Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise

David L. Donoho, *Member, IEEE*, Michael Elad, and Vladimir N. Temlyakov

***Abstract*—Overcomplete representations are attracting interest in signal processing theory, particularly due to their potential to generate sparse representations of signals. However, in general, the problem of finding sparse representations must be unstable in the presence of noise. This paper establishes the possibility of stable recovery under a combination of sufficient sparsity and favorable structure of the overcomplete system. Considering an ideal underlying signal that has a sufficiently sparse representation, it is assumed that only a noisy version of it can be observed. Assuming further that the overcomplete system is *incoherent*, it is shown that the optimally sparse approximation to the noisy data differs from the optimally sparse decomposition of the ideal noiseless signal by at most a constant multiple of the noise level. As this optimal-sparsity method requires heavy (combinatorial) computational effort, approximation algorithms are considered. It is shown that similar stability is also available using the basis and the matching pursuit algorithms. Furthermore, it is shown that these methods result in sparse approximation of the noisy data that contains only terms also appearing in the unique sparsest representation of the ideal noiseless sparse signal.**

***Index Terms*—Basis pursuit, greedy approximation, incoherent dictionary, Kruskal rank, matching pursuit, overcomplete representation, sparse representation, stability, stepwise regression, superresolution.**

I. Introduction

A. Overcomplete Representation

RESEARCHERS spanning a diverse range of viewpoints have recently advocated the use of *overcomplete* signal representations [27], [30], [1], [5], [4], [33], [37], [35]. Generally speaking, they suppose we have a signal vector $y \in R^n$, and a collection of vectors $\phi_i \in R^n$, $i = 1, \ldots, m$, with $m > n$ such vectors, so that the collection forms "more than a basis"; since [27] such collections are usually called *dictionaries*, and their elements are called *atoms*. We want a representation for our signal $y = \sum_i \alpha_i \phi_i$ as a linear combination of atoms in this dictionary.

Such representations differ from the more traditional basis representation because they offer a wider range of generating elements; potentially, this wider range allows more flexibility in signal representation, and more effectiveness at tasks like signal extraction and data compression. Proposals for overcomplete representations have included multiscale Gabor functions [27], [30], systems defined by algebraic codes [33], amalgams of wavelets and sinusoids [3], [4], [14], libraries of windowed cosines with a range of different widths and locations [5], [42], multiscale windowed ridgelets [32], systems generated at random [11], and amalgams of wavelets and linelike generating elements [22].

A number of interesting arguments, both heuristic and theoretical, have been advanced to support the benefits of overcompleteness; in theoretical neuroscience it has been argued that overcomplete representations are probably necessary for use in biological settings in the mammalian visual system [28]; in approximation theory, there are persuasive examples where approximation from overcomplete systems outperforms any known basis [2]; in signal processing, it has been reported that decomposition into separate transforms gives improved compression [1], [8] and improved equalization [6]; and in image processing, it has been shown that one can separate images into disjoint signal types using such decompositions [31], [32], [22].

At the same time, there is an apparent obstacle to overcomplete representations, based on elementary linear algebra. We can think of the atoms in our dictionary as columns in a matrix $\mathbf{\Phi}$, so that $\mathbf{\Phi}$ is n by m and $m > n$. A representation of $y \in R^n$ can be thought of as a vector $\alpha \in R^m$ satisfying $y = \mathbf{\Phi}\alpha$. However, linear algebra tells us that because $m > n$, the problem of representation is underdetermined. Hence, as is widely taught in elementary courses, there is no unique solution to the representation problem, and far more disturbingly, if the data are even slightly inaccurate, some familiar algorithms will be staggeringly unstable. That this can be a real issue was shown by Wohlberg [43] who considered a dictionary of sinusoids with frequencies spaced finer than the usual discrete Fourier frequencies, and documented the extreme ill-posedness that can result.

In this paper, we consider the impact of sparsity constraints on this situation, and study algorithms which can in certain circumstances generate sparse representations in an overcomplete dictionary. We derive rigorous bounds showing that, when the dictionary $\mathbf{\Phi}$ has a property of *mutual incoherence* (defined below), and when it offers a sufficiently sparse representation for the ideal noiseless signal, the algorithms are locally stable, i.e., under addition of small amounts of noise, the algorithms recover the ideal sparse representation with an error that grows at most proportionally to the noise level. Some of the algorithms are even globally stable, i.e., they recover the ideal noiseless reconstruction with an error at worst proportional to noise level

Manuscript received February 29, 2004; revised September 12, 2005. This work was supported in part by the National Science Foundation under Grants DMS 00-77261, 01-40698 (FRG), 02-00187, ANI-008584 (ITR) and by DARPA ACMP, and ONR-MURI

D. L. Donoho is with the Department of Statistics, Stanford University, Stanford, CA 94305-9025 USA (e-mail: donoho@stanford.edu).

M. Elad is with the Department of Computer Science, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il).

V. N. Temlyakov is with the Department of Mathematics, University of South Carolina, Columbia, SC 29208 USA (e-mail: temlyak@math.sc.edu).

Communicated by G. Battail, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2005.860430

even under the addition of arbitrary amounts of noise. Under sufficient sparsity the constants of proportionality are very reasonable.

In short, we show that, although the problem of recovering the underlying overcomplete representation is admittedly very ill-posed in general, when the underlying representation is sparse, and the dictionary is incoherent, the ill-posedness can disappear.

B. Sparse Representation

To fix ideas, consider the problem of finding the sparsest representation possible in an overcomplete dictionary $\mathbf{\Phi}$. As a measure of sparsity of a vector α, we take the so-called ℓ^0 norm $\|\alpha\|_0$, which is simply the number of nonzero elements in α. The sparsest representation is then the solution to the optimization problem

$$(P_0): \min_{\alpha} \|\alpha\|_0 \text{ subject to } y = \mathbf{\Phi}\alpha. \tag{1.1}$$

As stated, this seems to be a general combinatorial optimization problem, requiring that one enumerate all possible k-element collections of columns of $\mathbf{\Phi}$, for $k = 1, 2, \ldots, n$, looking for the smallest collection permitting representation of the signal. Such an algorithm would cost at least $O(2^m)$ flops to carry out in general, and at least $O(m^k)$ even when a sparse k-element representation existed. We therefore turn to approximations/relaxations of (P_0).

Orthogonal Greedy Algorithm. One heuristic approach builds up k-element approximate representations a step at a time, adding to an existing $(k-1)$-element approximation a new term chosen in a greedy fashion to minimize the resulting ℓ^2 error (over all possible choices of the single additional term). When stopped after $N \ll m$ stages, one gets a sparse approximate representation. In more detail, the procedure starts from an initial residual $r^{(0)} = y$ and a current decomposition $\hat{y}^0 = 0$; then for $k = 1, \ldots,$ it augments the decomposition $\hat{y}^{(k-1)} \rightarrow \hat{y}^{(k)}$ and updates the residual $\hat{r}^{(k-1)} \rightarrow \hat{r}^{(k)}$ stepwise, always maintaining $y = \hat{y}^{(k)} + \hat{r}^{(k)}$. We suppose that the dictionary has normalized atoms, so that each $\|\phi_i\|_2 = 1$. At the kth stage, the algorithm selects an atom to be added to the decomposition based on correlation with the current residual

$$i_k = \text{argmax}_{1 \le i \le m} \left| \left\langle r^{(k-1)}, \phi_i \right\rangle \right|;$$

it builds a decomposition consisting of the atoms selected through that stage

$$\hat{y}^{(k)} = \sum_{l=1}^{k} a_{i_l}^k \phi_{i_l} \tag{1.2}$$

where the coefficients $(a_{i_l}^k)$ are fitted by least squares to minimize $\|y - \hat{y}^{(k)}\|^2$; and it subtracts this model from y, obtaining a new residual

$$r^{(k)} = y - \hat{y}^{(k)}$$

which can be input to the next stage of the algorithm. At that point, the basic iteration is repeated. The algorithm can be stopped when the residual norm is below some predetermined threshold, or based on the number of atoms used.

In the setting of statistical modeling, greedy stepwise least squares is called *forward stepwise regression*, and has been widely practiced since the 1960s [7], [20]. When used in the signal processing setting, this goes by the name of *matching pursuit* (MP) [27]; actually we have described a variant called orthogonal matching pursuit (OMP) [29]. Following [9], we call this the orthogonal greedy algorithm (OGA).

Convex Relaxation. A more formal approach convexifies (P_0) by replacing the ℓ^0-norm with an ℓ^1-norm

$$(P_1): \min_{\alpha} \|\alpha\|_1 \quad \text{subject to } y = \mathbf{\Phi}\alpha. \tag{1.3}$$

This can be cast as a linear programming (LP) problem, for which solutions are available even in large-scale problems, owing to modern interior-point linear programming methods. This approach to overcomplete signal representation was called basis-pursuit (BP) in [4], which observed that it sometimes gave highly sparse solutions to problems known to have such sparse solutions, and showed that it could, in specific cases, outperform the greedy pursuit approach in generating sparse solutions.

Formal Justification. The key point about both OGA and BP is that they are much more practical than the direct solution of (P_0). Perhaps surprisingly, these approaches can, with certain conditions, correctly solve (P_0). Thus, practical methods can solve problems that otherwise on the surface seem computationally intractable. Previous work [9], [17], [39], [38], [15], [13], [14], [11], [19] established that both OGA and BP approaches can be successful for signals having sparse representations; under appropriate conditions on $\mathbf{\Phi}$ and y, these algorithms produce the globally optimal solution of (P_0). The concept of *mutual coherence* of the dictionary $\mathbf{\Phi}$ plays a major role in these results. It is defined, assuming that the columns of $\mathbf{\Phi}$ are normalized to unit ℓ^2-norm, in terms of the Gram matrix $\boldsymbol{G} = \mathbf{\Phi}^T\mathbf{\Phi}$. With $G(k, j)$ denoting entries of this matrix, the mutual coherence is

$$M = M(\mathbf{\Phi}) = \max_{1 \le k, j \le m, k \ne j} |G(k, j)|. \tag{1.4}$$

A dictionary is incoherent if M is small. There are overcomplete systems with $m \approx n^2$ and $M \approx 1/\sqrt{n}$ [33]. The results in [13], [14], [11], [19] showed that, if there exists a representation $y = \mathbf{\Phi}\alpha$ with sparsity $N = \|\alpha\|_0$, and N does not exceed a threshold $(1 + M^{-1})/2$ defined by M alone (we consider $M > 0$), then a) this is the unique sparsest representation, and b) these algorithms would find it. If, for example, $M = 1/\sqrt{n}$, this result promises, for large n, an ideal form of atomic decomposition even of fairly complex objects. In such cases, provided the object y is made from $< \sqrt{n}/2$ atoms in the dictionary, this sparse decomposition can be uniquely recovered.

C. Presence of Noise

In most practical situations it is not sensible to assume that the available data y obey precise equality $y = \mathbf{\Phi}\alpha$ with a sparse representation α. A more plausible scenario assumes *sparse approximate representation*: that there is an ideal noiseless signal

$x_0 \in R^n$ with a sparse representation, $x_0 = \mathbf{\Phi}\alpha_0$ with $\|\alpha_0\|_0$ small, but that we can observe only a noisy version $y = x_0 + z$, where $\|z\|_2 \leq \epsilon$.

