PROBABILISTIC ANALYSIS FOR BASIS SELECTION VIA ℓ_p DIVERSITY MEASURES

David P. Wipf and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407 USA
e-mail: dwipf@ucsd.edu, brao@ece.ucsd.edu

ABSTRACT

Finding sparse representations of signals is an important problem in many application domains. Unfortunately, when the signal dictionary is overcomplete, finding the sparsest representation is NP-hard without some prior knowledge of the solution. However, suppose that we have access to such information. Is it possible to demonstrate any performance bounds in this restricted setting? Herein, we will examine this question with respect to algorithms that minimize general ℓ_p-norm-like diversity measures. Using randomized dictionaries, we will analyze performance probabilistically under two conditions. First, when $0 \leq p < 1$, we will quantify (almost surely) the number and quality of every local minimum. Next, for the $p = 1$ case we will extend the deterministic results of Donoho and Elad (2003) by deriving explicit confidence intervals for a theoretical equivalence bound, under which the minimum ℓ_1-norm solution is guaranteed to equal the maximally sparse solution. These results elucidate our previous empirical studies applying ℓ_p measures to basis selection tasks.

1. INTRODUCTION

Sparse signal representations from overcomplete dictionaries find increasing relevance in a large number of application domains [1, 2, 3]. The canonical form of this problem is given by,

$$\boldsymbol{t} = \Phi \boldsymbol{w} + \boldsymbol{\epsilon}, \tag{1}$$

where $\Phi \in \Re^{N \times M}$ is a matrix whose columns represent a possibly overcomplete basis ($M > N$), $\boldsymbol{w}$ is the vector of weights to be learned, $\boldsymbol{\epsilon}$ is noise, and $\boldsymbol{t}$ is a signal vector. In this vein, we seek to find weight vectors whose entries are predominantly zero (i.e., small in the ℓ_0-norm sense) that nonetheless allow us to accurately represent $\boldsymbol{t}$.

Recently, a large body of important theoretical work has addressed problems like (1) in the low-noise limit, i.e., as $\boldsymbol{\epsilon} \rightarrow 0$. For example, Basis Pursuit finds the weight vector satisfying $\boldsymbol{t} = \Phi \boldsymbol{w}$ with minimum ℓ_1-norm via linear programming [2]. More generally, ℓ_p-norm-like diversity measures have been proposed (see e.g., [3, 4]) to find sparse solutions by solving

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} d_p(\boldsymbol{w}) \triangleq \sum_{i=1}^{M} |w_i|^p, \qquad \text{s.t. } \boldsymbol{t} = \Phi \boldsymbol{w} \tag{2}$$

with $p \in [0, 1]$. Note that this formulation encompasses Basis Pursuit as $p \rightarrow 1$, whereas when $p \rightarrow 0$, we are effectively minimizing the ℓ_0-norm directly. A descent algorithm called FOCUSS minimizes (2), at least locally, for all values of p [4].

Ideally, we would like to use $d_0(\boldsymbol{w}) = \|\boldsymbol{w}\|_0$; however, it is well known that the resultant minimization problem is NP-hard without some prior knowledge of the optimal solution [4]. Consequently, the FOCUSS algorithm, with $p = 0$, is susceptible to suboptimal local minima. In contrast, no local minima exist when $d_1(\boldsymbol{w})$ is used, although now the global minimum need not correspond with the maximally sparse solution to $\boldsymbol{t} = \Phi \boldsymbol{w}$.

While difficult in a general setting, if information is available pertaining to the optimal solution, the potential exists for establishing some bounds on the performance we can expect. Using randomized dictionaries, we will analyze performance probabilistically under two conditions. First, when $0 \leq p < 1$, we will quantify (almost surely) the number and quality of every local minimum. Specifically, we will show that the number of local minima must equal the number of sparse solutions and that for each local minimum $\boldsymbol{w}$, $d_0(\boldsymbol{w}) = N$. (Note: A *sparse solution* is formally defined as a solution with N or fewer nonzero entries, i.e., $d_0(\boldsymbol{w}) \leq N$).

