace. A familiar example is interpolation by means of fixed-knot, polynomial splines. For some functions, this elementary procedure works quite well. Later, various nonlinear methods were developed. One fundamental technique is to project each signal onto the best linear subspace induced 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. It yields tremendous gains over the linear method, especially when the input signals are compatible with the basis [DeV98, Tem02]. But, as noted, some functions fit into an orthonormal basis like a square peg fits a round hole. To deal with this problem, researchers have spent the last fifteen years developing redundant systems, called dictionaries, for analyzing and representing complicated functions. A Gabor dictionary, for example, consists of complex exponentials at different frequencies which are localized to short time intervals. It is used for joint time-frequency analysis [Grö01].

Redundant systems raise the awkward question of how to use them effectively for approximation. The problem of representing a signal with the best linear combination of m elements from a dictionary is called sparse approximation or highly nonlinear approximation. The core algorithmic question:

> For a given class of dictionaries, how does one design a fast algorithm which 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, called 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 those elements which have already been selected. This technique just extends the trivial greedy algorithm which succeeds for an orthonormal system. Basis Pursuit is a more sophisticated approach, which replaces the original sparse approximation problem by a linear programming problem. Empirical evidence suggests that BP is more powerful than OMP [CDS99]. Meanwhile, the major advantage of Orthogonal Matching Pursuit is that it has simple, fast implementations [DMA97, GMS03].

1.1. **Major Results.** I have developed theory for two distinct sparse approximation problems. The EXACT-SPARSE problem is to recover an exact superposition

$$\boldsymbol{x} = \sum_{k=1}^{m} b_k\, \boldsymbol{\varphi}_k$$

of m elements (called atoms) from a redundant dictionary. To state the first theorem, I define a matrix $\boldsymbol{\Phi}_{\text{opt}}$ whose columns are the m atoms that comprise the signal and write $\boldsymbol{\Phi}_{\text{opt}}^{+}$ for its pseudo-inverse.

Theorem A. *Suppose that $\boldsymbol{x}$ is a signal which can be expressed as a linear combination the m atoms in $\boldsymbol{\Phi}_{\text{opt}}$. Both Orthogonal Matching Pursuit and Basis Pursuit recover the m-term representation of $\boldsymbol{x}$ whenever*

$$\max_{\psi} \left\| \boldsymbol{\Phi}_{\text{opt}}^{+} \boldsymbol{\psi} \right\|_1 < 1, \tag{1.1}$$

where the maximization occurs over all atoms $\boldsymbol{\psi}$ which do not participate in the m-term representation.

This result is essentially the best possible for OMP, and it is also the best possible for BP in certain cases. It is remarkable that (1.1) is a natural sufficient condition for such disparate techniques to resolve sparse signals. This fact suggests that EXACT-SPARSE has tremendous structure. Now, Theorem A would not be very useful without a technique for checking when the condition holds. To that end, I define the Babel function, $\mu_1(m)$, of a dictionary, which equals the maximum absolute sum of inner products between a fixed atom and m other atoms. The Babel function provides a natural generalization of the cumulative coherence μm, where μ is the maximum absolute inner product between two atoms. If the Babel function grows slowly, we say that the dictionary is quasi-incoherent.

Theorem B. *The condition (1.1) holds for every superposition of m atoms from a dictionary whenever*

$$m < \tfrac{1}{2}\left(\mu^{-1} + 1\right)$$

or, more generally, whenever

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

If the dictionary consists of J concatenated orthonormal bases, (1.1) is in force whenever

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

Together, Theorems A and B unify all of the recent results for Basis Pursuit [EB02, DE02, GN02] and extend them to Orthogonal Matching Pursuit as well.

The second problem, SPARSE, requests the optimal m-term approximation of a general signal. Although EXACT-SPARSE and SPARSE are related, the latter is much harder to solve. Nevertheless,

Orthogonal Matching Pursuit is a provably good approximation algorithm for the sparse problem over a quasi-incoherent dictionary. To be precise, suppose that $\boldsymbol{a}_{\mathrm{opt}}$ is an optimal m-term approximant of the signal $\boldsymbol{x}$, and let $\boldsymbol{a}_k$ denote the k-th approximant produced by OMP.

Theorem C. *Orthogonal Matching Pursuit will recover an atom from the optimal m-term representation of an arbitrary signal $\boldsymbol{x}$ whenever*

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

Taking $\mu_1(m) \leq \frac{1}{3}$, it follows that Orthogonal Matching Pursuit will calculate an m-term approximant $\boldsymbol{a}_m$ that satisfies

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + 6m}\,\|\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}\|_2\,.$$

This theorem extends work of Gilbert, Muthukrishnan and Strauss [GMS03]. No comparable results are available for the Basis Pursuit paradigm.

2. Background

2.1. **Sparse Approximation Problems.** The standard sparse approximation problem[1] is set in the Hilbert space $\mathbb{C}^d$. A *dictionary* for $\mathbb{C}^d$ is a finite collection $\mathscr{D}$ of unit-norm vectors which spans the whole space. The elements 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 metaphysics. The whole dictionary structure is written as

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

The letter N will indicate the size of the dictionary.

The problem is to approximate a given signal $\boldsymbol{x} \in \mathbb{C}^d$ using a linear combination of m atoms from the dictionary. Since m is taken to be much smaller than the dimension, the approximant is *sparse*. Specifically, we seek a solution to the minimization problem

$$\min_{|\Lambda| = m}\ \min_{\boldsymbol{b}}\ \left\|\boldsymbol{x} - \sum_{\lambda \in \Lambda} b_\lambda\,\boldsymbol{\varphi}_\lambda\right\|_2,\tag{2.1}$$

where the index set $\Lambda \subset \Omega$ and $\boldsymbol{b}$ is a list of complex-valued coefficients. For a fixed Λ, the inner minimization of (2.1) 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 that I have outlined will be called $(\mathscr{D}, m)$-Sparse. Note that it is posed for an *arbitrary* vector 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 to transfer this burden to a preprocessing stage, since we are likely to use the same dictionary for many approximations. A second reason is that Davis, Mallat and Avellaneda have shown that solving or even approximating the solution of (2.1) is NP-hard if the dictionary is unrestricted [DMA97]. Nevertheless, it is not quixotic to seek algorithms for the sparse problem over a *particular* dictionary.

We shall also consider a second problem called $(\mathscr{D}, m)$-Exact-Sparse, where the input signal is restricted to be a linear combination of m atoms or fewer from $\mathscr{D}$. There are several motivations. Although natural signals are not perfectly sparse, one might imagine applications in which a sparse

---

[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.

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.

2.1.1. *Synthesis and Analysis.* Associated with each dictionary is the $d \times N$ *synthesis matrix* $\boldsymbol{\Phi}$ whose columns are atoms. The column order does not matter, so long as it is fixed. That is,

$$\boldsymbol{\Phi} \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\varphi}_{\omega_1} & \boldsymbol{\varphi}_{\omega_2} & \cdots & \boldsymbol{\varphi}_{\omega_N} \end{bmatrix}.$$

The synthesis matrix generates a superposition $\boldsymbol{x}$ from a vector $\boldsymbol{b}$ of coefficients: $\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{b}$. The redundancy of the dictionary permits the same signal to be synthesized from an infinite number of distinct coefficient vectors. The conjugate transpose $\boldsymbol{\Phi}^*$ of the synthesis matrix is called the *analysis matrix*. It maps a vector to a list of inner products with the dictionary: $\boldsymbol{b}' = \boldsymbol{\Phi}^*\boldsymbol{x}$. In general, *nota bene* that $\boldsymbol{\Phi}(\boldsymbol{\Phi}^*\boldsymbol{x}) \neq \boldsymbol{x}$ unless the dictionary is an orthonormal basis!

2.1.2. *Related Problems.* $(\mathscr{D}, m)$-SPARSE exemplifies a large class of linear sparse approximation problems [GMS03]. We shall continue to draw signals from $\mathbb{C}^d$, but now we measure the approximation error with a general norm $\|\cdot\|$. Fix a dictionary $\mathscr{D}$ consisting of N unit-norm vectors which span $\mathbb{C}^d$, and associate with $\mathscr{D}$ the linear synthesis operator $\boldsymbol{\Phi} : \mathbb{C}^N \to \mathbb{C}^d$. Then let $\|\cdot\|_{\text{sp}}$ be a function which measures the sparsity of a coefficient vector. The sparsity function does not need to be a norm, in spite of the notation.