Noise-Aware Variant of (P_0). We can adapt to this noisy setting by modifying (P_0) to include a noise allowance

$$(P_{0,\delta}) : \min_{\alpha} \|\alpha\|_0 \text{ subject to } \|y - \mathbf{\Phi}\alpha\|_2 \leq \delta. \tag{1.5}$$

Note that $(P_{0,0}) \equiv (P_0)$. Also, if we apply this with $\delta \geq \epsilon = \|y - x_0\|_2$, the problem has a sparse solution; in fact, the solution $\hat{\alpha}$ obeys $\|\hat{\alpha}\|_0 \leq \|\alpha_0\|_0$, or more formally, since α_0 is a feasible solution of $(P_{0,\epsilon})$

$$\|\hat{\alpha}\|_0 = \text{val}(P_{0,\delta}) \leq \text{val}(P_{0,\epsilon}) \leq \|\alpha_0\|_0. \tag{1.6}$$

In our notations, val of an optimization problem stands for its value at the solution. Note that $(P_{0,\delta})$ rarely has a unique solution, since once the sparsest solution is found, many feasible variants of it sharing the same support can be built.

Noise-Aware Variant of OGA. Just as with $(P_0), (P_{0,\delta})$ demands exorbitant computational efforts in general, and so again we may resort to heuristics and relaxations. On the one hand, OGA can be employed for approximating the solution of (1.5); the stepwise procedure can simply be stopped when the representation error gets below δ.

Noise-Aware Variant of (P_1). On the other hand, we can pursue a strategy of convexification, replacing the ℓ^0-norm in (1.5) by an ℓ^1-norm

$$(P_{1,\delta}) : \min_{\alpha} \|\alpha\|_1 \text{ subject to } \|y - \mathbf{\Phi}\alpha\|_2 \leq \delta. \tag{1.7}$$

This can be cast as a convex quadratic program which can be solved by many standard approaches, including iteratively reweighted least squares (IRLS) [23], interior-point algorithms [4], and active-set methods. It is also closely related to basis pursuit denoising (BPDN) [4], and to the LASSO technique employed in statistical regression to avoid overfitting when combining predictors [34]. Both those proposals amount to solving a corresponding convex optimization in Lagrangian form

$$(P'_{1,\lambda}) : \min_{\alpha} \|\alpha\|_1 + \|y - \mathbf{\Phi}\alpha\|_2^2/\lambda; \tag{1.8}$$

for appropriate $\lambda = \lambda(y, \delta)$ the solutions of $(P_{1,\delta})$ and $(P'_{1,\lambda(y,\delta)})$ are the same. Solving (1.8) with fixed λ leads to different results—see [40] for an analysis of this option.

We note that instead of ℓ^1, one can use ℓ^p-norm with $p < 1$ in order to better imitate $(P_{0,\delta})$ while losing convexity. This is the spirit behind the FOCUSS method [18].

D. Stability Properties

In this paper, we develop several results exhibiting stable recovery of sparse representations in the presence of noise. We now briefly sketch their statements.

First, we show that when sufficient sparsity is present, where "sufficient" is defined relative to the degree of mutual incoherence, solving $(P_{0,\delta})$ enables stable recovery. We suppose that we have a possibly overcomplete system $\mathbf{\Phi}$ with mutual coherence $M = M(\mathbf{\Phi})$. Suppose that we have a noisy signal y and that the ideal noiseless signal x_0 has a sparse representation α_0 with at most N nonzeros. We have that the noise is bounded, $\|y - x_0\|_2 \leq \epsilon$. Then if $N < (M^{-1} + 1)/2$, it follows that the solution $\hat{\alpha}_{0,\delta}$ of $(P_{0,\delta})$ obeys

$$\|\hat{\alpha}_{0,\delta} - \alpha_0\|_2 \leq \Lambda_0(M, N) \cdot (\epsilon + \delta), \qquad \forall \delta \geq \epsilon > 0 \tag{1.9}$$

with the stability coefficient $\Lambda_0(M, N)^2 = 1/(1 - M(2N - 1))$. The proportionality constant Λ_0 (which we also call the *stability coefficient*) can be quite moderate given sufficient sparsity. In words, provided the underlying object is sparsely represented and the noise level is known, recovery by explicitly imposing sparsity yields an approximation to the ideal sparse decomposition of the noiseless signal in which the error is at worst proportional to the input noise level.

Next, we develop a parallel result for ℓ^1 minimization. Making parallel assumptions, tightened so that the ideal noiseless signal x_0 has a sparse representation α_0 with $N < (M^{-1} + 1)/4$, we show that the solution $\hat{\alpha}_{1,\delta}$ of $(P_{1,\delta})$ obeys

$$\|\hat{\alpha}_{1,\delta} - \alpha_0\|_2 \leq \Lambda_1(M, N) \cdot (\epsilon + \delta), \qquad \forall \delta \geq \epsilon > 0 \tag{1.10}$$

where the stability coefficient

$$\Lambda_1(M, N)^2 = 1/(1 - M(4N - 1)).$$

In words, ℓ^1-based reconstruction in incoherent overcomplete systems has an error which is at worst proportional to the input noise level. The sparsity requirement is twice as stringent for the ℓ^1-based result as for the ℓ^0-based result.

By comparison, OGA obeys a *local* stability result. Again suppose a possibly overcomplete system with $M = M(\mathbf{\Phi})$, and an ideal noiseless signal x_0 having a representation with at most N atoms. Suppose that the smallest among the N nonzeros in the representation of x_0 has amplitude A. Assume that we know the noise level $\epsilon = \|y - x_0\|_2$ and run the OGA just until the representation error $\leq \epsilon$. Call the result of this greedy algorithm $\hat{\alpha}_{\text{OGA},\epsilon}$. Set

$$\Lambda_{\text{OGA}}^2(M, N) = 1/(1 - M(N - 1)).$$

Then if the noise is sufficiently weak

$$\epsilon \leq \frac{A}{2} \cdot (1 - M(2N - 1)) \tag{1.11}$$

the recovered representation $\hat{\alpha}_{\text{OGA},\epsilon}$ obeys

$$\|\hat{\alpha}_{\text{OGA},\epsilon} - \alpha_0\|_2 \leq \Lambda_{\text{OGA}}(M, N) \cdot \epsilon. \tag{1.12}$$

This is a local stability result because for large values of $\epsilon = \|y - x_0\|_2$ the condition (1.11) will necessarily fail.

Note the parallel nature of the bounds and the conclusions. Three quite different algorithms all obey stability results in which having N a fraction of M^{-1} is the key assumption.

E. Support Properties

A different fundamental question about efforts to obtain sparse representation is: *do we actually recover the* ***correct*** *sparsity pattern*? Our stability results do not address this question, since it is possible for a nonsparse representation to be close to a sparse representation in an ℓ^2 sense.

The question is fundamental and broadly significant. Throughout science and technology, it is habitual to fit sparse models to noisy data, and then simply *assume* that terms appearing in such fitted models are dependable features.

In this paper, we are able to shed some light on this situation. Our results show that, under appropriate conditions, the empirical representation $\hat{\alpha}$ is not only at least as sparse as the ideal sparse representation but *it only contains atoms also appearing in the ideal sparse representation*. Since that ideal sparse representation is, by our other results, unique and well-defined, these insights endow the empirical support of $\hat{\alpha}$ with a, perhaps surprising, significance.

Our first result concerns solution of $(P_{1,\delta})$ with $\delta = C \cdot \epsilon$. Here, $C = C(M, N) > 1$, and so we are solving the ℓ^1 minimization problem using an exaggeration of the noise level. It shows, with $M = M(\mathbf{\Phi})$ and $\|\alpha_0\|_0 \leq N$, that the solution $\hat{\alpha}_{1,\delta}$ has its support contained in the support of α_0. Here $C \approx \sqrt{N(1-\beta)}/(1-2\beta)$ for $\beta = MN < 1/2$ for high values of N, so ordinarily it requires considerable overstatement of the noise level to achieve this level of conservatism. However, it does provide the very interesting *epistemological* benefit that the atoms appearing in the representation have more than incidental meaning.

Our second result is obtained in the course of analyzing OGA; it shows that, under condition (1.11) and $NM < 1/2$, the ideal noiseless representation is unique, and the support of $\hat{\alpha}_{\mathrm{OGA}}$ is contained in the support of α_0.

F. Contents

The next sections supply the analysis behind the stability bounds just quoted, and a discussion of support properties. The final section extends this work in various directions.

- *Numerical Results*. We study the actual stability and support recovery behavior of the ℓ^1 and OGA on synthetic examples, finding typical behavior far more favorable than our theoretical bounds.
- *Superresolution*. We situate our work with respect to the problem of superresolution, in which astronomers, seismologists, spectroscopists, and others attempt to "deblur" sparse spike trains.
- *Geometry*. We develop a geometric viewpoint explaining why stability can sometimes be expected for the ℓ^1 penalization scheme, under conditions of sufficient sparsity.

We have recently learned that in parallel to our efforts, there are two similar contributions, handling stability and support recovery for the basis pursuit. J. A. Tropp has been working independently on some of the same problems [40], and so has J. J. Fuchs [16]. After some recent discussions with these authors and a careful study of these works we find that their methods and results have a rather different flavor, ensuring that the three separate works are of interest in studying sparse approximation under noise.

II. Stability Using $(P_{0,\epsilon})$

Suppose again the existence of an ideal noiseless signal $x_0 = \mathbf{\Phi}\alpha_0$ and noisy observations $y = x_0 + z$ with $\|y - x_0\|_2 \leq \epsilon$. Consider applying $(P_{0,\delta})$ with $\delta \geq \epsilon$ to obtain a sparse approximation to y. The following establishes the stability estimate mentioned in the introduction.

Theorem 2.1: Let the dictionary $\mathbf{\Phi}$ have mutual coherence $M = M(\mathbf{\Phi})$. Suppose the noiseless signal $x_0 = \mathbf{\Phi}\alpha_0$, where α_0 satisfies

$$\|\alpha_0\|_0 = N < (1/M + 1)/2. \tag{2.1}$$

Then

a) α_0 is the unique sparsest such representation of x_0; and
b) the reconstruction $\hat{\alpha}_{0,\delta}$ from applying $(P_{0,\delta})$ to the noisy data y approximates α_0

$$\|\hat{\alpha}_{0,\delta} - \alpha_0\|_2^2 \leq \frac{(\epsilon+\delta)^2}{1 - M(2N-1)}, \qquad \forall \delta \geq \epsilon > 0. \tag{2.2}$$

Claim a) actually follows from known results; e.g., see [13], [14] for the two-ortho case, and [11], [19] for general dictionaries. Claim b) requires the following.

Lemma 2.2: Let $M = M(\mathbf{\Phi})$, and let $N < 1/M + 1$. Every $n \times N$ submatrix formed by concatenating N columns from $\mathbf{\Phi}$ has the Nth singular value bounded below by $(1 - M(N-1))^{1/2}$.