Secondly, when $p = 1$, a theoretical *equivalence* bound has been established in [5] whereby the globally minimum $d_1(\boldsymbol{w})$ solution (easy to find) is equivalent to the globally minimum $d_0(\boldsymbol{w})$ solution (difficult to find although ultimately what we want). More concretely, this bound stipulates that if the (unknown) optimal solution $\boldsymbol{w}_0$ satisfies $d_0(\boldsymbol{w}_0) \triangleq D_0 < b$, then Basis Pursuit will always converge to $\boldsymbol{w}_0$; the bound b is dependent on the dictionary Φ. We expand on this result by deriving explicit confidence intervals for this bound in the case where Φ is a Gaussian random dictionary. This allows us to probabilistically evaluate the performance potential afforded by this bound.

Before we begin, we should mention that randomized dictionaries are of particular interest in signal processing and other disciplines [5, 6, 1]. Moreover, basis vectors from many real world measurements can often be modelled as random. In any event, randomized dictionaries capture a wide range of phenomena and therefore represent a viable benchmark for testing basis selection methods. At least we would not generally expect an algorithm to perform well with a random dictionary and poorly on everything else. Moreover, as we shall soon see, such dictionaries lend themselves to probabilistic analyses that suggest useful prescriptions for choosing a suitable p.

This research was supported by DiMI grant #22-8376 and Nissan.

2. PERFORMANCE ANALYSIS WITH $p < 1$

In this section, we address the issue of finding performance guarantees when $p < 1$. Local minima pose a clear impediment to achieving this goal, and therefore, quantifying the number and extent of such minima is crucial. While strictly deterministic results may be evasive in general, we will now address this issue probabilistically. To facilitate this goal, we first present the following result (see Appendix for proof):

Lemma 1 *If Φ satisfies the unique representation property (URP), i.e., every subset of N columns are linearly independent, then the set of sparse solutions of (2) with $p \in [0, 1)$ is equal to the set of local minima.*

Assuming the URP holds (as is the case almost surely for dictionaries formed from iid elements drawn from a continuous, bounded probability density), we can conclude from Lemma 1 that we need only determine the number of sparse solutions in counting local minima. Additionally, if we assume there exists only a single degenerate sparse solution (i.e., a single solution with $d_0(\boldsymbol{w}) < N$), then this solution is by definition the maximally sparse solution $\boldsymbol{w}_0$. Under these circumstances, it is a simple matter to show that the total number of sparse solutions, $\mathcal{L}$, is given by $\binom{M}{N} - \binom{M-D_0}{N-D_0} + 1$. But how do we know that our assumption of a single degenerate sparse solution is valid? The following Lemma addresses this question (see Appendix for proof):

Lemma 2 *Let $\Phi \in \Re^{N \times M}, M > N$ be constructed such that it satisfies the URP. Additionally, let $\boldsymbol{t} \in \Re^N$ satisfy $\boldsymbol{t} = \Phi \boldsymbol{w}_0$ for some $\boldsymbol{w}_0$ such that $d_0(\boldsymbol{w}_0) \triangleq D_0 < N$, with non-zero entries of $\boldsymbol{w}_0$ drawn independently and identically from a continuous, bounded density. Then there is almost surely no other solution $\boldsymbol{w} \neq \boldsymbol{w}_0$ such that $\boldsymbol{t} = \Phi \boldsymbol{w}$ and $d_0(\boldsymbol{w}) = D < N$.*

Given that the conditions of Lemma 2 are satisfied, we may then conclude that,

$$\mathrm{P}\left[\mathcal{L} = \binom{M}{N} - \binom{M - D_0}{N - D_0} + 1\right] = 1. \tag{3}$$