The *primal sparse approximation problem* requests the best approximation to a signal $\boldsymbol{x}$ subject to the condition that the coefficients in the approximation have sparsity less than m. That is,

$$\min_{\boldsymbol{b} \in \mathbb{C}^N} \|\boldsymbol{x} - \boldsymbol{\Phi}\boldsymbol{b}\| \qquad \text{subject to} \qquad \|\boldsymbol{b}\|_{\text{sp}} \leq m. \tag{2.2}$$

Interchanging the objective function and the constraint yields the *dual sparse approximation problem*: Find the sparsest set of coefficients that approximates the signal within a tolerance of ε.

$$\min_{\boldsymbol{b} \in \mathbb{C}^N} \|\boldsymbol{b}\|_{\text{sp}} \qquad \text{subject to} \qquad \|\boldsymbol{x} - \boldsymbol{\Phi}\boldsymbol{b}\| \leq \varepsilon. \tag{2.3}$$

The standard problem, $(\mathscr{D}, m)$-SPARSE, is a primal sparse approximation problem which measures error with the ℓ_2 norm and sparsity with the ℓ_0 quasi-norm[2].

2.2. Dictionary Analysis.

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

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

Roughly speaking, this number measures how much two vectors in the dictionary 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 is called a *multi-ONB*. For some d, it is possible to build a multi-ONB which contains d or

---

[2]The ℓ_0 quasi-norm of a vector equals the number of nonzero components.

even $(d + 1)$ bases yet retains the minimal coherence $\mu = d^{-1/2}$ possible [HSP02]. 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 *optimal Grassmannian frame*. See [SH02, ST03] 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 Matching Pursuit [MZ93]. The first theoretical developments appeared in Donoho and Huo's paper [DH01]. Stronger results for Basis Pursuit, phrased in terms of coherence, were provided in [EB02, DE02, GN02]. Most recently, Gilbert, Muthukrishnan and Strauss have exhibited an approximation algorithm for sparse problems over suitably incoherent dictionaries [GMS03].

2.2.2. *The Babel Function.* The coherence parameter does not offer a very subtle description of a dictionary 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 that end, I introduce the *Babel function*, which measures the maximum total coherence between a fixed atom and a collection of other atoms. In a sense, the Babel function indicates how much the atoms are "speaking the same language." It's 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 Babel function will arise naturally in my 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 [DE02].

Formally, the Babel function is defined by

$$\mu_1(m) \stackrel{\text{def}}{=} \max_{|\Lambda| = m} \max_{\boldsymbol{\psi}} \sum_{\Lambda} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle|, \tag{2.4}$$

where the vector $\boldsymbol{\psi}$ ranges over the atoms indexed by $\Omega \setminus \Lambda$. The subscript in the notation serves to distinguish the Babel function from the coherence and to remind us that it is an absolute sum. A close examination of the formula shows that $\mu_1(1) = \mu$ and that μ_1 is a non-decreasing function of m. Place the convention that $\mu_1(0) = 0$. When the Babel function of a dictionary grows slowly, we say informally that the dictionary is *quasi-incoherent*.

The next proposition shows that the Babel function is a direct generalization of the cumulative coherence.

Proposition 2.1. *If a dictionary has coherence μ, then $\mu_1(m) \leq \mu m$.*

Proof. Calculate that

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

$\square$

2.2.3. *An Example.* For a realistic dictionary where the atoms have analytic definitions, the Babel function is not too difficult to compute. As a simple example, consider a dictionary of decaying pulses. 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 the sequence

$$\varphi_k(t) = \begin{cases} 0, & 0 \leq t < k, \\ \beta^{t-k}\,\sqrt{1 - \beta^2}, & k \leq t. \end{cases}$$
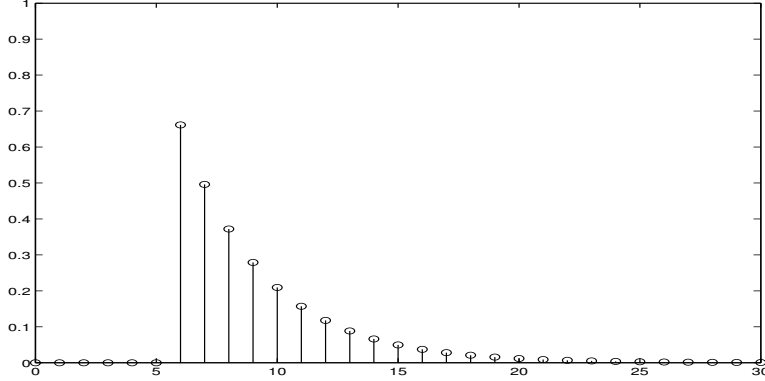
FIGURE 1. In captivity, a pulse at $k = 6$ with $\beta = 0.75$.

A specimen appears in Figure 1. It can be shown that the pulses 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 pulse has unit norm. It also follows that the coherence $\mu = \beta$. Now, here is the calculation of the Babel function in lurid detail:

$$\mu_1(m) = \max_{|\Lambda|=m} \max_{\boldsymbol{\psi}} \sum_{\Lambda} |\langle \boldsymbol{\psi}, \boldsymbol{\varphi}_\lambda \rangle| = \max_{|\Lambda|=m} \max_{k \notin \Lambda} \sum_{j \in \Lambda} |\langle \boldsymbol{\varphi}_k, \boldsymbol{\varphi}_j \rangle|$$

$$= \max_{|\Lambda|=m} \max_{k \notin \Lambda} \sum_{j \in \Lambda} \beta^{|k-j|}.$$

The maximum occurs, for example, when $\Lambda = \{0, 1, 2, \ldots, \lfloor \frac{m}{2} \rfloor - 1, \lfloor \frac{m}{2} \rfloor + 1, \ldots, m-1, m\}$ and $k = \lfloor \frac{m}{2} \rfloor$. The symbolic form of the Babel function depends on the parity of m.

$$\mu_1(m) = \begin{cases} \dfrac{2\beta\left(1 - \beta^{m/2}\right)}{1 - \beta} & \text{for } m \text{ even, and} \\[3ex] \dfrac{2\beta\left(1 - \beta^{(m-1)/2}\right)}{1 - \beta} + \beta^{(m+1)/2} & \text{for } m \text{ odd.} \end{cases}$$

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

2.2.4. *Spark.* The *spark* of a matrix is the least number of columns that form a linearly dependent set. Compare this against the matrix rank, which is the greatest number of linearly *independent* columns [DE02]. In coding theory, the spark of a codebook would be called the *distance* of the code. The following theorem is fundamental.

Theorem 2.2 (Donoho-Elad, Gribonval-Nielsen [DE02, GN02])**.** *All sparse representations over* m *atoms are unique if and only if* $m < \frac{1}{2} \text{ spark } \boldsymbol{\Phi}$.

We can use the Babel function and the coherence parameter to develop lower bounds on the spark of a dictionary. First, let $\boldsymbol{\Phi}_m$ be a matrix whose columns are m distinct atoms. The following lemma and its proof are essentially due to Donoho and Elad [DE02].

Lemma 2.3. *The squared singular values of* $\boldsymbol{\Phi}_m$ *exceed* $(1 - \mu_1(m-1))$.

Proof. Consider the Gram matrix $G \stackrel{\text{def}}{=} (\boldsymbol{\Phi}_m^* \boldsymbol{\Phi}_m)$. The Geršgorin Disc Theorem [HJ85] states that every eigenvalue of G lies in one of the m discs

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

The normalization of the atoms implies that $G_{kk} \equiv 1$. The sum is bounded above by $\sum_{j \neq k} |G_{jk}| = \sum_{j \neq k} |\langle \boldsymbol{\varphi}_{\lambda_k}, \boldsymbol{\varphi}_{\lambda_j} \rangle| \leq \mu_1(m - 1)$. The result follows since the eigenvalues of G equal the squared singular values of $\boldsymbol{\Phi}_m$. $\square$

If the singular values of $\boldsymbol{\Phi}_m$ are nonzero, then the m atoms which comprise the matrix are linearly independent, whence

Theorem 2.4 (Donoho-Elad [DE02])**.** *Lower bounds on the spark of a dictionary are*

$$\operatorname{spark} \boldsymbol{\Phi} \geq \min\{m : \mu_1(m - 1) \geq 1\}, \qquad and$$
$$\operatorname{spark} \boldsymbol{\Phi} \geq \mu^{-1} + 1.$$

The coherence result also appears in [GN02]. For structured dictionaries, better estimates are possible. For example,

Theorem 2.5 (Gribonval-Nielsen [GN02])**.** *If $\mathscr{D}$ is a μ-coherent dictionary consisting of L orthonormal bases,*

$$\operatorname{spark} \boldsymbol{\Phi} \geq \left[1 + \frac{1}{L - 1} \right] \mu^{-1}.$$

2.3. **Greedy Algorithms.** If the dictionary is an orthonormal basis, the sparse approximation problem has a straightforward solution. It is possible to build the approximation one term at a time by selecting at each step the atom which correlates most strongly with the residual signal. Greedy techniques for sparse approximation extend this idea to more general dictionaries.