Proof: This is equivalent to the claim that

$$\|\mathbf{\Phi}_0 v\|_2^2 \geq \|v\|_2^2 (1 - M(N-1)) \tag{2.3}$$

where $\mathbf{\Phi}_0$ is the concatenation of N columns from $\mathbf{\Phi}$ and v is any vector in $\boldsymbol{R}^N$. Assume without loss of generality these are the first N columns. Let $G = \mathbf{\Phi}_0^H \mathbf{\Phi}_0$ be the corresponding Gram matrix, and write

$$v^H G v = \|v\|_2^2 + \sum_{i \neq j} v(i) v(j) G(i,j). \tag{2.4}$$

Now

$$\begin{aligned} \left| \sum_{i \neq j} v(i) v(j) G(i,j) \right| &\leq M \sum_{i \neq j} |v(i) v(j)| \\ &= M \left(\sum_{i,j} |v(i) v(j)| - \|v\|_2^2 \right) \\ &= M \left(\|v\|_1^2 - \|v\|_2^2 \right) \leq \|v\|_2^2 (N-1) M. \end{aligned}$$

Using this inequality in the identity (2.4) gives (2.3). □

Apply this to $y = x_0 + z$, with $\hat{x}_{0,\delta} = \mathbf{\Phi}\hat{\alpha}_{0,\delta}$, and we have

$$\|x_0 - \hat{x}_{0,\delta}\|_2 \leq \|y - \hat{x}_{0,\delta}\|_2 + \|z\|_2 \leq \delta + \epsilon.$$

Also, note that, if $\delta \geq \epsilon$, then $\hat{x}_{0,\delta}$ is a linear combination of at most N columns of $\mathbf{\Phi}$. Thus, $x_0 - \hat{x}_{0,\delta}$ can be written as a linear combination of at most $2N$ columns from $\mathbf{\Phi}$. Applying Lemma 2.2, or, more properly, the inequality (2.3), gives the result. □

III. Stability Using $(P_{1,\epsilon})$

As in the Introduction, we are given a signal $y = \mathbf{\Phi}\alpha_0 + z$, where z is an additive noise, known to satisfy $\|z\|_2 \leq \epsilon$. We apply $(P_{1,\delta})$ to this signal (necessarily with $\delta \geq \epsilon$); i.e., we solve (1.7) and obtain a solution $\hat{\alpha}_{1,\delta}$. We study its deviation from the ideal representation α_0.

A. Stability Result

Theorem 3.1: Let the overcomplete system $\mathbf{\Phi}$ have mutual coherence $M(\mathbf{\Phi})$. If some representation of the noiseless signal $x_0 = \mathbf{\Phi}\alpha_0$ satisfies

$$N = \|\alpha_0\|_0 < (1/M + 1)/4 \tag{3.1}$$

then the deviation of the $(P_{1,\delta})$ representation from α_0, assuming $\delta \geq \epsilon$, can be bounded by

$$\|\hat{\alpha}_{1,\delta} - \alpha_0\|_2^2 \leq \frac{(\epsilon+\delta)^2}{1 - M(4N-1)}. \tag{3.2}$$

Proof: The stability bound can be posed as the solution to an optimization problem of the form

$$\max_{\alpha_0, z} \|\hat{\alpha} - \alpha_0\|_2^2 \text{ subject to } \left\{ \begin{array}{c} \hat{\alpha} = \arg\min_\alpha \|\alpha\|_1 \text{ subject to } \|\mathbf{\Phi}\alpha - y\|_2 \leq \delta \\ y = \mathbf{\Phi}\alpha_0 + z, \|z\|_2 \leq \epsilon, \|\alpha_0\|_0 \leq N. \end{array} \right\}. \tag{3.3}$$

Put in words, we consider all representation vectors α_0 of bounded support, and all possible realizations of bounded noise, and we ask for the largest error between the ideal sparse decomposition and its reconstruction from noisy data. Defining $v = \alpha - \alpha_0$, and similarly $w = \hat{\alpha} - \alpha_0$, we can rewrite the above problem as

$$\max_{\alpha_0, z} \|w\|_2^2 \text{ subject to } \left\{ \begin{array}{c} w = \arg\min_v \|\alpha_0 + v\|_1 \text{ subject to } \|\mathbf{\Phi}v - z\|_2 \leq \delta \\ \|z\|_2 \leq \epsilon, \|\alpha_0\|_0 \leq N. \end{array} \right\}. \tag{3.4}$$

We develop an upper bound on val (3.4) in a sequence of relaxations, each one expanding the feasible set and increasing the maximal value. To begin, note that if w is the minimizer of $\|\alpha_0 + v\|_1$ under these constraints, then relaxing the constraints to all w satisfying $\|\alpha_0 + w\|_1 \leq \|\alpha_0\|_1$ expands the feasible set. However, this is true only if $\delta \geq \epsilon$ since otherwise $v = 0$ is not a feasible solution. Thus, we consider

$$\left\{ w \;\middle|\; \begin{array}{c} \|\alpha_0 + w\|_1 \leq \|\alpha_0\|_1 \& \|\mathbf{\Phi}w - z\|_2 \leq \delta \\ \|z\|_2 \leq \epsilon, \|\alpha_0\|_0 \leq N \end{array} \right\}. \tag{3.5}$$

We now expand this set by exploiting the relation

$$\|\alpha_0 + w\|_1 - \|\alpha_0\|_1 \geq \|w\|_1 - 2\sum_{k \in \mathcal{S}} |w(k)|$$

where $\mathcal{S}$ is the support of the nonzeros in α_0 with complement $\mathcal{S}^c$, and we used $|a+b| - |a| \geq |a| - |b| - |a| = -|b|$. Therefore, we get a further increase in value by replacing the feasible set in (3.5) with

$$\left\{ w \;\middle|\; \begin{array}{c} \|w\|_1 \leq 2\sum_{k \in \mathcal{S}} |w(k)|, \|\mathbf{\Phi}w - z\|_2 \leq \delta \\ \|z\|_2 \leq \epsilon, \#\mathcal{S} \leq N \end{array} \right\}. \tag{3.6}$$

Writing this out yields a new optimization problem with still larger value

$$\max_{w, \mathcal{S}, z} \|w\|_2^2 \text{ subject to } \left\{ \begin{array}{c} \|w\|_1 \leq 2\sum_{k \in \mathcal{S}} |w(k)|, \|\mathbf{\Phi}w - z\|_2 \leq \delta \\ \|z\|_2 \leq \epsilon, \#\mathcal{S} \leq N \end{array} \right\}. \tag{3.7}$$

We next simplify our analysis by eliminating the noise vector z, using

$$\{w | \|\mathbf{\Phi}w - z\|_2 \leq \delta \& \|z\|_2 \leq \epsilon\} \subseteq \{w | \|\mathbf{\Phi}w\|_2 \leq \epsilon + \delta\}. \tag{3.8}$$

Expanding the feasible set of (3.7) using this observation gives

$$\max_{\mathcal{S}, w} \|w\|_2^2 \text{ subject to } \left\{ \begin{array}{c} \|w\|_1 < 2\sum_{k \in \mathcal{S}} |w(k)|, \;\; \|\mathbf{\Phi}w\|_2 \leq \Delta \\ \#\mathcal{S} \leq N \end{array} \right\} \tag{3.9}$$

where we introduced $\Delta = \epsilon + \delta$.

The constraint $\|\mathbf{\Phi}w\|_2 \leq \Delta$ is not posed in terms of the absolute values in the vector w, complicating the analysis; we now relax this constraint using incoherence of $\mathbf{\Phi}$. Again, the Gram matrix is $\boldsymbol{G} = \mathbf{\Phi}^T\mathbf{\Phi}$, and the mutual coherence is the maximal off-diagonal amplitude: $M = \max_{k \neq j} |G(k,j)|$. For a vector w, let $|w|$ be vector containing absolute values from w; similarly for matrices. Also, let $\mathbf{1}$ be the the m-by-m matrix of all ones. The constraint

$$\|\mathbf{\Phi}w\|_2^2 = w^T\boldsymbol{G}w \leq \Delta^2$$

can be relaxed

$$\begin{aligned} \Delta^2 &\geq w^T\boldsymbol{G}w = \|w\|_2^2 + w^T(\boldsymbol{G} - \boldsymbol{I})w \\ &\geq \|w\|_2^2 - |w|^T|\boldsymbol{G} - \boldsymbol{I}||w| \\ &\geq \|w\|_2^2 - M|w|^T|\mathbf{1} - \boldsymbol{I}||w| \\ &= (1+M)\|w\|_2^2 - M\|w\|_1^2. \end{aligned} \tag{3.10}$$

Using this, val (3.9) is bounded above by the value

$$\max_{\mathcal{S}, w} \|w\|_2^2 \text{ subject to } \left\{ \begin{array}{c} \|w\|_1 < 2\sum_{k \in \mathcal{S}} |w(k)| \\ (1+M)\|w\|_2^2 - M\|w\|_1^2 \leq \Delta^2 \\ \#\mathcal{S} \leq N \end{array} \right\}. \tag{3.11}$$

This problem is invariant under permutations of the entries in w which preserve membership in $\mathcal{S}$ and $\mathcal{S}^c$. It is also invariant under relabeling of coordinates. So assume that all nonzeros in α_0 are concentrated in the initial slots of the vector, i.e., that $\mathcal{S} = \{1, \ldots, N\}$.

Putting $w = (w_0, w_1)$ where w_0 gives the first N entries in w, and w_1 the remaining $m - N$ entries of w, we obviously have

$$\|w\|_2^2 = \|w_0\|_2^2 + \|w_1\|_2^2, \|w\|_1 = \|w_0\|_1 + \|w_1\|_1. \tag{3.12}$$

The ℓ^1 norm on R^k dominates the ℓ^2 norm and is dominated by $\sqrt{k}$ times the ℓ^2 norm. Thus,

$$\|w_0\|_1 \geq \|w_0\|_2 \geq \frac{\|w_0\|_1}{\sqrt{N}}, \qquad \|w_1\|_1 \geq \|w_1\|_2 \geq \frac{\|w_1\|_1}{\sqrt{m-N}}. \tag{3.13}$$

We define

$$\begin{aligned} &A = \|w_0\|_1, B = \|w_1\|_1 \\ &c_0 = \left(\frac{\|w_0\|_2}{\|w_0\|_1}\right)^2, \quad c_1 = \left(\frac{\|w_1\|_2}{\|w_1\|_1}\right)^2. \end{aligned} \tag{3.14}$$

Returning to the problem given in (3.11), and using our notations, we obtain a further reduction, from an opti-

mization problem on R^m to an optimization problem on $(A, B, c_0, c_1) \in R^4$

$$\max c_0A^2 + c_1B^2 \text{ subject to } \left\{ \begin{array}{c} A > B \\ (1+M)(c_0A^2 + c_1B^2) - M(A+B)^2 \le \Delta^2 \\ A, B \ge 0, \frac{1}{N} \le c_0 \le 1, 0 < c_1 \le 1 \end{array} \right\}. \quad (3.15)$$

We further define $B = \rho A$, where $0 \le \rho < 1$ and rewrite (3.15) as

$$\max (c_0 + \rho^2 c_1)A^2 \text{ subject to } \left\{ \begin{array}{c} (1+M)(c_0 + \rho^2 c_1)A^2 - M(1+\rho)^2A^2 \le \Delta^2 \\ A \ge 0, \frac{1}{N} \le c_0 \le 1, 0 < c_1 \le 1, 0 \le \rho < 1 \end{array} \right\}. \quad (3.16)$$

Define $\mu = (1+\rho)^2/(c_0+\rho^2 c_1)$. Then $1 \le \mu \le 4N$ over the region (3.16). Setting $V = A^2(c_0 + \rho^2 c_1)$, the first constraint defining that region takes the form

$$(1+M)V - M\mu V \le \Delta^2. \quad (3.17)$$

Since $\mu \le 4N$, the sparsity requirement (3.1) leads to

$$(1+M) - M\mu \ge (1+M) - 4NM > 0. \quad (3.18)$$

Hence,

$$V \le \frac{\Delta^2}{1 - M(\mu - 1)} \le \frac{\Delta^2}{1 - M(4N-1)} \quad (3.19)$$

as stated by (3.2) with the choice $\mu = 4N$.

