IMPROVED SPARSE APPROXIMATION OVER QUASI-INCOHERENT DICTIONARIES

J. A. Tropp[*] *A. C. Gilbert*[†] *S. Muthukrishnan*[‡] *M. J. Strauss*[§]

ABSTRACT

This paper discusses a new greedy algorithm for solving the sparse approximation problem over quasi-incoherent dictionaries. These dictionaries consist of waveforms that are uncorrelated "on average," and they provide a natural generalization of incoherent dictionaries. The algorithm provides strong guarantees on the quality of the approximations it produces, unlike most other methods for sparse approximation. Moreover, very efficient implementations are possible via approximate nearest-neighbor data structures.

1. INTRODUCTION

Sparse approximation is the problem of finding a concise representation of a given signal as a linear combination of a few elementary signals chosen from a rich collection. It has shown empirical promise in image processing tasks such as feature extraction, because the approximation cannot succeed unless it discovers structure latent in the image. For example, Starck, Donoho and Candès have used sparse approximation to extract features from noisy astronomical photographs and volumetric data [1]. Nevertheless, it has been difficult to establish that proposed algorithms actually solve the sparse approximation problem. This paper makes another step in that direction by describing a greedy algorithm that computes solutions with provable quality guarantees.

A *dictionary* $\mathscr{D}$ for the signal space $\mathbb{R}^d$ is a collection of vectors that spans the entire space. The vectors are called *atoms*, and we write them as φ_λ. The index λ may parameterize the time/scale or time/frequency localization of each atom, or it may be a label without any additional meaning. The number of atoms is often much larger than the signal dimension.

The *sparse approximation problem* with respect to $\mathscr{D}$ is to compute a good representation of each input signal as a short linear combination of atoms. Specifically, for an arbitrary signal $\boldsymbol{x}$, we search for an m-term superposition

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

which minimizes $\|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2$. We must determine both the optimal vectors, m atoms whose indices are listed by Λ_{opt}, as well as the optimal coefficients b_λ.

If $\mathscr{D}$ is an orthonormal basis, it is computationally easy to find $\boldsymbol{a}_{\text{opt}}$. For the indices Λ_{opt}, simply take m atoms with the largest inner products $|\langle \boldsymbol{x}, \varphi_\lambda \rangle|$ and form

$$\boldsymbol{a}_{\text{opt}} = \sum_{\Lambda_{\text{opt}}} \langle \boldsymbol{x}, \varphi_\lambda \rangle \; \varphi_\lambda .$$

Unfortunately, it can be difficult or impossible to choose an appropriate orthonormal basis for a given situation. For example, if the signals contain both harmonic and impulsive components, a single orthonormal basis will not represent them both efficiently. We have much more freedom with a redundant dictionary, since it may include a rich collection of waveforms which can provide concise representations of many different structures.

The price that we pay for additional flexibility is an increased cost to determine these concise representations. For general redundant dictionaries, it is computationally infeasible to search all possible m-term representations. In fact, if $\mathscr{D}$ is an arbitrary dictionary, finding the best m-term representation of an arbitrary signal is NP-hard [2]. There are algorithms with provable approximation guarantees for specific dictionaries, e.g. Villemoes' algorithm for Haar wavelet packets [3]. There are also some well-known heuristics, such as Matching Pursuit (MP) [4], Orthogonal Matching Pursuit (OMP) [5] and m-fold Matching Pursuit [6]. Several other methods rely on the Basis Pursuit paradigm, which advocates minimizing the ℓ_1 norm of the coefficients in the representation instead of minimizing the sparsity directly [7].

Some theoretical progress has already been made for dictionaries with low *coherence*. The coherence parameter μ equals the maximal inner product between two distinct atoms. For example, the union of spikes and sines is a dictionary with $\mu = \sqrt{2/d}$. The authors in [6] have presented an efficient two-stage algorithm for the approximate representation of any signal over a sufficiently incoherent

[*]Inst. for Comp. Eng. and Sci. (ICES), The University of Texas at Austin, Austin, TX 78712, jtropp@ices.utexas.edu.