2.3.1. *Matching Pursuit.* The simplest of the greedy procedures is Matching Pursuit (MP), which Mallat and Zhang introduced to the signal processing community [MZ93]. Matching Pursuit begins by setting the initial residual equal to the signal and making a trivial initial approximation. That is,

$$\boldsymbol{r}_0 \stackrel{\text{def}}{=} \boldsymbol{x}, \qquad \text{and} \qquad \boldsymbol{a}_0 \stackrel{\text{def}}{=} \boldsymbol{0}.$$

At step k, MP chooses an atom $\boldsymbol{\varphi}_{\lambda_k}$ that solves the easy optimization problem

$$\lambda_k \stackrel{\text{def}}{=} \arg\max_\omega |\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_\omega \rangle|. \tag{2.5}$$

Then, the procedure calculates a new approximation and a new residual:

$$\boldsymbol{a}_k \stackrel{\text{def}}{=} \boldsymbol{a}_{k-1} + \langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k} \rangle \, \boldsymbol{\varphi}_{\lambda_k}, \qquad \text{and}$$
$$\boldsymbol{r}_k \stackrel{\text{def}}{=} \boldsymbol{r}_{k-1} - \langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k} \rangle \, \boldsymbol{\varphi}_{\lambda_k}. \tag{2.6}$$

The residual can also be expressed as $\boldsymbol{r}_k = \boldsymbol{x} - \boldsymbol{a}_k$. If 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 [Jon87]. In fact, this convergence is exponential [DMA97].

Greedy techniques for sparse approximation were developed in the statistics community under the name Projection Pursuit Regression [FS81]. In the approximation commmunitity, MP is known as the Pure Greedy Algorithm [Tem02]. Qian and Chen [QC94] suggested the same algorithm for time-frequency analysis independently of Mallat and Zhang. For more history, theory and an comprehensive list of references, see Temlyakov's monograph [Tem02].

2.3.2. *Orthogonal Matching Pursuit.* Orthogonal Matching Pursuit (OMP) adds a least-squares minimization at each step to obtain the best approximation of the signal over the atoms which have already been chosen. This revision significantly improves the rate of convergence.

At each step of OMP, an atom is selected according to the same rule as MP, via (2.5). But the approximations are calculated differently. Let $\Lambda_k \stackrel{\text{def}}{=} \{\lambda_1, \ldots, \lambda_k\}$ list the atoms which have been selected at step k. Then the k-th approximant is

$$\boldsymbol{a}_k \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{a}} \|\boldsymbol{x} - \boldsymbol{a}\|_2, \qquad \boldsymbol{a} \in \text{span}\{\boldsymbol{\varphi}_\lambda : \lambda \in \Lambda_k\}. \tag{2.7}$$

This minimization can be performed incrementally with standard least-squares techniques. As before, the residual is calculated as $\boldsymbol{r}_k \stackrel{\text{def}}{=} \boldsymbol{x} - \boldsymbol{a}_k$. It is not difficult to check that the residual equals zero after d steps.

Orthogonal Matching Pursuit was developed independently by many researchers. The earliest reference appears to be a 1989 paper of Chen, Billings and Luo [CBL89]. The first signal processing papers on OMP arrived around 1993 [PRK93, DMZ94].

2.3.3. *OMP and the Sparse Problem.* Gilbert, Muthukrishnan and Strauss have shown that Orthogonal Matching Pursuit is an approximation algorithm for $(\mathscr{D}, m)$-SPARSE when the dictionary is suitably incoherent [GMS03]. One version of their result is

Theorem 2.6 (Gilbert-Muthukrishnan-Strauss [GMS03])**.** *Let $\mathscr{D}$ have coherence μ, and assume that $m < \frac{1}{8\sqrt{2}}\mu^{-1} - 1$. For an arbitrary signal $\boldsymbol{x}$, Orthogonal Matching Pursuit generates an m-term approximant $\boldsymbol{a}_m$ which satisfies*

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

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

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

2.3.4. *Weak Greedy Algorithms.* Orthogonal Matching Pursuit has a cousin called Weak Orthogonal Matching Pursuit (WOMP) that makes a cameo appearance in this article. Instead of selecting the optimal atom at each step, WOMP settles for one which is nearly-optimal. Specifically, it finds an index λ_k so that

$$|\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k}\rangle| \geq \alpha \, \max_\omega |\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_\omega\rangle|, \tag{2.8}$$

where $\alpha \in (0, 1]$ is a fixed *weakness parameter*. Once the new atom is chosen, the approximation is calculated as before, via (2.7). WOMP(α) has essentially the same convergence properties as OMP [Tem02].

2.4. **Other Related Work.** This section contains a brief survey of other major results on sparse approximation, without making any claims to be comprehensive. I am particularly interested in theory about whether or not each algorithm is provably correct.

2.4.1. *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 a certain objective function over a subclass of the orthogonal bases contained in the dictionary. Then it performs the best m-term approximation with respect to the selected basis [CW92]. Although BOB frequently produces good results, it does not offer any guarantees on the quality of approximation. Later, Villemoes developed an algorithm for the Haar wavelet packet dictionary, that produces provably good approximations with a low time cost [Vil97]. Villemoes' result is a serious coup, even though Haar wavelets have limited applicability.

2.4.2. *Basis Pursuit.* The other major approach to sparse approximation is the Basis Pursuit (BP) paradigm, developed by Chen, Donoho and Saunders. Strictly speaking, BP is not an algorithm but a principle. The key idea is to replace the original primal problem,

$$\min_{\boldsymbol{b}} \|\boldsymbol{x} - \boldsymbol{\Phi}\,\boldsymbol{b}\|_2 \qquad \text{subject to} \qquad \|\boldsymbol{b}\|_0 = m,$$

by a variant of the dual problem,

$$\min_{\boldsymbol{b}} \|\boldsymbol{b}\|_1 \qquad \text{subject to} \qquad \boldsymbol{\Phi}\,\boldsymbol{b} = \boldsymbol{x},$$

and hope that the solutions coincide [CDS99].

At least two algorithms have been proposed for solving the Basis Pursuit problem. The original paper advocates interior-point methods of linear programming [CDS99]. More recently, Sardy, Bruce and Tseng have suggested another procedure called Block Coordinate Relaxation [SBT00]. Both techniques are computationally intensive.

At present, the Basis Pursuit paradigm offers no approximation guarantees for the general sparse approximation problem. There is, however, a sequence of intriguing results for $(\mathscr{D}, m)$-EXACT-SPARSE. In their seminal paper [DH01], Donoho and Huo established a connection between uncertainty principles and sparse approximation. In particular, they proved

Theorem 2.7 (Donoho-Huo [DH01]). *Let $\mathscr{D}$ be a union of two orthonormal bases with coherence μ. If $m < \frac{1}{2}(\mu^{-1} + 1)$, then Basis Pursuit can recover any superposition of m atoms from the dictionary.*

In [EB02], Elad and Bruckstein made some improvements to the bound on m, which turn out to be optimal [FN]. More recently, the theorem of Donoho and Huo has been extended to multi-ONBs and arbitrary incoherent dictionaries [DE02, GN02]. Donoho and Elad have also developed a generalized notion of incoherence that is equivalent to the Babel function defined in this article. I discuss these results elsewhere in the text.

3. RECOVERING SPARSE SIGNALS

In this section, I consider the restricted problem $(\mathscr{D}, m)$-EXACT-SPARSE. The major result is a single sufficient condition under which both Orthogonal Matching Pursuit and Basis Pursuit recover a given linear combination of m atoms from the dictionary. Then, I show how to check 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.

3.1. **The Exact Recovery Condition.** Imagine that a given signal $\boldsymbol{x}$ has a representation over m atoms, say

$$\boldsymbol{x} = \sum_{\Lambda_{\text{opt}}} b_\lambda\,\boldsymbol{\varphi}_\lambda,$$

where $\Lambda_{\text{opt}} \subset \Omega$ is an index set of size 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 an exact representation using fewer than m atoms.