How does this result affect our performance analysis? For an arbitrary initialization $\boldsymbol{w}$ and $M > N$ (i.e., Φ is overcomplete) , we cannot guarantee (i.e., with probability one) that the FOCUSS algorithm (or any other descent method) will avoid converging to one of these local minima. Moreover, even if we allow for reinitializations (a method that has been shown to be extremely successful in [4]), an exhaustive search of these extrema to find the global solution is not feasible. In fact, the only consolation we can provide is in the special case where $D_0 = 1$. Although local minima still exist under these circumstances, if we initialize FOCUSS at the ℓ_2-norm solution, it will typically always converge to the global minima. In all other cases (i.e., $D_0 > 1$), there is always some potential that we will converge to one of the $\mathcal{L} - 1$ suboptimal local minima, each with suboptimal diversity $d_0(\boldsymbol{w}) = N$.

3. PERFORMANCE ANALYSIS WHEN $p = 1$

As previously mentioned, when $p = 1$, a single minimum exists that may or may not correspond with the maximally sparse solution $\boldsymbol{w}_0$. However, in [5], a substantial result is derived that dictates when the minimum $d_1(\boldsymbol{w})$ solution is sufficient.

Theorem 1 *(Equivalence Theorem [5]) Given an arbitrary dictionary Φ with columns $\boldsymbol{\phi}_i$ normalized such that $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1, \forall i = 1, \ldots, M$, and given $G \triangleq \Phi^T \Phi$ and $\kappa \triangleq \max_{i \neq j} |G_{i,j}|$, if the sparsest representation of a signal by $\boldsymbol{t} = \Phi \boldsymbol{w}_0$ satisfies $d_0(\boldsymbol{w}_0) \leq b \triangleq 1/2(1 + 1/\kappa)$, then the Basis Pursuit solution (which minimizes the $p = 1$ case) is guaranteed to equal $\boldsymbol{w}_0$.*

This is a potentially powerful result since it specifies a computable condition by which the minimum $d_1(\boldsymbol{w})$ solution is guaranteed to produce $\boldsymbol{w}_0$. Therefore, it behooves us to determine if Basis Pursuit and its attendant theoretical bound possess significant superiority over the $p < 1$ case by analyzing this bound in practical problems of interest. As a step in this direction, we will derive confidence intervals for equivalence bound b in the case where the elements of Φ are zero-mean, iid Gaussian random variables. This handles the empirical studies in [4, 8] and approximates many other relevant situations. Moreover, we may potentially extend these results to handle other randomized dictionary types.

Ideally, for a given N,M and confidence level α, we would like to find a critical value C such that $\mathrm{P}(b < C) = 1 - \alpha$. Such information is extremely useful in assessing the applicability of Theorem 1 when b is a random variable. For example, suppose we find that $C = 4.7$ for some values of N, M and $\alpha = 0.05$. This implies that there is a 95% chance that the bound b will be less than 4.7. Therefore, if for some reason we expect that $d_0(\boldsymbol{w}_0) \geq 5 > C$, then there is a very low probability that Theorem 1 can apply.

To form confidence intervals, we need to know something about the distribution of b. We begin the analysis by noting that the Gramian matrix G is proportional to the sample covariance (with a known mean of zero) of the columns of Φ. Furthermore, since we have used the ℓ_2-norm to standardize the columns, we see that G represents the exact sample correlation matrix of Φ, i.e., $G_{i,j} = \mathrm{corr}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = \rho_{i,j}$. The distribution of each element $\rho_{i,j}$ is not very accessible; however, when the mean is known

$$\tau_{i,j} = g(\rho_{i,j}) \triangleq \frac{\rho_{i,j}\sqrt{\mathrm{dof}}}{\sqrt{1 - \rho_{i,j}^2}} \tag{4}$$

has a t-distribution with $\mathrm{dof} = N - 1$ degrees of freedom.