The requirement (3.18) puts a restriction on N and M, being free parameters of the problem. Using $\mu = 4N$ leads to the sparsity requirement in (3.1), since $(1+M) - 4NM > 0$. □

B. Interpretation and Comments

Theorem 3.1 prompts several remarks.

- Setting $\epsilon = \delta = 0$ puts us in the noiseless case $(P_{1,0})$. In that setting, Theorem 3.1 tells us that if $N < (1 + M^{-1})/4$, there will be zero error in finding the unique sparsest representation—i.e., solving the ℓ^1 optimization problem $(P_{1,0})$ solves the ℓ^0 problem $(P_{0,0})$. As the ℓ^1 problem is convex and the ℓ^0 problem combinatorial in general, this is by itself significant. The same general phenomenon described has been observed before in [13], [14], [11], [19]. The sharpest results, in [11], [19], established that this phenomenon occurs for any sparsity N smaller than $(1 + M^{-1})/2$, which means that the new result is slack by a factor of 2 in the $\epsilon = 0$ case. Perhaps a tighter inequality could be achieved with more care.
- If the signal is not noisy (i.e., $\epsilon = 0$) but nevertheless $(P_{1,\delta})$ is employed with $\delta > 0$, an approximate solution is assured, with a bound on the deviation of the approximate representation from the ideal noiseless representation. So in "needlessly" going from $(P_{1,0})$ to $(P_{1,\delta})$ we tolerate errors in the decomposition, but the errors are controlled.
- In practice, if the signal is noisy (i.e., $\epsilon > 0$) and we set $\delta = 0$ as if there were no noise, some degree of stability is still obtained! However, our results do not cover this case, and further work is required to establish this kind of stability.

IV. Support Recovery With ℓ^1

So far, we have concentrated on the recovery of x_0. We now consider whether we can correctly recover the *support* of x_0. Our approach applies $(P_{1,\delta})$ with a specially chosen $\delta \gg \epsilon$.

Theorem 4.1: Suppose that $y = x_0 + z$ where $x_0 = \mathbf{\Phi}\alpha_0$, $\|\alpha_0\|_0 \le N$, and $\|z\| \le \epsilon$. Let $M = M(\mathbf{\Phi}) > 0$ and suppose $\beta \equiv MN < 1/2$. Set

$$\gamma = \frac{\sqrt{(1-\beta)} + (1-\beta)/\sqrt{N}}{1 - 2\beta}. \quad (4.1)$$

Solve $(P_{1,\delta})$ with exaggerated noise level $\delta > \gamma \cdot \sqrt{N} \cdot \epsilon$. Then $\mathrm{supp}(\hat{\alpha}_{1,\delta}) \subset \mathrm{supp}(\alpha_0)$.

As an example, if $\beta = 1/4$, exaggerating the noise level by a factor $\sqrt{3N} + 3/2$ leads to partial support recovery. This pronounced inflation of the noise level might cause (in an extreme case) a zero solution. Still, from the following proof it seems that $\sqrt{N}$ dependence is intrinsic to the problem.

Proof: For later use, we let $\eta = \eta(\mathbf{\Phi})$ be the smallest Nth singular value of any submatrix $\mathbf{\Phi}_0$ of $\mathbf{\Phi}$ containing N columns from $\mathbf{\Phi}$. By our assumptions and Lemma 2.2

$$\eta \ge (1 - M(N-1))^{1/2} > (1-\beta)^{1/2}. \quad (4.2)$$

We need notation for the convex functional $L(\alpha) = \|\alpha\|_1$ and for the quadratic functional $Q(\alpha) = \|y - \mathbf{\Phi}\alpha\|_2^2$.

Now let $\mathcal{S}$ be the support of the ideal noiseless representation α_0, and consider the support-constrained optimization problem $(P_{1,\delta,\mathcal{S}})$ where feasible vectors α must be supported in $\mathcal{S}$. Let α_1 be a solution of this problem. We claim that, in fact, α_1 is actually the solution of the *support-unconstrained* problem $(P_{1,\delta})$, i.e., $\alpha_1 = \hat{\alpha}_{1,\delta}$. Observe, first of all, that unless $\alpha_1 \ne 0$, there is nothing to prove. Now consider perturbations u of α_1 i.e., representations of the form $\alpha_1 + t \cdot u$, for $t > 0$ small. We will show that a perturbation which does not increase the ℓ^1 objective, definitely violates the constraint. Formally

$$L(\alpha_1 + tu) \le L(\alpha_1) \quad \text{for all sufficiently small } t > 0 \quad (4.3)$$

implies

$$Q(\alpha_1 + tu) > Q(\alpha_1) \quad \text{for all sufficiently small } t > 0. \quad (4.4)$$

It follows that α_1 is locally optimal for $(P_{1,\delta})$. By convexity, this local condition implies global optimality; and global optimality implies that the solution has support $\mathcal{S}_0 \subset \mathcal{S}$ as claimed.

We now note that, for $t > 0$

$$Q(\alpha_1 + tu) - Q(\alpha_1) = -2t\langle r, \mathbf{\Phi}u\rangle + t^2\|\mathbf{\Phi}u\|^2 \quad (4.5)$$

where $r \equiv y - \mathbf{\Phi}\alpha_1$ and $\mathcal{S}_0 = \mathrm{supp}(\alpha_1) \subset \mathcal{S}$, while for sufficiently small positive t

$$\begin{aligned} &L(\alpha_1 + tu) - L(\alpha_1) \\ &\quad = \sum_{j \in \mathcal{S}_0} (|\alpha_1(j) + tu(j)| - |\alpha_1(j)|) + t \cdot \sum_{j \in \mathcal{S} \setminus \mathcal{S}_0} |u(j)| \\ &\quad = t \cdot \left(\sum_{j \in \mathcal{S}_0} \sigma(j)u(j) + \sum_{j \in \mathcal{S} \setminus \mathcal{S}_0} |u(j)| \right) \end{aligned} \quad (4.6)$$

where $\sigma(j) = \mathrm{sgn}(\alpha_1(j))$ for $j \in \mathcal{S}_0$, and 0 otherwise, and we used the identity $|a+b|-|a| = \mathrm{sgn}(a)b$, valid for $|b| < |a|$. Let $\dot{Q}(u)$ and $\dot{L}(u)$ denote the terms of order t in (4.5) and (4.7); we plan to show that

$$\dot{L}(u) \le 0 \Longrightarrow \dot{Q}(u) > 0, \qquad \forall u \in R^m; \tag{4.7}$$

this will show that (4.3) implies (4.4).

We first work out the consequence of α_1 solving $(P_{1,\delta,\mathcal{S}})$. Since $\alpha_1 \neq 0$, there is a nonempty subspace of vectors supported in $\mathcal{S}_0$; (4.5) and (4.7) show that Q and L are both differentiable at α_1. The fact that α_1 solves the constrained optimization problem $(P_{1,\delta,\mathcal{S}})$, implies (by classical constrained optimization/Lagrange multiplier ideas) that for some $\lambda > 0$

$$\nabla Q = -\lambda \nabla L, \qquad \text{at } \alpha = \alpha_1.$$

This implies that for vectors u supported in $\mathcal{S}_0$ we must have $\dot{Q}(u) = -\lambda \dot{L}(u)$, or

$$-2\langle r, \mathbf{\Phi} u\rangle = -\lambda \cdot \sum_{j \in \mathcal{S}_0} \sigma(j) u(j). \tag{4.8}$$

We introduce the notation $\langle\, , \rangle$ for inner product in R^m, $\langle\, , \rangle_0$ for inner product restricted to coordinates in $\mathcal{S}_0$, and $\langle\, , \rangle_1$ for inner product restricted to coordinates in $\mathcal{S}_0^c$. We introduce the notation $||||_{1,0}$ for the ℓ^1 norm restricted to $\mathcal{S}_0$, etc. To illustrate the notation, we have that, if u is a vector supported in $\mathcal{S}_0$, $\langle r, \mathbf{\Phi} u\rangle = \langle \mathbf{\Phi}^T r, u\rangle_0$. Then (4.8) says that, for all u

$$2\langle \mathbf{\Phi}^T r, u\rangle_0 = \lambda \langle \sigma, u\rangle_0. \tag{4.9}$$

We now show that α_1 is the unique global optimum of $(P_{1,\delta})$—i.e., the original problem, without the support constraint. We write

$$\langle r, \mathbf{\Phi} u\rangle = \langle \mathbf{\Phi}^T r, u\rangle_0 + \langle \mathbf{\Phi}^T r, u\rangle_1;$$

then from (4.9)

$$\dot{Q}(u) = -2\langle \mathbf{\Phi}^T r, u\rangle = -\lambda \langle \sigma, u\rangle_0 - 2\langle \mathbf{\Phi}^T r, u\rangle_1$$

while

$$\dot{L}(u) = \langle \sigma, u\rangle_0 + ||u||_{1,1}.$$

Hence, $\dot{L}(u) \le 0$ implies $||u||_{1,1} \le -\langle \sigma, u\rangle_0$, and also

$$\dot{Q}(u) \ge \lambda ||u||_{1,1} - 2\langle \mathbf{\Phi}^T r, u\rangle_1.$$

Hence, (4.7) follows from

$$2\langle \mathbf{\Phi}^T r, u\rangle_1 < \lambda ||u||_{1,1}. \tag{4.10}$$

Later we will show that

$$\langle \mathbf{\Phi}^T r, u\rangle_1 \le ||u||_{1,1} \cdot \left(\epsilon + \delta \frac{M\sqrt{N}}{\sqrt{1-\beta}} \right). \tag{4.11}$$

Thus, (4.10) follows from

$$\epsilon + \delta \frac{M\sqrt{N}}{\sqrt{1-\beta}} < \lambda/2. \tag{4.12}$$

Recalling (4.9), and choosing $u = \mathbf{\Phi}^T r$, shows that

$$\lambda \ge 2 \cdot ||\mathbf{\Phi}^T r||_{2,0}/\sqrt{N}.$$

Using $||r||_2 = \delta$ and the definition (4.2) of η

$$||\mathbf{\Phi}^T r||_{2,0} \ge \eta \cdot (\delta - \epsilon) > \sqrt{1-\beta} \cdot (\delta - \epsilon)$$

and so

$$\lambda/2 > \frac{\sqrt{1-\beta} \cdot (\delta - \epsilon)}{\sqrt{N}}. \tag{4.13}$$

In short, for (4.12) and hence (4.10) to follow, we should require

$$\epsilon + \delta \frac{M\sqrt{N}}{\sqrt{1-\beta}} < (\delta - \epsilon) \frac{\sqrt{1-\beta}}{\sqrt{N}} \tag{4.14}$$

which can be rearranged to

$$\gamma = \frac{\delta}{\epsilon\sqrt{N}} \ge \frac{\sqrt{(1-\beta)} + (1-\beta)/\sqrt{N}}{1-2\beta}. \tag{4.15}$$

This holds because our definition of γ makes this an equality.