From the dictionary synthesis matrix, extract the $d \times m$ matrix $\boldsymbol{\Phi}_{\text{opt}}$ whose columns are the atoms listed in Λ_{opt}:

$$\boldsymbol{\Phi}_{\text{opt}} \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\varphi}_{\lambda_1} & \boldsymbol{\varphi}_{\lambda_2} & \cdots & \boldsymbol{\varphi}_{\lambda_m} \end{bmatrix},$$

where λ_k ranges over Λ_{opt}. Then, the signal can also be expressed as

$$\boldsymbol{x} = \boldsymbol{\Phi}_{\text{opt}}\boldsymbol{b}_{\text{opt}}$$

for a vector of m complex coefficients, $\boldsymbol{b}_{\text{opt}}$. Since the optimal atoms are linearly independent, $\boldsymbol{\Phi}_{\text{opt}}$ has full column-rank. Define a second matrix $\boldsymbol{\Psi}_{\text{opt}}$ whose columns are the $(N - m)$ atoms indexed

by $\Omega \setminus \Lambda_{\mathrm{opt}}$. Then $\boldsymbol{\Psi}_{\mathrm{opt}}$ contains the atoms which *do not* participate in the optimal representation. Using this notation, I state

Theorem 3.1 (Exact Recovery for OMP). *A sufficient condition for Orthogonal Matching Pursuit to resolve $\boldsymbol{x}$ completely in m steps is that*

$$\max_{\boldsymbol{\psi}} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{+} \boldsymbol{\psi} \right\|_1 < 1, \tag{3.1}$$

where $\boldsymbol{\psi}$ ranges over the columns of $\boldsymbol{\Psi}_{\mathrm{opt}}$.

*A fortiori, Orthogonal Matching Pursuit is a correct algorithm for $(\mathscr{D}, m)$-*EXACT-SPARSE *whenever the condition* (3.1) *holds for every superposition of m atoms from $\mathscr{D}$.*

I call (3.1) the Exact Recovery Condition. It guarantees that no spurious atom can masquerade as part of the signal well enough to fool Orthogonal Matching Pursuit. Theorem 3.10 of the sequel shows that (3.1) is essentially the best possible for OMP. Incredibly, (3.1) also provides a natural sufficient condition for Basis Pursuit to recover a sparse signal, which I prove in Section 3.2.

Proof. After the first k steps, suppose that Orthogonal Matching Pursuit has recovered an approximant $\boldsymbol{a}_k$ which is a linear combination of k atoms listed in Λ_{opt}. We would like to develop a condition which can guarantee that the next atom is also optimal.

Observe that the vector $\boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k$ lists the inner products between $\boldsymbol{r}_k$ and the optimal atoms. So the number $\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}$ equals the largest of these inner products in magnitude. Similarly, $\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}$ expresses the largest inner product between the residual and any non-optimal atom. In consequence, to see whether the largest inner product occurs at an optimal atom, we just need to examine the quotient

$$\rho(\boldsymbol{r}_k) \stackrel{\text{def}}{=} \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}}. \tag{3.2}$$

On account of the selection criterion (2.5), we see that a greedy choice[3] will recover another one of the optimal atoms if and only if $\rho(\boldsymbol{r}_k) < 1$.

Notice that the ratio (3.2) bears a suspicious resemblance to an induced matrix norm. Before we can apply the usual norm bound, the term $\boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k$ must appear in the numerator. By assumption, $\boldsymbol{r}_k = \boldsymbol{x} - \boldsymbol{a}_k$ lies in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$, and the matrix $(\boldsymbol{\Phi}_{\mathrm{opt}}^{+})^{*} \boldsymbol{\Phi}_{\mathrm{opt}}^{*}$ is a projection onto the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$. Therefore,

$$\begin{aligned}
\rho(\boldsymbol{r}_k) &= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}} \\
&= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^{*} (\boldsymbol{\Phi}_{\mathrm{opt}}^{+})^{*} \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{r}_k \right\|_{\infty}} \\
&\leq \left\| \boldsymbol{\Psi}_{\mathrm{opt}}^{*} (\boldsymbol{\Phi}_{\mathrm{opt}}^{+})^{*} \right\|_{\infty, \infty}.
\end{aligned}$$

---

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

I use $\|\cdot\|_{p,p}$ to denote the norm for linear operators mapping $(\mathbb{C}^d, \|\cdot\|_p)$ onto 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 can take a conjugate transpose and switch norms.

$$
\begin{aligned}
\rho(\boldsymbol{r}_k) &\leq \left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{\Phi}_{\mathrm{opt}}^+)^* \right\|_{\infty,\infty} \\
&= \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^+ \boldsymbol{\Psi}_{\mathrm{opt}} \right\|_{1,1} \\
&= \max_{\boldsymbol{\psi}} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^+ \boldsymbol{\psi} \right\|_1 ,
\end{aligned}
$$

where the maximimation occurs over the columns of $\boldsymbol{\Psi}_{\mathrm{opt}}$.

Assuming that $\boldsymbol{r}_k$ lies in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$, the relation $\rho(\boldsymbol{r}_k) < 1$ will obtain whenever

$$
\max_{\boldsymbol{\psi}} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^+ \boldsymbol{\psi} \right\|_1 < 1. \tag{3.3}
$$

Suppose that (3.3) holds. Since the initial residual $\boldsymbol{r}_0$ lies in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$, a greedy selection recovers an optimal atom at each step. Each residual is orthogonal to the atoms which have already been selected, so OMP will never choose the same atom twice. It follows that m steps of OMP will retrieve all m atoms which comprise $\boldsymbol{x}$. Therefore, $\boldsymbol{a}_m = \boldsymbol{x}$. $\qquad \square$

An immediate consequence of the proof technique is a result for Weak Orthogonal Matching Pursuit.

Corollary 3.2. *A sufficient condition for* $WOMP(\alpha)$ *to resolve* $\boldsymbol{x}$ *completely in* m *steps is that*

$$
\max_{\boldsymbol{\psi}} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^+ \boldsymbol{\psi} \right\|_1 < \alpha, \tag{3.4}
$$

where $\boldsymbol{\psi}$ *ranges over the columns of* $\boldsymbol{\Psi}_{\mathrm{opt}}$.

3.2. **Recovery via Basis Pursuit.** It is even easier to prove that the Exact Recovery Condition is sufficient for Basis Pursuit to recover a sparse signal. This theorem will allow me to unify all the recent results about BP.

Theorem 3.3 (Exact Recovery for BP). *A sufficient condition for Basis Pursuit to recover the optimal representation of a sparse signal* $\boldsymbol{x} = \boldsymbol{\Phi}_{\mathrm{opt}} \boldsymbol{b}_{\mathrm{opt}}$ *is that*

$$
\max_{\boldsymbol{\psi}} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^+ \boldsymbol{\psi} \right\|_1 < 1, \tag{3.5}
$$

where $\boldsymbol{\psi}$ *ranges over the atoms which do not participate in* $\boldsymbol{\Phi}_{\mathrm{opt}}$.

A fortiori, Basis Pursuit is a correct algorithm for $(\mathscr{D}, m)$-EXACT-SPARSE *whenever the condition* (3.5) *holds for every superposition of* m *atoms from* $\mathscr{D}$.

The proof requires 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* $\{A_k\}$ *do not have identical* ℓ_1 *norms. Then* $\|A\boldsymbol{v}\|_1 < \|A\|_{1,1} \|\boldsymbol{v}\|_1$.

Proof. Calculate that

$$
\begin{aligned}
\|A\boldsymbol{v}\|_1 &\leq \sum_{j,k} |A_{jk}| \, |\boldsymbol{v}_k| = \sum_k \|A_k\|_1 \, |\boldsymbol{v}_k| \\
&< \max_k \|A_k\|_1 \sum_k |\boldsymbol{v}_k| = \|A\|_{1,1} \|\boldsymbol{v}\|_1 .
\end{aligned}
$$

$\qquad \square$

Proof of Theorem 3.3. Suppose that the optimal representation of a signal is $\boldsymbol{x} = \boldsymbol{\Phi}_{\mathrm{opt}}\boldsymbol{b}_{\mathrm{opt}}$, and assume that the Exact Recovery Condition (3.5) holds for this signal $\boldsymbol{x}$.

Now, let $\boldsymbol{x} = \boldsymbol{\Phi}_{\mathrm{alt}}\boldsymbol{b}_{\mathrm{alt}}$ be a different representation of the signal with nonzero coefficients. It follows that $\boldsymbol{\Phi}_{\mathrm{alt}}$ contains at least one atom $\boldsymbol{\psi}_0$ which does not appear in $\boldsymbol{\Phi}_{\mathrm{opt}}$. According to (3.5), $\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\psi}_0\right\|_1 < 1$. Meanwhile, $\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\varphi}\right\|_1 \leq 1$ for every other atom $\boldsymbol{\varphi}$, optimal or non-optimal.