[†]AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, agilbert@research.att.com.

[‡]AT&T Labs-Research & Rutgers University, 180 Park Avenue, Florham Park, NJ 07932, muthu@research.att.com. Supported in part by NSF CCR 00-87022 and NSF ITR 0220280.

[§]AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, mstrauss@research.att.com.

0-7803-7750-8/03/$17.00 ©2003 IEEE

dictionary. This is the first known algorithm which provably approximates the solution to the sparse problem for any class of general dictionaries. In addition, this algorithm is highly efficient. For a suitably incoherent dictionary, it is also known that Basis Pursuit can resolve the subclass of signals which have an *exact* sparse representation [8].

This article offers a number of improvements to [6]. Specifically, we present a modified version of the algorithm in [6], which calculates significantly more accurate sparse representations for incoherent dictionaries and also applies to a much larger class of redundant dictionaries. Unlike an incoherent dictionary where all the inner products are small, the dictionaries we consider only need to have small inner products "on average." In addition, our analysis is simpler. Of course, the new algorithm can be implemented just as efficiently as the ones in [6].

2. ALGORITHM AND ANALYSIS

2.1. Overall results

For an incoherent dictionary, we have

Theorem 1 *Fix a dictionary $\mathscr{D}$ with coherence μ. We seek an m-term representation of an arbitrary signal $\boldsymbol{x}$, where $m < \frac{1}{2}\mu^{-1}$. There is an algorithm that produces an m-term representation $\boldsymbol{a}_m$ for $\boldsymbol{x}$ with error*

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

In comparison, the algorithm of [6] requires that $m < \frac{1}{32}\mu^{-1}$ and produces approximations with error

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + 2064\mu m^2}\, \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2 .$$

When $m \leq \mu^{-1/2}$, the resulting constant of approximation is about 46. Meanwhile, the algorithm described here produces approximations with error

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq 3\, \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2$$

so long as $4 \leq m \leq \mu^{-1/2}$.

Theorem 1 is a special case of a result for general dictionaries. To state the full theorem, we need to borrow a definition from [9]. Let $\boldsymbol{\Phi}$ be a matrix whose columns are the atoms of $\mathscr{D}$. Then the Gram matrix $G \stackrel{\text{def}}{=} \boldsymbol{\Phi}^* \boldsymbol{\Phi}$ contains all the inner products between atoms[1]. Let $|G|$ denote its entrywise absolute value. Now, we define the *Babel function* $\boldsymbol{\mu}_1(m)$ of the dictionary to be the maximum sum of any m (nondiagonal) elements from a single row of $|G|$. In other words, the Babel function quantifies the maximum total coherence between a fixed atom and a collection of m other atoms[2]. The Babel function is a more subtle way of describing the dictionary than the coherence, since coherence only reflects the largest inner product. Clearly,

$$\boldsymbol{\mu}_1(m) \leq \mu m. \tag{1}$$

That is, the cumulative coherence always dominates the Babel function. When the Babel function grows slowly, we say informally that the dictionary is *quasi-incoherent*.

Theorem 2 *So long as $\boldsymbol{\mu}_1(m) < \frac{1}{2}$, our algorithm produces an m-term approximation $\boldsymbol{a}_m$ which satisfies*

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2 \leq \sqrt{1 + \frac{2m\, \boldsymbol{\mu}_1(m)}{(1 - 2\boldsymbol{\mu}_1(m))^2}}\, \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2 ,$$

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

Obviously, Theorem 1 follows directly from Theorem 2 by application of the bound (1).