It only remains to demonstrate (4.11). Write $r = r_0 + r_1$, where r_1 is the component of r not in the span of $(\phi_j : j \in \mathcal{S}_0)$, which has norm $\le \epsilon$, while r_0 is the component of r in the span of the $(\phi_j : j \in \mathcal{S}_0)$, with norm $\le \delta$. Hence,

$$\langle \mathbf{\Phi}^T r, u\rangle_1 = \langle \mathbf{\Phi}^T r_0, u\rangle_1 + \langle \mathbf{\Phi}^T r_1, u\rangle_1$$

$$\langle \mathbf{\Phi}^T r_1, u\rangle_1 \le ||u||_{1,1} \left(\max_{j \in \mathcal{S}_0^c} |(\mathbf{\Phi}^T r_1)(j)| \right) \le ||u||_{1,1} \cdot \epsilon$$

and

$$\begin{aligned} \langle \mathbf{\Phi}^T r_0, u\rangle_1 &= \langle \mathbf{\Phi}_0 v, \mathbf{\Phi}_1 u\rangle \\ &= \sum_{i \in \mathcal{S}_0, j \in \mathcal{S}_0^c} G(i,j) v(i) u(j) \le M\sqrt{N} ||u||_{1,1} ||v||_2 \end{aligned} \tag{4.16}$$

where $\mathbf{\Phi}_0 v = r_0$, $G(i,j)$ are entries of the Gram matrix $\mathbf{\Phi}^T \mathbf{\Phi}$ (known to be in the range $[-M, M]$) and $\mathbf{\Phi}_0$ is the submatrix of Φ with columns from $\mathcal{S}$ only. The definition of η yields

$$\eta ||v||_2 \le ||r_0||_2$$

and since $||r_0||_2 \le \delta$, (4.16) gives

$$\langle \mathbf{\Phi}^T r_0, u\rangle_1 \le \delta \cdot M\sqrt{N}/\eta \cdot ||u||_{1,1}$$

giving (4.11). □

V. Local Stability of the OGA

Observe that both $(P_{0,\epsilon})$ and $(P_{1,\epsilon})$ refer to global optimization problems, while the OGA described in the Introduction is based on greedy stagewise approximation. Paralleling this distinction, the stability result we now develop for OGA is a local one, valid only for sufficiently small $\epsilon < \epsilon^*(\alpha_0)$, depending on the representation coefficients.

For ease of exposition, we shall hereafter assume that the order of the columns ϕ_1, $\phi_2, \ldots$ in the overcomplete system matrix $\mathbf{\Phi}$ has been chosen so that in the ideal noiseless signal $x_0 = \mathbf{\Phi}\alpha_0$, the first N entries in α_0 are the nonzero entries, and that these are ordered

$$|\alpha_0(1)| \ge |\alpha_0(2)| \ge \cdots \ge |\alpha_0(N)|. \tag{5.1}$$

Theorem 5.1: Suppose the ideal noiseless signal x_0 has a representation $x_0 = \mathbf{\Phi}\alpha_0$ satisfying[1]

$$N = \|\alpha_0\|_0 \leq \frac{1+M}{2M} - \frac{1}{M} \cdot \frac{\epsilon}{|\alpha_0(N)|}. \tag{5.2}$$

Denote by $\hat{\alpha}_{\mathrm{OGA},\epsilon}$ the result of greedy stepwise least-squares fitting applied on the noisy signal y, which stops as soon as the representation error $\leq \epsilon$. Then

a) $\hat{\alpha}_{\mathrm{OGA},\epsilon}$ has the correct sparsity pattern

$$\mathrm{supp}(\hat{\alpha}_{\mathrm{OGA},\epsilon}) = \mathrm{supp}(\alpha_0); \tag{5.3}$$

b) $\hat{\alpha}_{\mathrm{OGA},\epsilon}$ approximates the ideal noiseless representation

$$\|\hat{\alpha}_{\mathrm{OGA},\epsilon} - \alpha_0\|_2^2 \leq \frac{\epsilon^2}{1 - M(N-1)}. \tag{5.4}$$

Note that the argument assumes the noise level ϵ is known, to enable the stopping rule in the algorithm. This parallels the assumption $\delta = \epsilon$, which we relaxed in Theorems 2.1 and 3.1 by using $\delta \geq \epsilon$.

The general idea—that the support properties of α_0 and $\hat{\alpha}_{\mathrm{OGA},\epsilon}$ are the same—seems worthy of its own study. In the Technical Report [12] on which this paper is based, we call this the *trapping* property, and develop it further.

We break our analysis in two stages, considering claims (5.3) and (5.4) in turn.

A. Getting the "Correct" Support

Lemma 5.2: Suppose that we have a signal y satisfying $y = x_0 + z$ where x_0 admits sparse synthesis $x_0 = \mathbf{\Phi}\alpha_0$ using at most N atoms, where

$$N < \frac{1+M}{2M} - \frac{1}{M} \cdot \frac{\epsilon}{\|\alpha\|_\infty} \tag{5.5}$$

and where $\|z\| \leq \epsilon$. Then the first step of the greedy algorithm selects an atom index from among the $\leq N$ nonzeros in α_0.

Proof: The greedy algorithm operates by projecting y onto each atom ϕ_k in turn, selecting an atom index where the projection magnitude is largest. The lemma will follow from

$$\max_{1 \leq k \leq N} |\langle y, \phi_k \rangle| > \max_{k > N} |\langle y, \phi_k \rangle|. \tag{5.6}$$

We now develop a lower bound on the left-hand side and an upper bound on the right-hand side which guarantees this. Assuming that the largest amplitude entry in α_0 occurs in slot 1, the left-hand side of (5.6) is bounded below by

$$\begin{aligned} |\langle y, \phi_1 \rangle| &= |\langle x_0 + z, \phi_1 \rangle| \\ &\geq \left| \left\langle \sum_{j=1}^{N} \alpha_0(j)\phi_j, \phi_1 \right\rangle \right| - |\langle z, \phi_1 \rangle| \\ &\geq |\alpha_0(1)| - \sum_{j=2}^{N} |\alpha_0(j)| \cdot |\langle \phi_j, \phi_1 \rangle| - \epsilon \\ &\geq |\alpha_0(1)| - |\alpha_0(1)| \cdot (N-1)M - \epsilon. \end{aligned} \tag{5.7}$$

[1]This inequality is equivalent to the one posed in (1.11).

We used $\|\phi_j\|_2^2 = 1$ for all j; $|\langle \phi_j, \phi_1 \rangle| \leq M$ for $j \neq 1$; $\|z\|_2 \leq \epsilon$; and the ordering of the $|\alpha(k)|$. The right-hand side of (5.6) can be bounded above by the same approach, leading to, for $k > N$

$$\begin{aligned} |\langle y, \phi_k \rangle| &= |\langle x_0 + z, \phi_k \rangle| \\ &\leq \sum_{j=1}^{N} |\alpha_0(j)| \cdot |\langle \phi_j, \phi_k \rangle| + |\langle z, \phi_k \rangle| \\ &\leq |\alpha_0(1)| \cdot NM + \epsilon. \end{aligned} \tag{5.8}$$

Imposing (5.5) and using the two bounds (5.7) and (5.8), we see that

$$|\alpha_0(1)| - |\alpha_0(1)| \cdot (N-1)M - \epsilon > |\alpha_0(1)| \cdot NM + \epsilon. \tag{5.9}$$

Relation (5.6) follows; the greedy algorithm therefore chooses at Stage 1 one of the nonzeros in the ideal representation of the noiseless signal. □

To continue to later stages, we need the following.

Lemma 5.3: Let $x_0 = \sum_{i=1}^{N} \alpha_0(i)\phi_i$ and $y = x_0 + z$ with $\|z\|_2 \leq \epsilon$. Let $\mathcal{S}_k$ be a set of k indices in $\{1, \ldots, m\}$, m being the number of columns in $\mathbf{\Phi}$. Let β_k be a vector of m coefficients with k nonzeros located at the indices in $\mathcal{S}_k$. Define a new signal y_k by subtracting k atoms with nonzero coefficients in β_k

$$y_k = y - \sum_{i \in \mathcal{S}_k} \beta_k(i)\phi_i.$$

Similarly, define

$$x_k = x_0 - \sum_{i \in \mathcal{S}_k} \beta_k(i)\phi_i.$$

Then

- if $\mathcal{S}_k \subset \{1, \ldots, N\}$, where $N < (1 + M^{-1})/2$, x_k has a unique sparsest representation $x_k = \mathbf{\Phi}\alpha_k$ made of at most N atoms; these are all atoms originally appearing in the representation of x_0;
- the new signal y_k can be viewed as a superposition of x_k and noise z_k, with noise level $\epsilon_k = \|z_k\|_2 \leq \epsilon$.

Proof: Define the vector

$$\alpha_k(i) = \begin{cases} \alpha_0(i) - \beta_k(i), & i \in \mathcal{S}_k \\ \alpha_0(i), & i \notin \mathcal{S}_k. \end{cases}$$

Then clearly $x_k = \mathbf{\Phi}\alpha_k$. Also, $\mathrm{supp}(\alpha_k) \subseteq \mathrm{supp}(\alpha)$. Since then

$$\|\alpha_k\|_0 \leq \|\alpha_0\|_0 \leq N < (1 + M^{-1})/2$$

we conclude that α_k is the unique sparsest representation of x_k. Moreover

$$\begin{aligned} \epsilon_k &= \|y_k - x_k\| \\ &= \|\left(y - \sum \beta_k(i)\phi_i\right) - \left(x_0 - \sum a_k(i)\phi_i\right)\|_2 \\ &= \|y - x_0\|_2 = \|z\|_2 \leq \epsilon. \end{aligned}$$

Hence we have established the two claims. □

The impact of the preceding two lemmas is that selection of a term, followed by the formation of the residual signal, leads us to a situation like before, where the ideal noiseless signal has no more atoms than before, and the noise level is the same.