First, assume that the columns of $\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\Phi}_{\mathrm{alt}}$ do not have identical ℓ_1 norms. We may use the lemma to calculate that

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

If perchance the columns of $\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\Phi}_{\mathrm{alt}}$ all have the same ℓ_1 norm, that norm must equal $\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\psi}_0\right\|_1 < 1$. 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 non-optimal coefficients for representing the signal has strictly larger ℓ_1 norm than the optimal coefficients. Therefore, Basis Pursuit will recover the optimal representation. $\square$

3.3. Babel Function 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 superposition, so long as m is not too large.

Theorem 3.5. *Suppose that μ_1 is the Babel function of $\mathscr{D}$. The Exact Recovery Condition holds whenever*

$$\mu_1(m-1) + \mu_1(m) < 1. \tag{3.6}$$

*Thus, Orthogonal Matching Pursuit and Basis Pursuit are correct algorithms for $(\mathscr{D}, m)$-*SPARSE *whenever (3.6) is in force. In other words, this condition guarantees that either procedure will recover an arbitrary superposition of m atoms from $\mathscr{D}$.*

One interpretation of this theorem is that the Exact Recovery Condition holds for quasi-incoherent dictionaries. The result for Basis Pursuit is slightly stronger than the most general theorem in [DE02], which is equivalent to Corollary 3.6 of the sequel.

Proof. Begin the calculation by expanding the pseudo-inverse and applying a norm bound.

$$\begin{aligned}
\max_{\boldsymbol{\psi}}\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^+\boldsymbol{\psi}\right\|_1 = \max_{\boldsymbol{\psi}}\left\|(\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\Phi}_{\mathrm{opt}})^{-1}\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\psi}\right\|_1 \\
\leq \left\|(\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\Phi}_{\mathrm{opt}})^{-1}\right\|_{1,1}\max_{\boldsymbol{\psi}}\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\psi}\right\|_1.
\end{aligned} \tag{3.7}$$

The Babel function offers a tailor-made estimate of the second term of (3.7);

$$\max_{\boldsymbol{\psi}}\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\psi}\right\|_1 = \max_{\boldsymbol{\psi}}\sum_{\Lambda_{\mathrm{opt}}}\left|\langle\boldsymbol{\psi},\boldsymbol{\varphi}_\lambda\rangle\right| \leq \mu_1(m). \tag{3.8}$$

Bounding the first term of (3.7) requires more sophistication. We develop the inverse as a von Neumann series and use Banach algebra methods to estimate its norm. First, notice that $(\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\Phi}_{\mathrm{opt}})$ has a unit diagonal because all atoms are normalized. So the off-diagonal part A satisfies

$$\boldsymbol{\Phi}_{\mathrm{opt}}^*\boldsymbol{\Phi}_{\mathrm{opt}} = \mathsf{I}_m + A.$$

Each column of A lists the inner products between one atom of $\boldsymbol{\Phi}_{\mathrm{opt}}$ and the remaining $(m-1)$ atoms. By definition of the Babel function,

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

Whenever $\|A\|_{1,1} < 1$, the von Neumann series $\sum(-A)^k$ converges to the inverse $(I_m + A)^{-1}$ [Kre89]. In this case, we may compute

$$\left\|(\boldsymbol{\Phi}_{\text{opt}}^* \, \boldsymbol{\Phi}_{\text{opt}})^{-1}\right\|_{1,1} = \left\|(I_m + A)^{-1}\right\|_{1,1} = \left\|\sum_{k=0}^{\infty}(-A)^k\right\|_{1,1}$$

$$\leq \sum_{k=0}^{\infty}\|A\|_{1,1}^k = \frac{1}{1 - \|A\|_{1,1}} \tag{3.9}$$

$$\leq \frac{1}{1 - \mu_1(m-1)}.$$

Introduce the estimates (3.8) and (3.9) into inequality (3.7) to obtain

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

We reach the result by applying Theorems 3.1 and 3.3. $\qquad\square$

A corollary follows directly from basic facts about the Babel function.

Corollary 3.6. *Orthogonal Matching Pursuit and Basis Pursuit both recover every superposition of m atoms from $\mathscr{D}$ whenever one of the following conditions is satisfied:*

$$m < \tfrac{1}{2}(\mu^{-1} + 1), \qquad \text{or, more generally,} \tag{3.10}$$

$$\mu_1(m) < \tfrac{1}{2}. \tag{3.11}$$

The incoherence condition is the best possible. It would fail for any $\lceil\tfrac{1}{2}(\mu^{-1} + 1)\rceil$ atoms chosen from an optimal Grassmannian frame with $N = d + 1$ vectors. The bound (3.10) appears in both [DE02, GN02] with reference to Basis Pursuit. The second bound also appears in [DE02].

To see the difference between the two conditions in Corollary 3.6, let us return to the dictionary of decaying pulses from Section 2.2.3. Recall that

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

Set $\beta = \tfrac{1}{5}$. Then the incoherence condition (3.10) requires that $m < 3$. On the other hand, $\mu_1(m) < \tfrac{1}{2}$ for every m. Therefore, (3.11) shows that OMP or BP can recover any (finite) linear combination of pulses!

3.4. **Structured Dictionaries.** If the dictionary has special form, better estimates are possible.

Theorem 3.7. *Suppose that $\mathscr{D}$ consists of J concatenated orthonormal bases with overall coherence μ. Let $\boldsymbol{x}$ be a superposition of p_j atoms from the j-th basis, $j = 1, \ldots, J$. Without loss of generality, assume that $0 < p_1 \leq p_2 \leq \cdots \leq p_J$. The Exact Recovery Condition holds whenever*

$$\sum_{j=2}^{J}\frac{\mu\, p_j}{1 + \mu\, p_j} < \frac{1}{2\,(1 + \mu\, p_1)}. \tag{3.12}$$

In which case both Orthogonal Matching Pursuit and Basis Pursuit recover the sparse representation.

The major theorem of Gribonval and Nielsen's paper [GN02] is that (3.12) is a sufficient condition for Basis Pursuit to succeed in this setting. When $J = 2$, we retrieve the major theorem of Elad and Bruckstein's paper [EB02]:

Corollary 3.8. *Suppose that $\mathscr{D}$ consist of two orthonormal bases with coherence μ, and let $\boldsymbol{x}$ be a signal consisting of p atoms from the first basis and q atoms from the second basis, where $p \leq q$. The Exact Recovery Condition holds whenever*

$$2\,\mu^2\,pq + \mu q < 1. \tag{3.13}$$

Feuer and Nemirovsky have shown that the bound (3.13) is the best possible for BP [FN]. It follows by contraposition that the present result on the Exact Recovery Condition is the best possible for a two-ONB. For an arbitrary m-term superposition from a multi-ONB, revisit the calculations of Gribonval and Nielsen [GN02] to discover

Corollary 3.9. *If $\mathscr{D}$ 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 all m-term superpositions.

The proof of Theorem 3.7 could be used to torture prisoners, so it is cordoned off in Appendix A where it won't hurt anyone.

3.5. **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 absurdem*. Therefore, the conditions of Theorem 3.5, Corollary 3.6 and Corollary 3.9 guarantee that 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 $m \geq \frac{1}{2}\operatorname{spark}\boldsymbol{\Phi}$.

Uniqueness does not prove that 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 $m \geq (\sqrt{2} - \frac{1}{2})\,\mu^{-1}$. Within this pocket[4] lie uniquely determined signals which cannot be recovered by Orthogonal Matching Pursuit, as this partial converse of Theorem 3.1 shows.