To use these results for the purpose at hand, we note that the following holds for any constant C such that $p(b < C) = 1 - \alpha$:

$$\begin{aligned} \mathrm{P}(b < C) &= \mathrm{P}\left(\kappa > \frac{1}{1 - 2C}\right) \\ &= \mathrm{P}\left(\max_{i \neq j} |\rho_{i,j}| > \frac{1}{1 - 2C}\right) \\ &= \mathrm{P}\left(\max_{i \neq j} |\tau_{i,j}| > g\left[\frac{1}{1 - 2C}\right]\right), \end{aligned} \tag{5}$$

where the first step follows from Theorem 1, which defines b as a monotonically decreasing function of κ. The third step follows since $g(\cdot)$ is a monotonically increasing function. By defining $C' \triangleq g(\frac{1}{1-2C})$, we can find a $1-\alpha$ confidence interval by solving

$$\mathrm{P}\left(\max_{i \neq j} |\tau_{i,j}| < C'\right) = \alpha \tag{6}$$

for C' and then inverting the definition of C' to find C. So how do we find the critical value C'?

To begin, we note that there are $n \triangleq (M^2 - M)/2$ unique off-diagonal terms in G and therefore, n terms $\tau_{i,j}$ over which we are taking a maximum. By construction, these elements are identically distributed; however, they are unfortunately not all independent of one another (the positive-definiteness of G dictates that there must be some dependency between elements $\rho_{i,j}$ and therefore between elements $\tau_{i,j}$). There are two ways to address this issue: (I) find a strict (non-approximate) upper-bound on the true critical value or (II), adopt an approximation that, based on Monte-Carlo verification, is extremely good. We address these options in turn.

3.1. Method I

Suppose it is sufficient that we find a critical value $\bar{C}$ (via some $\bar{C}'$) that represents a *strict* (non-approximate) upper bound on the true critical value,i.e., such that,

$$\mathrm{P}(b < \bar{C}) > \mathrm{P}(b < C) = 1 - \alpha. \tag{7}$$

We can accomplish this goal by only using the $M/2$ elements $\tau_{i,j}$ that are in fact exactly independent, e.g., $\tau_{1,2}$, $\tau_{3,4}$, ..., $\tau_{M-1,M}$ (assuming M is even), and ignoring all other terms. We note that these $M/2$ terms are independent since they each use unique, independently distributed columns of Φ. Now we are employing a maximization over fewer elements and therefore,

$$\max_{i=1,3,5,\ldots} |\tau_{i,i+1}| < \max_{i \neq j} |\tau_{i,j}|. \tag{8}$$

which implies that $\bar{C} > C$ (since $\bar{C}' < C'$) or equivalently, that $\mathrm{P}(b < \bar{C}) > 1 - \alpha$ as desired. Since we are now dealing with iid random variables, we can find $\bar{C}'$ by solving

$$\mathrm{P}\left(\max_{i=1,3,5,\ldots} |\tau_{i,j}| < \bar{C}'\right) = \mathrm{P}\left(|\tau_{1,2}| < \bar{C}'\right)^{M/2} = \alpha, \tag{9}$$

which is equivalent to solving

$$\mathrm{P}\left(\tau_{1,2} > \bar{C}'\right) = \frac{1 - \alpha^{2/M}}{2}. \tag{10}$$

The solution to (10) is given by $\bar{C}' = T_{N-1}^{-1}[(1-\alpha^{2/M})/2]$, where $T_{N-1}^{-1}(\cdot)$ denotes the inverse t-distribution with $N-1$ degrees of freedom. We may then compute $\bar{C}$ by inverting $\bar{C}' \triangleq g(\frac{1}{1-2\bar{C}})$.