We wish to repeatedly apply these lemmas. Starting with $\alpha = \alpha_0$, we will get an α_1 and an i_1; we then hope to apply the observations again, getting α_2 and i_2, etc. If we are allowed to continue invoking these lemmas for N steps, we produce in this way series $\alpha_0, \ldots, \alpha_N$, and $i_1, \ldots, i_N$. Naturally, the sets $\mathcal{S}_k = \{i_1, \ldots, i_k\}$ are nested.

Note, however, that a series of conditions must be satisfied for the repeated use of the first lemma. At the kth iteration, we need the following analog of (5.5):

$$N \leq \frac{1+M}{2M} - \frac{1}{M} \cdot \frac{\epsilon}{\|\alpha_{k-1}\|_\infty}. \tag{5.10}$$

This will follow from our original assumption (5.2) and the fact that if α_k differs from α_0 in at most k places, and α_0 is ordered as in (5.1), then

$$\|\alpha_k\|_\infty \geq |\alpha_0(k+1)|.$$

The above, along with the ordering assumption on the coefficients in α_0, show that $\|\alpha_k\|_\infty \geq |\alpha_0(N)|$ for $1 \leq k \leq N$, and so the sequence of conditions (5.10) is implied by the final one at $k = N$, which is a consequence of (5.2). Hence, assumption (5.2) allows us to repeatedly apply Lemmas 5.2 and 5.3, and conclude that atom indices selected at stages $1 \leq k \leq N$ obey $1 \leq i_k \leq N$: only correct atoms are selected.

In fact, we can say much more. The coefficient sequence β_k generated at stage k solves the least-squares problem

$$\min_a \left\| y - \sum_{i \in \mathcal{S}_k} \beta(i) \phi_i \right\|_2. \tag{5.11}$$

This ensures that the signal y_k is actually orthogonal to each atom selected at stages $1, \ldots k$. Hence, OGA is forced to select from among the atom indices in $\{1, \ldots, N\}$ always one of the previously unselected ones. It therefore by stage N selects all N atom indices in $\{1, \ldots, N\}$. Now by assumption, the residual at that stage has ℓ^2 norm $\leq \epsilon$. Hence, the stopping criterion must be reached by stage N. At the same time, by inspecting (5.9) we see that at stages $k < N$, each selected term $|\alpha_0(i_k)| > \epsilon$. This implies that the stopping criterion cannot be met before stage N, as $\|y_{k-1}\|_2^2 \geq |\alpha_0(i_k)|^2 > \epsilon^2$. Thus, we have proved the following.

Lemma 5.4: OGA stops after precisely N steps.

This, in turn, proves Claim a) of the theorem, (5.3). □

B. Stability Result

Now we turn to Claim b) of Theorem 5.1. We may partition $\mathbf{\Phi} = [\mathbf{\Psi}\, \mathbf{\Omega}]$, where $\mathbf{\Psi}$ denotes the first N columns in $\mathbf{\Phi}$ and $\mathbf{\Omega}$ the remainder. The OGA solves (5.11) with $\mathcal{S}_N = \{1, \ldots, N\}$, or

$$\hat{\alpha}_{\mathrm{OGA},\epsilon} = \arg\min_\alpha \|\mathbf{\Psi}\alpha - y\|_2^2 = \mathbf{\Psi}^+ y \tag{5.12}$$

where $\mathbf{\Psi}^+$ denotes the Moore–Penrose generalized inverse of $\mathbf{\Psi}$. Recall that the signal has a representation $y = \mathbf{\Phi}\alpha_0 + z$ with $\|z\|_2 \leq \epsilon$. Thus, using the formula above we have

$$\hat{\alpha}_{\mathrm{OGA},\epsilon} = \mathbf{\Psi}^+ y = \mathbf{\Psi}^+([\mathbf{\Psi}\mathbf{\Omega}]\alpha_0 + z). \tag{5.13}$$

We may partition $\alpha_0 = [\beta, 0]$ where β contains the first N entries of α, and similarly $\hat{\alpha}_{\mathrm{OGA},\epsilon} = [\hat{\beta}, 0]$. We obtain

$$\hat{\beta} = \beta + \mathbf{\Psi}^+ z. \tag{5.14}$$

The vector $\mathbf{\Psi}^+ z$ represents reconstruction error, and we have the error bound

$$\|\hat{\alpha}_{\mathrm{OGA},\epsilon} - \alpha_0\|_2 = \|\mathbf{\Psi}^+ z\|_2 \leq \|\mathbf{\Psi}^+\|_2 \cdot \|z\|_2 \leq \epsilon/\eta \tag{5.15}$$

where we bounded the norm of $\mathbf{\Psi}^+$ in terms η, the smallest singular value of $\mathbf{\Psi}$. Lemma 2.2 gives $\eta^2 \geq 1 - M(N-1)$. Expression (5.4) follows. □

C. Relation to N-Term Approximation

There has recently been a great deal of interest in the greedy algorithm as a method to generate near-best N-term approximations. Relevant literature includes [17], [35], [36], [38], [39]. The questions asked in this literature concern the quality of N-term approximations $x_N = \mathbf{\Phi}\alpha_N$, where $\|\alpha_N\|_0 = N$, to approximate a general object x. More specifically, let $x_{N,\mathrm{OGA}}$ be the N-term approximation to a vector x by the orthogonal greedy algorithm run through N steps, and $x_{N,0}$ be the optimal N-term approximation, obtained by

$$x_{N,0} = \mathrm{argmin}\{\|x - \mathbf{\Phi}\alpha_N\|_2 : \|\alpha_N\|_0 \leq N\}.$$

The central question is to compare the approximation errors of the two approaches

$$\epsilon_{N,0} \equiv \|x - x_{N,0}\| \approx \epsilon_{N,\mathrm{OGA}} \equiv \|x - x_{N,\mathrm{OGA}}\|.$$

In this direction, the first result was provided by Gilbert *et al.* [17]

$$\epsilon_{N,\mathrm{OGA}} \leq 8\sqrt{N}\epsilon_{N,0}, \qquad N < 1/(8M).$$

This was then improved by Tropp to [39]

$$\epsilon_{N,\mathrm{OGA}} \leq (1 + 6\sqrt{N})^{1/2} \cdot \epsilon_{N,0}, \qquad N < 1/(8M).$$

These results show that, at least in its initial stages, the greedy algorithm performs quite well compared to the optimal algorithm.

The results we developed in this section have only indirect connection to this problem. Note that all our estimates concern errors such as $\|\alpha_0 - \hat{\alpha}_{\mathrm{OGA}}\|$ on the *representation* scale, rather than errors such as $\|x_0 - x_{\mathrm{OGA}}\|$, that are measured on the *reconstruction* scale. Nevertheless, in an incoherent dictionary, the two are connected for small N. It follows that the ideas in this paper are very similar to ideas underlying the N-term approximation results cited above.

VI. Excessive Pessimism?

The conditions for stability developed here are unduly restrictive. We used worst case reasoning exclusively, deriving conditions which must apply to *every* dictionary, *every* sparse representation, and *every* bounded noise vector. The bounds we have proven are consequently very loose and do not describe typical behavior; the sparsity conditions we have posed are much too strict. To illustrate this, we conducted numerous experiments to

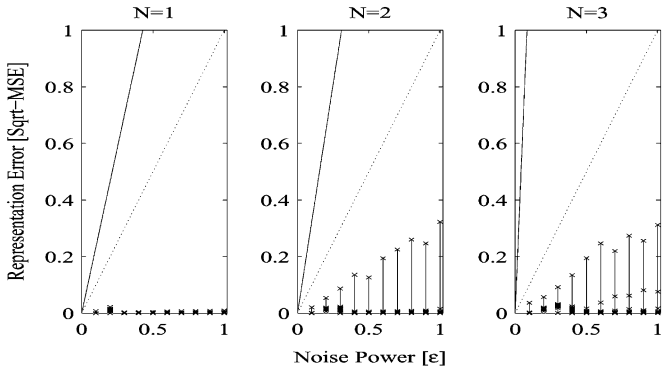

Fig. 1. ℓ^1 method, $\delta = \epsilon$: Representation error $\|\alpha_0 - \hat{\alpha}_{1,\epsilon}\|_2$ versus noise level ϵ. Solid lines depict the bound from Theorem 3.1. Different panels display results with support sizes $N = 1, 2, 3$.

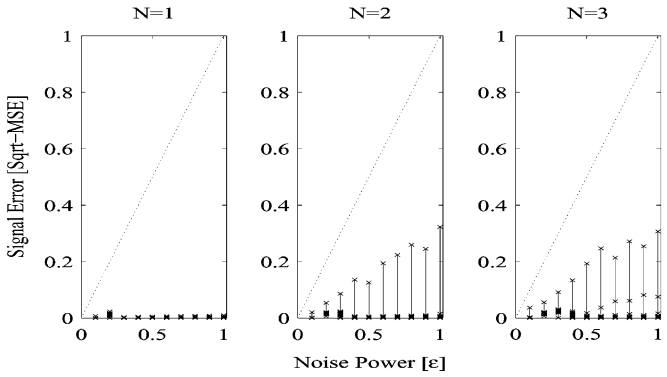

Fig. 2. ℓ^1 method, $\delta = \epsilon$: Signal error $\|\mathbf{\Phi}\alpha_0 - \mathbf{\Phi}\hat{\alpha}_{1,\epsilon}\|_2$ as a function of the noise level ϵ. Dotted lines indicate input noise level.

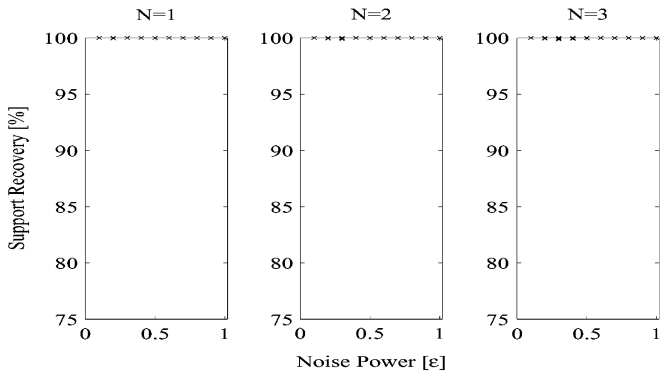

Fig. 3. ℓ^1 method, $\delta = \epsilon$: Support recovery success in percent.

study the stability of various algorithms in concrete cases. We present several representative results in this section.