Theorem 3.10 (Exact Recovery Converse for OMP). *Assume that m-term superpositions are unique but that the Exact Recovery Condition (3.1) fails for a given signal $\boldsymbol{x}$ with optimal synthesis matrix $\boldsymbol{\Phi}_{\mathrm{opt}}$. Then there are signals in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$ which Orthogonal Matching Pursuit cannot recover.*

Proof. If the Exact Recovery Condition fails, then

$$\max_{\boldsymbol{\psi}}\left\|\boldsymbol{\Phi}_{\mathrm{opt}}^{+}\boldsymbol{\psi}\right\|_1 \geq 1. \tag{3.14}$$

Now, notice that every signal $\boldsymbol{x}$ which has a representation over the atoms in $\boldsymbol{\Phi}_{\mathrm{opt}}$ yields the same two matrices $\boldsymbol{\Phi}_{\mathrm{opt}}$ and $\boldsymbol{\Psi}_{\mathrm{opt}}$ by the uniqueness of m-term representations. Next, choose $\boldsymbol{y}_{\mathrm{bad}} \in \mathbb{C}^m$ to be a vector for which equality holds in the estimate

$$\frac{\left\|\boldsymbol{\Psi}_{\mathrm{opt}}^{*}(\boldsymbol{\Phi}_{\mathrm{opt}}^{+})^{*}\boldsymbol{y}\right\|_{\infty}}{\|\boldsymbol{y}\|_{\infty}} \leq \left\|\boldsymbol{\Psi}_{\mathrm{opt}}^{*}(\boldsymbol{\Phi}_{\mathrm{opt}}^{+})^{*}\right\|_{\infty,\infty}.$$

Optimal synthesis matrices have full column rank, so $\boldsymbol{\Phi}_{\mathrm{opt}}^{*}$ maps the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$ onto $\mathbb{C}^m$. Therefore, we can find a signal $\boldsymbol{x}_{\mathrm{bad}}$ in the column span of $\boldsymbol{\Phi}_{\mathrm{opt}}$ for which $\boldsymbol{\Phi}_{\mathrm{opt}}^{*}\boldsymbol{x}_{\mathrm{bad}} = \boldsymbol{y}_{\mathrm{bad}}$. Working backward from (3.14) through the proof of the Exact Recovery Theorem, we discover that $\rho(\boldsymbol{x}_{\mathrm{bad}}) \geq 1$. In conclusion, if we run Orthogonal Matching Pursuit with $\boldsymbol{x}_{\mathrm{bad}}$ as input, it chooses a

---

[4]See the article of Elad and Bruckstein [EB02] for a very enlightening graph that delineates the regions of uniqueness and recovery for two-ONB dictionaries.

non-optimal atom in the first step. Since $\boldsymbol{\Phi}_{\text{opt}}$ provides the unique m-term representation of $\boldsymbol{x}_{\text{bad}}$, the initial incorrect selection damns OMP from obtaining the m-term representation of $\boldsymbol{x}_{\text{bad}}$. $\square$

4. Recovering General Signals

The usual goal of sparse approximation is the analysis or compression of natural signals. But the assumption that a signal has an exact, sparse representation must be regarded as Platonic because these signals do not exist in the wild.

Proposition 4.1. *If $m < d$, the collection of signals which have an exact representation using m atoms forms a set of measure zero in $\mathbb{C}^d$.*

Proof. The signals which lie in the span of m distinct atoms form an m-dimensional hyperplane, 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 union of m-dimensional hyperplanes. This union has measure zero in $\mathbb{C}^d$. $\square$

It follows that a generic signal does not have an exact, 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 $(\mathscr{D}, m)$-Exact-Sparse is very different from proving that it succeeds for $(\mathscr{D}, m)$-Sparse. Nevertheless, the analysis in Section 3.1 suggests that Orthogonal Matching Pursuit may be able to recover vectors from the optimal representation even when the signal is not perfectly sparse.

4.1. OMP as an Approximation Algorithm.
Let $\boldsymbol{x}$ be an arbitrary signal, and suppose that $\boldsymbol{a}_{\text{opt}}$ is an optimal m-term approximation to $\boldsymbol{x}$. That is, $\boldsymbol{a}_{\text{opt}}$ is a solution to the minimization (2.1). Note that $\boldsymbol{a}_{\text{opt}}$ may not be unique. We write

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

for an index set Λ_{opt} of size m. Once again, denote by $\boldsymbol{\Phi}_{\text{opt}}$ 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 which is linearly dependent on the others could be replaced by a linearly independent atom to improve the quality of the approximation. Let $\boldsymbol{\Psi}_{\text{opt}}$ be the matrix whose columns are the $(N - m)$ remaining atoms.

Now I formulate a condition under which Orthogonal Matching Pursuit recovers optimal atoms.

Theorem 4.2 (General Recovery). *Assume that $\mu_1(m) < \frac{1}{2}$, and suppose that $\boldsymbol{a}_k$ consists only of atoms from an optimal representation $\boldsymbol{a}_{\text{opt}}$ of the signal $\boldsymbol{x}$. At step $(k+1)$, Orthogonal Matching Pursuit will recover another atom from $\boldsymbol{a}_{\text{opt}}$ whenever*

$$\|\boldsymbol{x} - \boldsymbol{a}_k\|_2 > \sqrt{1 + \frac{m \, (1 - \mu_1(m))}{(1 - 2 \, \mu_1(m))^2}} \, \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2 \, . \tag{4.1}$$

I call (4.1) 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 which optimal approximation appears in the calculation. But the principal corollary of Theorem 4.2 is that OMP is an approximation algorithm for $(\mathscr{D}, m)$-Sparse.

Corollary 4.3. *Assume that $\mu_1(m) < \frac{1}{2}$, and let $\boldsymbol{x}$ be a completely arbitrary signal. Orthogonal Matching Pursuit produces an m-term approximant $\boldsymbol{a}_m$ which satisfies*

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + C(\mathscr{D}, m)} \; \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2 , \tag{4.2}$$

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

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

Proof. At step $(K+1)$, imagine that (4.1) fails. 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. $\square$

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 Voltaire said, "The best is the enemy of the good."

Placing a restriction on the Babel function leads to a simpler statement of the result, which generalizes and improves the work in [GMS03].

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 which satisfy*

$$\|\boldsymbol{x} - \boldsymbol{a}_k\|_2 \leq \sqrt{1 + 6m} \; \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2 . \tag{4.3}$$

The constant here is not small, so it is better to regard this as a qualitative theorem on the performance of OMP. See [GMST03] for another greedy algorithm with a much better constant of approximation. At present, Basis Pursuit offers no approximation guarantees.

Let us return again to the example of Section 2.2.3. 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 Babel function condition demonstrates that the approximation constant is never more than $\sqrt{1 + 6m}$.

Another consequence of the analysis is a corollary for Weak Orthogonal Matching Pursuit.

Corollary 4.5. *Weak Orthogonal Matching Pursuit with parameter α can calculate m-term approximants which satisfy*

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

If, for example, $\mu_1(m) \leq \frac{1}{3}$, then WOMP($\frac{3}{4}$) has an approximation constant no worse than $\sqrt{1 + 24m}$.

4.2. **Proof of the General Recovery Theorem.**

Proof. After k steps, suppose that Orthogonal Matching Pursuit has recovered an approximant $\boldsymbol{a}_k$ which is a linear combination of k atoms listed in Λ_{opt}. The residual is $\boldsymbol{r}_k = \boldsymbol{x} - \boldsymbol{a}_k$, and the condition for recovering another optimal atom is

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

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

$$\begin{aligned}
\rho(\boldsymbol{r}_k) &= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* \boldsymbol{r}_k \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* \boldsymbol{r}_k \right\|_\infty} \\
&= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_k) \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_k) \right\|_\infty} \\
&= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}) + \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}) + \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty} \\
&\leq \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}) \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty} + \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty} \\
&\overset{\mathrm{def}}{=} \rho_{\mathrm{err}}(\boldsymbol{r}_k) + \rho_{\mathrm{opt}}(\boldsymbol{r}_k).
\end{aligned} \tag{4.4}$$

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

To bound $\rho_{\mathrm{opt}}(\boldsymbol{r}_k)$, repeat the arguments of Section 3.3, *mutatis mutandis*. This yields

$$\rho_{\mathrm{opt}}(\boldsymbol{r}_k) \leq \frac{\mu_1(m)}{1 - \mu_1(m-1)} \leq \frac{\mu_1(m)}{1 - \mu_1(m)}. \tag{4.5}$$

Meanwhile, $\rho_{\mathrm{err}}(\boldsymbol{r}_k)$ has the following simple estimate:

$$\begin{aligned}
\rho_{\mathrm{err}}(\boldsymbol{r}_k) &= \frac{\left\| \boldsymbol{\Psi}_{\mathrm{opt}}^* (\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}) \right\|_\infty}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty} \\
&= \frac{\max_{\boldsymbol{\psi}} |\boldsymbol{\psi}^* (\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}})|}{\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_\infty} \\
&\leq \frac{\max_{\boldsymbol{\psi}} \left\| \boldsymbol{\psi} \right\|_2 \left\| \boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}} \right\|_2}{m^{-1/2} \left\| \boldsymbol{\Phi}_{\mathrm{opt}}^* (\boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k) \right\|_2} \\
&\leq \frac{\sqrt{m} \left\| \boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}} \right\|_2}{\sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}}) \left\| \boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k \right\|_2}.
\end{aligned} \tag{4.6}$$

Since $\boldsymbol{\Phi}_{\mathrm{opt}}$ has full column rank, $\sigma_{\min}(\boldsymbol{\Phi}_{\mathrm{opt}})$ is nonzero.