We can easily construct a dictionary for which we need the more general theorem. Let each atom be a linear combination of two impulses:

$$\boldsymbol{\varphi}_k = \tfrac{\sqrt{35}}{6}\, \boldsymbol{\delta}_k + \tfrac{1}{6}\, \boldsymbol{\delta}_{k+1} \qquad \text{for } k = 1, \ldots, d.$$

Then the coherence $\mu = \frac{\sqrt{35}}{36}$, which means that Theorem 1 applies only when $m \leq 3$. Meanwhile, the Babel function $\boldsymbol{\mu}_1(m) = \frac{\sqrt{35}}{18} < \frac{1}{3}$ for every $m \geq 2$. Therefore, the general theorem shows that approximation succeeds for any m, and the error bound is

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

Another consequence of Theorem 2 is that the algorithm can recover any signal which has an exact m-term representation, so long as $\boldsymbol{\mu}_1(m) < \frac{1}{2}$. In fact, the analysis of [9] shows that Orthogonal Matching Pursuit alone can accomplish this task. Donoho and Elad have proven that Basis Pursuit can recover exactly sparse signals under an identical condition [10]. But they have not offered an approximation guarantee for general signals, and the algorithms associated with the Basis Pursuit paradigm are typically very slow. The algorithm described here is significantly faster.

2.2. A Structural Lemma

An important ingredient in the analysis is our generalization of Parseval's Theorem to an arbitrary dictionary. A similar lemma appears implicitly in the analysis of [6].

[1] Here and elsewhere, * denotes the conjugate transpose.

[2] The subscript in the notation serves to distinguish the Babel function from the coherence parameter and to remind us that it represents the ℓ_1 norm, i.e. an absolute sum.

Lemma 3 (Approximate Parseval) *Let Λ be a collection of m atoms. Let P_Λ be the orthogonal projector onto the span of the atoms listed by Λ. For every signal $\boldsymbol{x}$, we have*

$$\frac{\sum_\Lambda |\langle \boldsymbol{x}, \varphi_\lambda \rangle|^2}{1 + \mu_1(m)} \leq \|P_\Lambda \boldsymbol{x}\|_2^2 \leq \frac{\sum_\Lambda |\langle \boldsymbol{x}, \varphi_\lambda \rangle|^2}{1 - \mu_1(m)}.$$

Proof. Define the matrix Φ_Λ whose columns are the m atoms listed in Λ. Then the orthogonal projector onto the span of the atoms in Λ can be written as

$$P_\Lambda = (\Phi_\Lambda^* \Phi_\Lambda)^{-1/2} \Phi_\Lambda^*.$$

Denote the smallest and largest singular values of Φ_Λ by $\sigma_{\min}$ and $\sigma_{\max}$. It follows that the smallest and largest eigenvalues of $(\Phi_\Lambda^* \Phi_\Lambda)^{-1/2}$ are respectively equal to $1/\sigma_{\max}$ and $1/\sigma_{\min}$. Therefore,

$$\frac{\|\Phi_\Lambda^* \boldsymbol{x}\|_2^2}{\sigma_{\max}^2} \leq \|P_\Lambda \boldsymbol{x}\|_2^2 \leq \frac{\|\Phi_\Lambda^* \boldsymbol{x}\|_2^2}{\sigma_{\min}^2}.$$

Since the squared singular values of Φ_Λ are equal to the eigenvalues of $G_\Lambda \stackrel{\text{def}}{=} \Phi_\Lambda^* \Phi_\Lambda$, we can estimate $\sigma_{\min}$ and $\sigma_{\max}$ by examining this Gram matrix. The Geršgorin Disc Theorem [11] states that every eigenvalue of a square matrix A lies in one of the m discs

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

Since G_Λ is a principal submatrix of the dictionary Gram matrix G, we can use the Babel function to bound the sum:

$$\sum_{j \neq k} |(G_\Lambda)_{jk}| \leq \mu_1(m).$$

All atoms have unit norm, so the diagonal of G_Λ is identically one. We may conclude that

$$\sigma_{\min}^2 \geq 1 - \mu_1(m) \qquad \text{and}$$
$$\sigma_{\max}^2 \leq 1 + \mu_1(m).$$

Finish by writing $\|\Phi_\Lambda^* \boldsymbol{x}\|_2^2 = \sum_\Lambda |\langle \boldsymbol{x}, \varphi_\lambda \rangle|^2$. □

The Babel function of an orthonormal basis is zero. In this case, Lemma 3 reduces to Parseval's Theorem.