We worked with a dictionary $\mathbf{\Phi} = [\boldsymbol{I}, \boldsymbol{H}]$, concatenating two orthonormal bases—the standard and Hadamard bases for signals of length $n = 128$ each, yielding $M = 1/\sqrt{128}$. We used randomly generated ideal representations α_0 satisfying the conditions of Theorem 3.1; since $(1 + M^{-1})/4 < 3.07$, we use $\|\alpha_0\|_0 = N = 1, 2, 3$. The nonzero entries of α_0 were located in uniform random positions, and the values of those entries were drawn from a normal distribution with zero mean and unit variance. The ideal noiseless signal $x_0 = \mathbf{\Phi}\alpha_0$ was normalized to have a unit ℓ^2-norm, so as to guarantee fixed signal-to-noise ratio (SNR) in our experiments. The signal x_0 was contaminated with zero-mean white Gaussian noise z, rescaled to enforce a specified noise level $\epsilon = \|z\|_2$, obtaining $y = \mathbf{\Phi}\alpha_0 + z$. The noise power values tested are $\epsilon = 0.1, 0.2, \ldots, 1.0$. We performed 1000 trials at each combination of N and ϵ—all together, 30 000 such realizations were generated, and to each we applied both $(P_{1,\delta})$ and OGA.

A. Experiments With ℓ^1 Penalization

We solved $(P_{1,\epsilon})$ numerically; thus, we assume the noise level ϵ is known and this knowledge is exploited in the recovery process. We used the IRLS algorithm [23] with a line search for the proper Lagrange multiplier. For each trial we calculated i) the representation error $\|\alpha_0 - \hat{\alpha}_{1,\delta}\|_2$, for comparison to theoretical bounds; ii) the denoising effect $\|\mathbf{\Phi}\alpha_0 - \mathbf{\Phi}\hat{\alpha}_{1,\delta}\|_2$, for comparison to the input noise power; and iii) the support match, measured as the relative energy in $\hat{\alpha}$ on the true support of α_0. This measure is robust against disagreements over the definition of "zero" entries in a numerical solution.

Fig. 1 presents the representation error results. For every ϵ value, the results are presented by plotting the tenths (11 points, starting from the minimum, and jumping by 100 sorted experiments' result, till the maximum). The results show a consistent stability with linearly growing error as a function of the noise level. The bounds of Theorem 3.1 are shown as solid lines, and the input noise amplitude is indicated by a dotted line. As expected, there is a large gap between the bound and the actual results. Accumulated marks near the axis imply that most of the experiments show stronger stability.

Fig. 2 presents the signal error results, showing the effect of denoising achieved. The results, organized similar to the ones in Fig. 1, show a very effective denoising.

Fig. 3 presents results on support recovery. The vertical axes describe the range [75, 100]%. As can be seen, $(P_{1,\epsilon})$ has nearly perfect success in recovering the support. Our tests probe the region beyond the range covered by Theorem 4.1. We have used $\delta = \epsilon$, rather than $\delta \approx 1.6\epsilon$, $\delta \approx 2\epsilon$, and $\delta \approx 3.1\epsilon$ for $N = 1, 2, 3$, respectively.

B. Experiments With Greedy Optimization

We now compare reconstruction errors for OGA with the above results, and with the bounds in Theorem 5.1 (inequality (5.4)). Paralleling the experiment for ℓ^1 with $\delta = \epsilon$, we display theoretical bounds and empirical errors in Fig. 4. Evidently, OGA behaves stably with results somewhat weaker than those obtained by the ℓ^1 penalization. Again, the upper bounds provided in Theorem 5.1 are seen to be exactly that—overestimates of the reconstruction error. Note that in this experiment we have used $N = 1, 2, 3$, totally disregarding the condition as posed in (5.2), tying the allowed sparsity to the coefficients' amplitude. Yet, the results show stability.

Fig. 5 shows the denoising effect, and as can be seen, the results resemble those found in the representation errors. Fig. 6 presents the support recovery success rates. Those are weaker than in the ℓ^1 case; as the noise strengthens, we violate the condition (5.2), and eventually see breakdown in the support recovery.

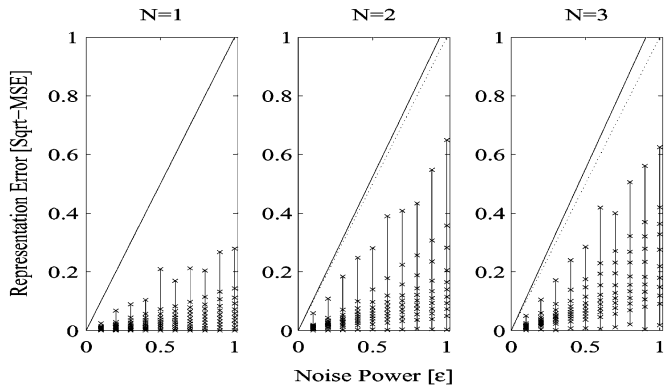

Fig. 4. OGA: Representation error $\|\alpha_0 - \hat{\alpha}_{\mathrm{OGA}}\|_2$ versus noise level ϵ. Solid lines depict bounds in Theorem 5.1.

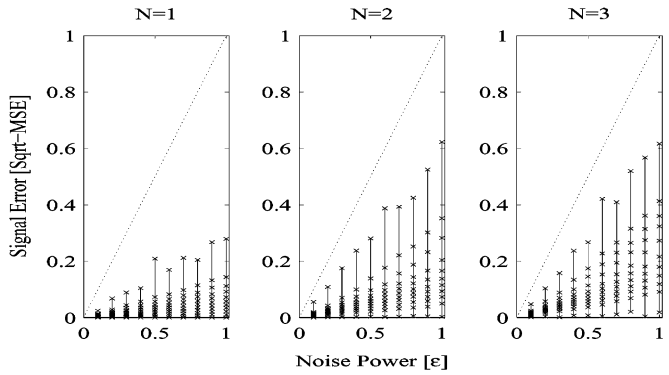

Fig. 5. OGA: Signal error $\|\mathbf{\Phi}\alpha_0 - \mathbf{\Phi}\hat{\alpha}_{1,\epsilon}\|_2$ versus noise level ϵ. Dotted lines depict input noise level. Different panels display results for support sizes $N = 1, 2, 3$.

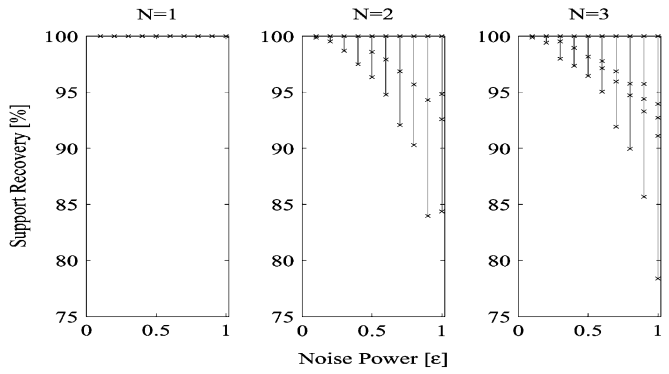

Fig. 6. OGA: Support recovery success in percent.

C. Geometric Heuristics

As indicated above, our very general reasoning is to blame for the looseness of our theoretical bounds; by developing bounds valid for a wide range of dictionaries and a wide range of sparsely represented signals, we are forced to consider the behavior for the worst possible combination of dictionary, signal, and noise.

We might get tighter results, by developing tools adapted to each specific $(\mathbf{\Phi}, \alpha_0)$ combination. Unfortunately, the closer we get to case-by-case analysis, the more difficult it becomes to get an intellectually digestible overview of the situation. At least for ℓ^1 minimization, it seems clear to the authors that, even at values

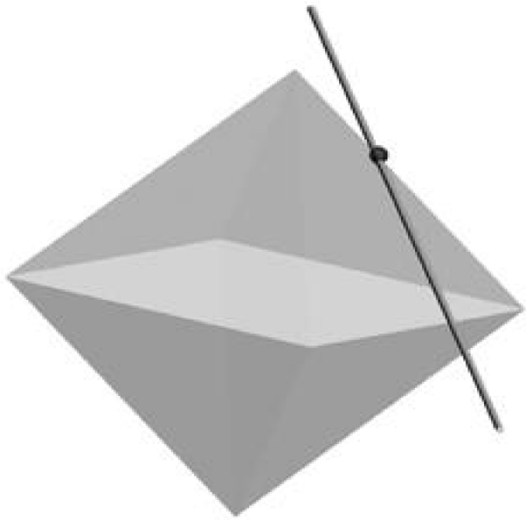

Fig. 7. Geometry favorable to unique ℓ^1 decomposition. Intersection of A_{x_0} with $B_1(R_0)$ in a unique point. This point is the unique solution $\hat{\alpha}_{1,0}$.

N far greater than those covered in Theorem 3.1 the following will generally be true.

- A sparse vector α_0 generating $x_0 = \mathbf{\Phi}\alpha_0$ will be the unique solution of $(P_{1,0})$ and
- the solution of $(P_{1,\epsilon})$ based on noisy data $y = x_0 + z$ with noise level ϵ will stably recover α_0.

It is less clear to us that we can expect the solution to the ℓ^1 problem to agree with the solution to the ℓ^0 problem with the same degree of generality.

Some insight may be gleaned by considering the geometry of minimal ℓ^1 decomposition; see Fig. 7 below. The minimal ℓ^1 decomposition in an overcomplete system is the point in the subspace $A_{x_0} = \{\alpha : x_0 = \mathbf{\Phi}\alpha\}$ having the smallest ℓ^1 norm. Denote this norm by $R_0 = \mathrm{val}(P_{1,0})$. Alternatively, if we consider the collection of balls $B_1(R) = \{\alpha : \|\alpha\|_1 \leq R\}$ in R^m, it is the "first point in A_{x_0}" to "meet" the family of balls as R grows from 0 to R_0. When this meeting occurs, if it is in a unique point, then the ℓ^1 decomposition is unique. Now note that if α_0 has few nonzeros, then it sits in a low-dimensional face of $B_1(R_0)$. Denote by F_{α_0} the smallest dimensional face of $B_1(R_0)$ containing α_0 in its interior.

Fig. 7 shows clearly a situation where F_{α_0} is *transversal* to A_{x_0}—the two subspaces meet nicely in a single point. More than this: all the faces of $B_1(R_0)$ touching F_{α_0} intersect A_{x_0} tranversally. Now the cleanness of these intersections imply that α_0 is the unique ℓ^1 minimizer in A_{x_0}.

A key observation is that the faces of the ball $B_1(R_0)$ run through a finite list of specific orientations. If we take a *generic* $\mathbf{\Phi}$, there would never be a fortuitous alignment of any subspace A_{x_0} with any of the low-dimensional faces of $B_1(R_0)$; hence, transversal intersections should be generic, and we can expect to have unique ℓ^1 minimizers *except* when $\|\alpha_0\|_0$ and $\dim A_{x_0}$ demand nonuniqueness.