Now, we can develop a concrete condition under which OMP retrieves optimal atoms. In the following calculation, assume that $\mu_1(m) < \frac{1}{2}$. Then combine inequalities (4.4), (4.5) and (4.6), and estimate the singular value with Lemma 2.3. We discover that $\rho(\boldsymbol{r}_k) < 1$ whenever

$$\frac{\sqrt{m} \left\| \boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}} \right\|_2}{\sqrt{1 - \mu_1(m)} \left\| \boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k \right\|_2} + \frac{\mu_1(m)}{1 - \mu_1(m)} < 1.$$

Some algebraic manipulations yield the inequality

$$\left\| \boldsymbol{a}_{\mathrm{opt}} - \boldsymbol{a}_k \right\|_2 > \frac{\sqrt{m\,(1 - \mu_1(m))}}{1 - 2\,\mu_1(m)} \left\| \boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}} \right\|_2.$$

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

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

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

Remark 4.6. The term $\sqrt{m}$ is an unpleasant aspect of (4.6), but it cannot be avoided. 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 $\boldsymbol{b}_{\mathrm{opt}}$ might improve the estimate, but this is a severe restriction. An astute reader will notice that I 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 $(\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}})$. 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 shall retreat from this battle, which should be fought on behalf of a particular application.

5. OMP with Approximate Nearest Neighbors

Gilbert, Muthukrishnan and Strauss have discussed how to use an approximate nearest neighbor (ANN) data structure to develop a fast implementation of Orthogonal Matching Pursuit (ANNOMP) for unstructured dictionaries. Indyk has also suggested this application for ANN data structures. Since the paper [GMS03] already describes the essential details of the technique, I will just say a few words on atom selection and conclude with a new theorem on using ANNOMP with quasi-incoherent dictionaries.

5.1. **Atom Selection.** Let $\widetilde{\boldsymbol{r}}_k$ denote the normalized residual. With an approximate nearest neighbor data structure, we can find an atom $\boldsymbol{\varphi}_{\lambda_k}$ which satisfies

$$\|\widetilde{\boldsymbol{r}}_{k-1} \pm \boldsymbol{\varphi}_{\lambda_k}\|_2^2 \leq (1 + \eta) \min_{\omega} \|\widetilde{\boldsymbol{r}}_{k-1} \pm \boldsymbol{\varphi}_{\omega}\|_2^2 \tag{5.1}$$

for a fixed number $\eta > 0$. Each sign is taken to minimize the respective norm. Rearranging (5.1) yields a guarantee on the inner products.

$$|\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\lambda_k}\rangle| \leq (1 + \eta) \max_{\omega} |\langle \boldsymbol{r}_{k-1}, \boldsymbol{\varphi}_{\omega}\rangle| - \eta \|\boldsymbol{r}_{k-1}\|_2. \tag{5.2}$$

Due to the absolute loss in the last term, this condition does not quite yield a weak greedy algorithm. Unfortunately, a greedy algorithm which uses the selection procedure (5.2) will not generally converge to the signal. It can be shown, however, that the algorithm makes progress until

$$\max_{\omega} |\langle \widetilde{\boldsymbol{r}}_K, \boldsymbol{\varphi}_{\omega}\rangle| \leq \eta.$$

In other words, it yields a residual $\boldsymbol{r}_K$ which is essentially orthogonal to every atom in the dictionary. Which means that no *single* atom can represent a significant part of the signal.

5.2. **Quasi-Incoherent Dictionaries.** For a quasi-incoherent dictionary, we can develop approximation guarantees for ANNOMP which parallel Corollary 4.4, so long as the parameter η is sufficiently small.

Theorem 5.1 (ANNOMP). *Suppose that $\mu_1(m) \leq \frac{1}{3}$, and set $\eta = \frac{1}{5\sqrt{m}}$. Then Orthogonal Matching Pursuit implemented with an approximate nearest neighbor data structure calculates m-term approximants that satisfy*

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + 24m} \|\boldsymbol{x} - \boldsymbol{a}_{\mathrm{opt}}\|_2.$$

Implementing ANNOMP with Indyk's nearest neighbor data structure [Ind00] requires preprocessing time and space $O(N (1/\eta)^{O(d)} \operatorname{polylog}(dN))$. Subsequently, each m-term representation can be calculated in $O(m^2 d + md \operatorname{polylog}(dN))$ time and $O(md)$ additional space, which is quite good considering that we have placed no restrictions on the dictionary beyond quasi-incoherence. A more sophisticated greedy algorithm based on approximate nearest neighbors appears in [GMST03], but no additional approximation guarantees are presently available.

APPENDIX A. PROOF OF THEOREM 3.7

Theorem 3.7. *Suppose that $\mathscr{D}$ consists of J concatenated orthonormal bases with overall coherence μ. Let $\boldsymbol{x}$ be a superposition of p_j atoms from the j-th basis, $j = 1, \ldots, J$. Without loss of generality, assume that $0 < p_1 \le p_2 \le \cdots \le p_J$. Then the Exact Recovery Condition holds whenever*

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

In which case both Orthogonal Matching Pursuit and Basis Pursuit recover the sparse representation.

Proof. Permute the columns of the optimal synthesis matrix so that $\boldsymbol{\Phi}_{\mathrm{opt}} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \boldsymbol{\Phi}_J \end{bmatrix}$, where the p_j columns of submatrix $\boldsymbol{\Phi}_j$ are the atoms from the j-th basis. Suppose that there are a total of m atoms. The goal is to provide a good upper bound for $\left\| \boldsymbol{\Phi}_{\mathrm{opt}}^{+} \boldsymbol{\psi} \right\|_1 = \left\| (\boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{\Phi}_{\mathrm{opt}})^{-1} \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{\psi} \right\|_1$, where $\boldsymbol{\psi}$ is a non-optimal vector. We shall develop this matrix-vector product explicitly under a worst-case assumption on the size of the matrix and vector entries.

The Gram matrix has the block form

$$G \stackrel{\mathrm{def}}{=} \boldsymbol{\Phi}_{\mathrm{opt}}^{*} \boldsymbol{\Phi}_{\mathrm{opt}} = \left[\begin{array}{c|c|c|c} \mathsf{I}_{p_1} & -A_{12} & \cdots & -A_{1J} \\ \hline -A_{21} & \mathsf{I}_{p_2} & \cdots & -A_{2J} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline -A_{J1} & -A_{J2} & \cdots & \mathsf{I}_{p_J} \end{array} \right] \stackrel{\mathrm{def}}{=} \mathsf{I}_m - A,$$

where the entries of A are bounded in magnitude by μ. Then we have the entrywise inequality

$$\left| G^{-1} \right| = \left| \mathsf{I}_m + \sum_{k=1}^{\infty} A^k \right| \le \mathsf{I}_m + \sum_{k=1}^{\infty} |A|^k,$$

where $|A|$ is the entrywise absolute value of the matrix. Therefore, we are at liberty in our estimates to assume that every off-diagonal-block entry of A equals μ. To proceed, creatively rewrite the Gram matrix as $G = (\mathsf{I}_m + \mu B) - (A + \mu B)$, where B is the block matrix

$$B = \left[\begin{array}{c|c|c|c} \mathbf{1}_{p_1} & 0 & \cdots & 0 \\ \hline 0 & \mathbf{1}_{p_2} & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & \mathbf{1}_{p_J} \end{array} \right].$$

We have used $\mathbf{1}$ to denote the matrix with unit entries. By the foregoing, we have the entrywise bound

$$\left| G^{-1} \right| \le \left((\mathsf{I}_m + \mu B) - \mu \mathbf{1}_m \right)^{-1},$$

which yields

$$\left| G^{-1} \right| \le \left(\mathsf{I}_m - \mu \left(\mathsf{I}_m + \mu B \right)^{-1} \mathbf{1}_m \right)^{-1} \left(\mathsf{I}_m + \mu B \right)^{-1}. \tag{A.1}$$

Now, we shall work out the inverses from the right-hand side of (A.1). Using von Neumann series, compute that

$$(\mathsf{I}_m + \mu B)^{-1} = \left[\begin{array}{c|c|c|c} \mathsf{I}_{p_1} - \frac{\mu}{1+\mu p_1} \mathbf{1}_{p_1} & 0 & \cdots & 0 \\ \hline 0 & \mathsf{I}_{p_2} - \frac{\mu}{1+\mu p_2} \mathbf{1}_{p_2} & \cdots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \cdots & \mathsf{I}_{p_J} - \frac{\mu}{1+\mu p_J} \mathbf{1}_{p_J} \end{array} \right]. \tag{A.2}$$