2.3. Analysis of two-phase greedy pursuit

The overall algorithm is a two-phase greedy pursuit. First, we initialize $\boldsymbol{a}_0 = \boldsymbol{0}$ and perform Orthogonal Matching Pursuit until we reach a K-term representation $\boldsymbol{a}_K$ with a reasonable error guarantee. (The optimal error is necessary to determine K. For now, assume an oracle provides it. Section 2.4 discusses how to avoid the trip to Delphi.) In the second stage, we use $(m - K)$-fold Matching Pursuit to acquire the remaining atoms. The algorithm returns the best approximation to the signal over the m chosen atoms. This procedure is different from the one given in [6], where the algorithm returns the sum of the best approximation of the signal over the first K atoms and the best approximation of the K-term residual over the last $(m - K)$ atoms.

Now, we sketch the analysis. We require the following.

Theorem 4 *[9] For any signal $\boldsymbol{x}$, Orthogonal Matching Pursuit can calculate an approximant $\boldsymbol{a}_K$ that consists of $0 \leq K \leq m$ atoms from the optimal m-term approximant $\boldsymbol{a}_{\text{opt}}$ and that satisfies the error bound*

$$\|\boldsymbol{x} - \boldsymbol{a}_K\|_2^2 \leq \left[1 + \frac{m\,(1 - \mu_1(m))}{(1 - 2\mu_1(m))^2} \right] \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2. \quad (2)$$

After the first phase is complete, we have selected K optimal atoms, Λ_K, and the K-term approximant satisfies the error bound (2). The second phase chooses $\widehat{K} \stackrel{\text{def}}{=} (m - K)$ atoms that have the largest inner products with the residual. More precisely, we find $\widehat{K}$ indices, $\Lambda_{\widehat{K}}$, to maximize

$$\sum_{\Lambda_{\widehat{K}}} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|.$$

We may assume that the atoms chosen in the second phase are distinct from those chosen in the first phase because the residual $(\boldsymbol{x} - \boldsymbol{a}_K)$ is orthogonal to each atom that participates in the approximant $\boldsymbol{a}_K$. The m atoms we have selected are indexed by $\Lambda_m \stackrel{\text{def}}{=} \Lambda_K \cup \Lambda_{\widehat{K}}$, and

$$\sum_{\Lambda_m} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2 \geq \sum_{\Lambda_{\text{opt}}} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2.$$

Finally, the algorithm returns $\boldsymbol{a}_m$, which is the best approximation to the signal using the atoms in Λ_m.

Theorem 5 *The m-term approximation produced by this two-phase algorithm satisfies*

$$\|\boldsymbol{x} - \boldsymbol{a}_m\|_2^2 \leq \left[1 + \frac{2m\,\mu_1(m)}{(1 - 2\mu_1(m))^2} \right] \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2.$$

Proof. The difference between the actual error and the optimal error is

$$\begin{aligned} &\|\boldsymbol{x} - \boldsymbol{a}_m\|_2^2 - \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2 \\ &\quad = \|\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_K\|_2^2 - \|\boldsymbol{a}_m - \boldsymbol{a}_K\|_2^2 \\ &\quad = \|P_{\text{opt}}(\boldsymbol{x} - \boldsymbol{a}_K)\|_2^2 - \|P_m(\boldsymbol{x} - \boldsymbol{a}_K)\|_2^2, \end{aligned}$$

where P_{opt} is the orthogonal projector onto the span of the optimal atoms and P_m is the orthogonal projector onto the span of the m chosen atoms. Applying Lemma 3,

$$\begin{aligned} &\|\boldsymbol{x} - \boldsymbol{a}_m\|_2^2 - \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2 \\ &\quad \leq \frac{\sum_{\Lambda_{\text{opt}}} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2}{1 - \mu_1(m)} - \frac{\sum_{\Lambda_m} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2}{1 + \mu_1(m)} \\ &\quad \leq \frac{\sum_{\Lambda_{\text{opt}}} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2}{1 - \mu_1(m)} - \frac{\sum_{\Lambda_{\text{opt}}} |\langle \boldsymbol{x} - \boldsymbol{a}_K, \varphi_\lambda \rangle|^2}{1 + \mu_1(m)}, \end{aligned}$$

since the atoms in Λ_{opt} carry less energy than those in Λ_m. Applying Lemma 3 again,