What about stability? A geometric explanation of stability for $\hat{\alpha}_{1,\epsilon}$ is illustrated in Fig. 8. Because of $\|\hat{\alpha}_{1,\epsilon}\|_1 \leq \|\alpha_0\|_1$, $\hat{\alpha}_{1,\epsilon}$ must belong to the cone C_{1,α_0} with vertex at α_0 consisting of all points α such that for small $t > 0$, $\|(1-t)\alpha + t\alpha_0\|_1 \leq \|\alpha_0\|_1$. On the other hand, because of $\|x_0 - \mathbf{\Phi}\hat{\alpha}_{1,\epsilon}\|_2 \leq 2\epsilon$, $\hat{\alpha}_{1,\epsilon}$ must

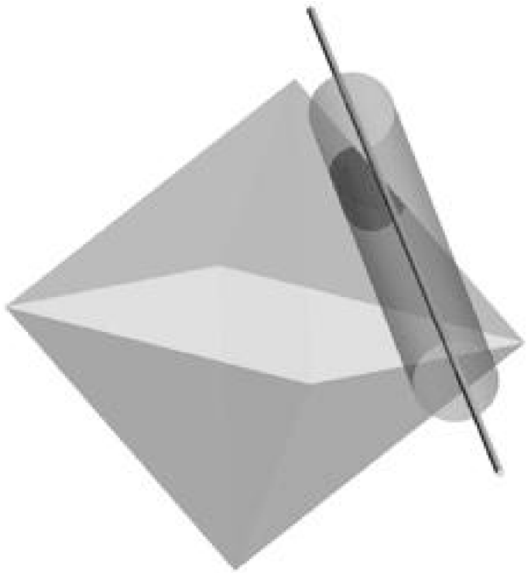

Fig. 8. Geometry favorable to stable ℓ^1 decomposition. Intersection of $A_{x_0,2\epsilon}$ with $B_1(R_0)$ in a tubular wedge. $\hat{\alpha}_{1,\epsilon}$ must lie in wedge. Small size of the wedge indicates stability.

belong to the cylinder $A_{x_0,2\epsilon}$ consisting of all vectors α obeying $\|x_0 - \mathbf{\Phi}\alpha\|_2 \leq 2\epsilon$. In short, for a cone C_{1,α_0} and a cylinder $A_{x_0,2\epsilon}$, we have

$$\hat{\alpha}_{1,\epsilon} \in \{C_{1,\alpha_0} \cap A_{x_0,2\epsilon}\}.$$

Roughly speaking, the size of this intersection is controlled by the angle between A_{x_0} and C_{1,α_0}. That this angle can be positive we know already; because that is the content of the transversality we have already discussed.

There is an analytical framework to quantify the above heuristic notions. There is a stability estimate adapted to a specific $(\mathbf{\Phi}, \alpha_0)$ pair

$$\|\hat{\alpha}_{1,\epsilon} - \alpha_0\| \leq \Lambda_1(\mathbf{\Phi}, \alpha_0) \cdot 2\epsilon, \qquad \epsilon > 0$$

where

$$\Lambda_1(\mathbf{\Phi}, \alpha_0) = \sup\left\{ \frac{\|\gamma - \alpha_0\|}{\|\mathbf{\Phi}(\gamma - \alpha_0)\|} : \|\gamma\|_1 \leq \|\alpha_0\|_1 \right\}.$$

Equivalently

$$\Lambda_1(\mathbf{\Phi}, \alpha_0) = \sup\left\{ \frac{\|v\|}{\|\mathbf{\Phi}v\|} : v \in \dot{C} \right\}$$

where $\dot{C}$ denotes the tangent cone to C_{1,α_0} at α_0, i.e., the collection of vectors v such that $\alpha_0 + tv \in C_{1,\alpha_0}$ for all sufficiently small $t > 0$.

This last display makes the point that we are trying to optimize a ratio of quadratic forms subject to membership in a cone. This makes us say that Λ_1 is akin to the secant of the angle between A_{x_0} and C_{1,α_0}. Unfortunately, to our knowledge, the problem of finding the angle between a cone and a subspace does not have a convenient computational solution. Hence, although the bound depends intimately on $\mathbf{\Phi}$ and α_0, we know of no way to easily compute this dependence at the moment.

D. Domain of Applicability

An *apparent* application of the results of this paper concerns the problem of resolving a spectrum at a resolution finer than the usual Rayleigh spacing. As a simple model, we could consider the complex-valued dictionary with atoms

$$\phi_i(k) \propto \exp\left\{ \sqrt{-1}\frac{2\pi}{\nu n} i(k-1) \right\}, \qquad k = 1, \ldots, n, \;\; i = 1, \ldots, \nu n.$$

Here ν, an integer > 1 is the *superresolution* factor, and the implicit constant of proportionality is chosen to enforce the normalization $\|\phi_i\|_2 = 1$. In this overcomplete system, the frequencies are spaced $\frac{2\pi}{\nu n}$ apart, which is ν times as closely as the usual spacing $\frac{2\pi}{n}$ of the Fourier frequencies, hence, the term superresolution. If we simply chose $\nu = 1$, we would have an orthogonal dictionary. If we choose $\nu = 2$, we get an overcomplete system with $m = 2n$. It would be very attractive to be able to solve this problem, getting finer frequency resolution out of a given signal length. However, Wohlberg [43] showed that in general this problem will lead to extreme ill-posedness, even under sparsity constraints.

We remark that, while superresolution is an attractive and important problem, this is *not* the setting we envisioned for applying our results. In the superresolving case, with $\nu = 2$, the dictionary has mutual coherence $M = M_n = \frac{1}{n \sin(\pi/n)}$, which tends to π^{-1} as n increases. This is quite large, and the dictionary is coherent rather than incoherent. It yields the sparsity threshold $(1 + M^{-1})/2 \approx 2.07$ which allows to disentangle at most two atoms, at any n!

The kind of situation we have in mind for applying our results is quite different; we are interested in cases where the mutual incoherence is comparable, for large n, to some power $n^{-\beta}$, so that, at least for large n, there is the potential to disentangle fairly complex superpositions of many atoms. Previous work has given several examples of this type of situation: random dictionaries [11], Grassmannian frames [33], and dictionaries for $3 - D$ voxel data made of digital points, lines, and planes [11].

For those interested in supperresolution, we remark that, in our opinion, the analysis in [43] adopts a framework which is unduly pessimistic. The careful theoretical work on superresolution [10] explains that stable superresolution is possible using sparsity; however, the notion of sparsity needs to be *adapted* to the setting. Specifically, it becomes important to define sparsity in terms of "number of nonzeros per Rayleigh interval" rather than simply "number of nonzeros." When this definitional convention is adopted, it is possible to prove that sufficient sparsity again enables superresolution, in agreement with a considerable body of empirical work, also cited in [10].

ACKNOWLEDGMENT

The authors would like to thank Matthew Ahlert and Inam Ur Rahman for help in preparing the figures.

REFERENCES

[1] A. P. Berg and W. B. Mikhael, "A survey of mixed transform techniques for speech and image coding," in *Proc. 1999 IEEE Int. Symp. Circuits and Systems*, vol. 4, Orlando, FL, Jun. 1999, pp. 106–109.

[2] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and the problem of approximating piecewise C^2 images with piecewise C^2 edges," *Commun. Pure Appl. Math.*, vol. 57, pp. 219–266, Feb. 2004.

[3] S. S. Chen, "Basis Pursuit," Ph.D. dissertation, Stanford Univ., Stanford, CA, Nov. 1995.

[4] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[5] R. Coifman, Y. Meyer, and V. Wickerhauser, "Adapted wave form analysis, wavelet-packets and applications," in *Proc. 2nd Int. Conf. Industrial and Applied Mathematics (ICIAM 91)*, Philadelphia, PA, 1992, pp. 41–50.

[6] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 374–377, Mar. 2002.

[7] C. Daniel and F. S. Wood, *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd ed. New York: Wiley, 1980.

[8] V. E. DeBrunner, L. X. Chen, and H. J. Li, "Lapped multiple basis algorithms for still image compression without blocking effect," *IEEE Trans. Image Process.*, vol. 6, pp. 1316–1322, Sep. 1997.

[9] R. A. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 12, pp. 213–227, 1996.

[10] D. L. Donoho, "Superresolution via sparsity constraints," *SIAM J. Math. Anal.*, vol. 23, pp. 1309–1331, 1992.

[11] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," in *Proc. Nat. Acad. Sci.*, vol. 100, 2003, pp. 2197–2202.

[12] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise," Dep. Statistics, Stanford Univ., http://www-stat.stanford.edu/~donoho/Reports/2004/StableSparse.pdf, Tech. Rep., 2004.

[13] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[14] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of $\Re^N$ bases," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep. 2002.

[15] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.

[16] ——, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inf. Theory*, submitted for publication.

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

[18] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[19] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[20] T. Hastie, R. Tibshirani, and J. H. Friedman, *Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[21] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1985.

[22] X. Huo, *Sparse Image Representation via Combined Transforms*. Stanford, CA: Stanford Univ., Nov., 1999.

[23] L. A. Karlovitz, "Construction of nearest points in the ℓ^p, p even and ℓ^1 norms," *J. Approx.Theory*, vol. 3, pp. 123–127, 1970.

[24] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Alg. Its Applic.*, vol. 18, no. 2, pp. 95–138, 1977.

[25] T. G. Kolda, "Orthogonal tensor decompositions," *SIAM J. Matrix Anal. Its Applic.*, vol. 23, no. 1, pp. 243–255, 2001.

[26] X. Liu and N. D. Sidiropoulos, "Cramer-Rao lower bounds for low-rank decomposition of multidimensional arrays," *IEEE Trans. Signal Process.*, vol. 49, no. 9, pp. 2074–2086, Sep. 2001.

[27] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[28] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Res.*, vol. 37, pp. 311–325, 1997.

[29] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993, pp. 40–44.

[30] S. Qian and D. Chen, "Signal representation using adaptive normalized Gaussian functions," *Signal Process.*, vol. 36, pp. 1–11, 1994.

[31] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition: Separation of texture from piecewise smooth content," presented at the SPIE Meeting, San-Diego, CA, Aug. 2003.

[32] J.-L. Starck, E. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 131–141, Jun. 2002.

[33] T. Strohmer and R. Heath Jr, "Grassmannian frames with applications to coding and communications," *Appl. Comp. Harm. Anal.*, vol. 14, pp. 257–275, May 2003.

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B.*, vol. 58, no. 1, pp. 267–288, 1996.

[35] V. N. Temlyakov, "Greedy algorithms and m-term approximation," *J. Approx. Theory*, vol. 98, pp. 117–145, 1999.

[36] V. N. Temlyakov, "Weak greedy algorithms," *Adv. Comput. Math.*, vol. 5, pp. 173–187, 2000.

[37] ——, "Greedy algorithms with regard to multivariate systems with special structure," *Constr. Approx.*, vol. 16, pp. 399–425, 2000.

[38] ——, "Nonlinear methods of approximation," *Found. Comput. Math.*, vol. 3, pp. 33–107, 2003.

[39] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[40] ——, "Just relax: Convex programming methods for subset selection and sparse approximation," *IEEE Trans. Inf. Theory*, submitted for publication.

[41] R. Venkataramani and Y. Bresler, "Optimal sub-nyquist nonuniform sampling and reconstruction of multiband signals," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2301–2313, Oct. 2001.

[42] M. V. Wickerhauser, *Adapted Wavelet Analysis From Theory to Software*. Wellesley, MA: A K Peters Ltd., 1994.

[43] B. Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *IEEE Trans. Signal Process.*, vol. 51, no. 12, pp. 3053–3060, Dec. 2003.