Meanwhile, the series development of the other inverse is

$$\left(I_m - \mu\left(I_m + \mu B\right)^{-1}\mathbf{1}_m\right)^{-1} = I_m + \sum_{k=1}^{\infty}\left(\mu\left(I_m + \mu B\right)^{-1}\mathbf{1}_m\right)^k. \tag{A.3}$$

Next, use (A.2) to calculate the product

$$\mu\left(I_m + \mu B\right)^{-1}\mathbf{1}_m = \begin{bmatrix} \frac{\mu}{1+\mu p_1}\mathbf{1}_{p_1} \\ \hline \frac{\mu}{1+\mu p_2}\mathbf{1}_{p_2} \\ \hline \vdots \\ \hline \frac{\mu}{1+\mu p_J}\mathbf{1}_{p_J} \end{bmatrix} \begin{bmatrix} \mathbf{1}_{p_1}^t & \mathbf{1}_{p_2}^t & \ldots & \mathbf{1}_{p_J}^t \end{bmatrix} \stackrel{\text{def}}{=} \boldsymbol{v}\,\mathbf{1}_m^t.$$

Nota bene that $\mathbf{1}$ indicates the *column vector* with unit entries. On the other hand, we have the inner product

$$\mathbf{1}_m^t\,\boldsymbol{v} = \sum_{j=1}^{J}\frac{\mu p_j}{1 + \mu p_j}.$$

Therefore, the series in (A.3) collapses to

$$\sum_{k=1}^{\infty}\left(\boldsymbol{v}\,\mathbf{1}_m^t\right)^k = \left(\boldsymbol{v}\,\mathbf{1}_m^t\right)\sum_{k=1}^{\infty}\left(\mathbf{1}_m^t\,\boldsymbol{v}\right)^{k-1} = \frac{1}{1 - \sum_{j=1}^{J}\frac{\mu p_j}{1+\mu p_j}}\,\boldsymbol{v}\,\mathbf{1}_m^t. \tag{A.4}$$

In consequence of (A.3) and (A.4), the inverse of the Gram matrix satisfies

$$\left|G^{-1}\right| \le \left(I_m + \frac{1}{1 - \sum_{j=1}^{J}\frac{\mu p_j}{1+\mu p_j}}\,\boldsymbol{v}\,\mathbf{1}_m^t\right)\left(I_m + \mu B\right)^{-1}. \tag{A.5}$$

Now, assume that the vector $\boldsymbol{\psi}$ is drawn from basis number Z. So

$$\left|\boldsymbol{\Phi}_{\text{opt}}^{*}\boldsymbol{\psi}\right| \le \begin{bmatrix} \mu\,\mathbf{1}_{p_1}^t & \ldots & \mathbf{0}_{p_Z}^t & \ldots & \mu\,\mathbf{1}_{p_J}^t \end{bmatrix}^t. \tag{A.6}$$

At last, we are prepared to calculate the product we care about. First,

$$\left|\boldsymbol{\Phi}_{\text{opt}}^{+}\boldsymbol{\psi}\right| = \left|\left(\boldsymbol{\Phi}_{\text{opt}}^{*}\boldsymbol{\Phi}_{\text{opt}}\right)^{-1}\boldsymbol{\Phi}_{\text{opt}}^{*}\boldsymbol{\psi}\right| \le \left|G^{-1}\right|\left|\boldsymbol{\Phi}_{\text{opt}}^{*}\boldsymbol{\psi}\right|. \tag{A.7}$$

We shall work from right to left. Equations (A.2) and (A.6) imply

$$\left(I_m + \mu B\right)^{-1}\left|\boldsymbol{\Phi}_{\text{opt}}^{*}\boldsymbol{\psi}\right| \le \begin{bmatrix} \frac{\mu}{1+\mu p_1}\mathbf{1}_{p_1}^t & \ldots & \mathbf{0}_{p_Z}^t & \ldots & \frac{\mu}{1+\mu p_J}\mathbf{1}_{p_J}^t \end{bmatrix}^t. \tag{A.8}$$

Introducing (A.5) and (A.8) into (A.7) yields

$$\left|\boldsymbol{\Phi}_{\text{opt}}^{+}\boldsymbol{\psi}\right| \le \begin{bmatrix} \frac{\mu}{1+\mu p_1}\mathbf{1}_{p_1} \\ \hline \vdots \\ \hline \mathbf{0}_{p_Z} \\ \hline \vdots \\ \hline \frac{\mu}{1+\mu p_J}\mathbf{1}_{p_J} \end{bmatrix} + \frac{\sum_{j\ne Z}\frac{\mu p_j}{1+\mu p_j}}{1 - \sum_{j=1}^{J}\frac{\mu p_j}{1+\mu p_j}} \begin{bmatrix} \frac{\mu}{1+\mu p_1}\mathbf{1}_{p_1} \\ \hline \vdots \\ \hline \frac{\mu}{1+\mu p_Z}\mathbf{1}_{p_Z} \\ \hline \vdots \\ \hline \frac{\mu}{1+\mu p_J}\mathbf{1}_{p_J} \end{bmatrix}. \tag{A.9}$$

Finally, apply the ℓ_1 norm to inequality (A.9) to reach

$$\left\|\boldsymbol{\Phi}_{\text{opt}}^{+}\boldsymbol{\psi}\right\|_1 \le \frac{\sum_{j\ne Z}\frac{\mu p_j}{1+\mu p_j}}{1 - \sum_{j=1}^{J}\frac{\mu p_j}{1+\mu p_j}}. \tag{A.10}$$

The bound (A.10) is weakest when $Z = 1$. And the theorem follows. $\square$

References

[CBL89]   S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *Intl. J. Control*, 50(5):1873–1896, 1989.

[CDS99]   S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comp.*, 20(1):33–61, 1999. Electronic.

[CW92]   R. R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best-basis selection. *IEEE Trans. Inform. Th.*, 1992.

[DE02]   D. L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. Draft, Dec. 2002.

[DeV98]   R. A. DeVore. Nonlinear approximation. *Acta Numerica*, pages 51–150, 1998.

[DH01]   D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Th.*, 47:2845–2862, Nov. 2001.

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

[DMZ94]   G. Davis, S. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Eng.*, July 1994.

[EB02]   M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Th.*, 48(9):2558–2567, 2002.

[FN]   A. Feuer and A. Nemirovsky. On sparse representations in pairs of bases. Accepted to *IEEE Trans. IT*, Nov. 2002.

[FS81]   J. H. Friedman and W. Stuetzle. Projection Pursuit Regressions. *J. Amer. Statist. Soc.*, 76:817–823, 1981.

[GMS03]   A. C. Gilbert, M. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *The 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, Jan. 2003.

[GMST03]   A. C. Gilbert, S. Muthukrishnan, M. J. Strauss, and J. A. Tropp. Improved sparse approximation over quasi-incoherent dictionaries. In submission, 2003.

[GN02]   R. Gribonval and M. Nielsen. Sparse representations in unions of bases. Technical Report 1499, Institut de Recherche en Informatique et Systèmes Aléatoires, Nov. 2002.

[Grö01]   K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston, 2001.

[HJ85]   R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ. Press, 1985.

[HSP02]   R. Heath, T. Strohmer, and A. J. Paulraj. On quasi-orthogonal signatures for CDMA systems. In *Proceedings of the 2002 Allerton Conference on Communication, Control and Computers*, 2002.

[Ind00]   P. Indyk. *High-Dimensional Computational Geometry*. PhD thesis, Stanford, 2000.

[Jon87]   L. K. Jones. On a conjecture of Huber concerning the convergence of Projection Pursuit Regression. *Ann. Stat.*, 15(2):880–882, 1987.

[Kre89]   E. Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1989.

[MZ93]   S. Mallat and Z. Zhang. Matching Pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

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

[QC94]   S. Qian and D. Chen. Signal representation using adaptive normalized Gaussian functions. *Signal Process.*, 36:329–355, 1994.

[SBT00]   S. Sardy, A. G. Bruce, and P. Tseng. Block Coordinate Relaxation methods for nonparametric wavelet denoising. *Comp. and Graph. Stat.*, 9(2), 2000.

[SH02]   T. Strohmer and R. Heath. Grassmannian frames with applications to coding and communication. In submission, 2002.

[ST03]   M. Sustik and J. A. Tropp. Existence of real Grassmannian frames. In preparation, 2003.

[Tem02]   V. Temlyakov. Nonlinear methods of approximation. *Foundations of Comp. Math.*, July 2002.

[Vil97]   L. F. Villemoes. Best approximation with Walsh atoms. *Constr. Approx.*, 13:329–355, 1997.

Texas Institute for Computational Engineering and Sciences (TICES), The University of Texas at Austin (C0200), Austin, TX 78712

E-mail address: jtropp@ticam.utexas.edu