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{a}_m\|_2^2 - \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2 &\leq \left[\frac{1+\mu_1(m)}{1-\mu_1(m)} - 1\right] \|P_{\text{opt}}(\boldsymbol{x} - \boldsymbol{a}_K)\|_2^2 \\ &= \frac{2\mu_1(m)}{1-\mu_1(m)} \|\boldsymbol{a}_{\text{opt}} - \boldsymbol{a}_K\|_2^2 \\ &= \frac{2\mu_1(m)}{1-\mu_1(m)} \left[\|\boldsymbol{x} - \boldsymbol{a}_K\|_2^2 - \|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2\right].\end{aligned}$$

The first theorem provides a bound on $\|\boldsymbol{x} - \boldsymbol{a}_K\|_2^2$ in terms of $\|\boldsymbol{x} - \boldsymbol{a}_{\text{opt}}\|_2^2$. Combining the two estimates and rearranging, we reach the stated result. □

2.4. Implementation

Implementing the algorithm which we have described requires foreknowledge of the optimal error. There are two ways to escape the need for omniscience. For the first option, simply execute the algorithm $(m+1)$ times, switching from the first phase to the second at each $K = 0, \ldots, m$. Then select the best of the representations. A second option is to guess the optimal error. This can be accomplished by running the algorithm with guesses taken from a geometric progression ranging from ε, the machine precision, to $\|\boldsymbol{x}\|_2$. This requires only $\log_2 \|\boldsymbol{x}\|_2 / \varepsilon$ attempts, and we may use the best of the representations. Both techniques are embarrassingly parallel, although efficient serial versions are also possible.

Both phases of the algorithm require the determination of maximum inner products. At each step, the Orthogonal Matching Pursuit phase chooses an atom with maximal inner product against the residual. The second phase can be implemented by selecting a maximal inner product, removing that atom from the dictionary and iterating. Therefore, in both stages, we use a data structure that preprocesses the dictionary and supports two queries: return an atom whose absolute inner product with the residual is maximal and delete an atom. An identity for unit vectors states that

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 1 - \tfrac{1}{2} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 .$$

To find the maximum absolute inner product between a signal $\boldsymbol{x}$ and the dictionary, we can normalize the signal as $\widetilde{\boldsymbol{x}}$ and solve

$$\min_{\lambda} \|\boldsymbol{\varphi}_\lambda \pm \widetilde{\boldsymbol{x}}\|_2^2 .$$

This minimization can be performed approximately using a nearest-neighbor data structure for vectors under the Euclidean metric [12]. Building the data structure requires time and space $\text{poly}(|\mathscr{D}|/\eta)$, where η is the precision required in the approximation. But each query costs only $d + \text{polylog}(|\mathscr{D}|/\eta)$ units of time[3]. This query does not necessarily return a vector with the largest inner product, but it always returns a vector that is nearly as good. The analysis of our algorithm changes slightly if we use a nearest-neighbor data structure to estimate maximum inner products. The details are somewhat technical, so we will relegate them to a longer version of this article.

3. REFERENCES

[1] J. L. Starck, D. L. Donoho, and E. J. Candès, "Astronomical image representation by the Curvelet Transform," *Astronomy and Astrophysics*, 2002, Accepted.

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

[3] L. F. Villemoes, "Best approximation with walsh atoms," *Constr. Approx.*, vol. 13, pp. 329–355, 1997.

[4] S. Mallat and Z. Zhang, "Matching pursuits with time frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

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

[6] A. C. Gilbert, S. 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.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999, electronic.

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

[9] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," ICES Report 0304, The University of Texas at Austin, 2003.

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

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

[12] P. Indyk, *High-Dimensional Computational Geometry*, Ph.D. thesis, Stanford, 2000.

[3] Saying that a function is $\text{poly}(t)$ means that it lies in $O(t^\alpha)$ for some $\alpha > 0$. Similarly, each function in $\text{polylog}(t)$ lies in $O(\log^\alpha t)$ for some $\alpha > 0$.