The utility of these results is that we are now able to determine areas of N, M-space where, with high probability, the Basis Pursuit equivalence bound cannot apply. For example, we used this procedure to compute confidence intervals for various values of N, M as shown in Figure 1. In general, we have found that for practical values of these variables the equivalence bound tends to be prohibitively tight, i.e., there is a high probability that b will be too small. For example, simulation studies were performed in [4, 8] comparing basis selection efficacy using $p = 1$ versus other methods. In these experiments, $N = 20$, $M \in [30, 100]$, and the optimal solution $\boldsymbol{w}_0$ satisfied $d_0(\boldsymbol{w}_0) = 4$ or 7. Over the course of numerous trials, the $p = 1$ solution sometimes failed to equal $\boldsymbol{w}_0$. We can now reconcile these results with Theorem 1 by noting that, with high probability, $b < 4.0$ (when $N = 20$ and $M \in [30, 100]$) and therefore, the equivalence result is not applicable.

In general, use of $\bar{C}$ establishes a (probably) necessary but not sufficient condition for applying Theorem 1 for Gaussian random dictionaries: basically, if $d_0(\boldsymbol{w}_0) > \bar{C}$ as defined above, then there is very little probability that b will be greater than $d_0(\boldsymbol{w}_0)$ and we cannot apply the theorem. For (probably) sufficient conditions, we must turn to the second method.

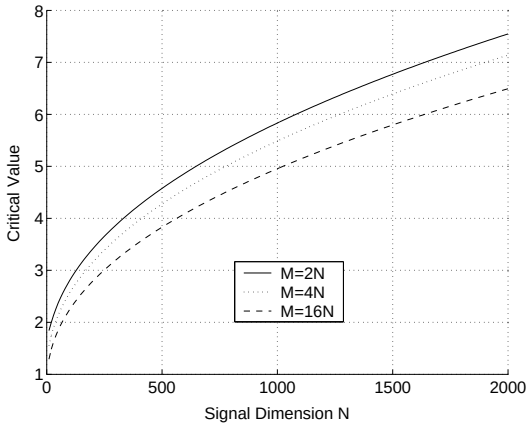

Fig. 1. $(95+)\%$ *critical values for equivalence bound* b. For a given value of N, M, there is a *greater* than 95% chance that b will be *less* than the corresponding critical value $\bar{C}$.

3.2. Method II

Suppose that we simply assume that all n off-diagonal terms of G are independent (actually, we only need to assume that the distribution of $\max_{i \neq j} |\tau_{i,j}|$ is the same as if we had such independence). We may then repeat the above analysis with $M/2$ replaced by n, the net result of which is to push each of the curves in Figure 1 down towards the x-axis. Now suppose we want to utilize this method to find a (probably) sufficient condition for using Theorem 1. We can accomplish this by setting α arbitrarily high, e.g., $\alpha = 0.95$. This allows us to find critical values C such that, with high probability, b will be *greater* than C or equivalently, with high probability the equivalence result will apply when $d_0(\boldsymbol{w}_0) < C$.

In Table 1, we have tabulated five approximate critical values obtained in this manner. We have also explored this approximation via numerous Monte Carlo simulations (e.g., using many different combinations of N,M and α, five of which are shown below), which have revealed that the critical values so obtained are statistically indistinguishable from the exact critical values, indicating that our independence assumption is quite benign. Thus, we have a simple, accurate way of computing confidence intervals that does *not* require MC techniques, which become too computationally expensive when N and M are much greater than 200.

N, M	10,40	20,40	20,80	100,200	200,400
C	1.0444	1.1586	1.1285	1.6152	1.9596
	(4.9%)	(5.1%)	(5.0%)	(4.9%)	(5.0%)

Table 1. *Approximate 5% critical values for equivalence bound* b. For a given value of N, M, there is approximately a 5% chance that b will be *less* than the corresponding critical value. The number in parenthesis represents the percentage of Monte-Carlo trials (out of $100,000$) that are less than the corresponding critical value.

4. CONCLUSIONS

In this paper we have addressed the issue of performance bounds when using ℓ_p measures to find sparse representations from over-

complete dictionaries. We have employed a probabilistic line in handling both the $p < 1$ case, where we quantified the number and quality of each local minima, and the $p = 1$ case, where somewhat stronger results were possible in the form of equivalence bound confidence intervals, under which the minimum $p = 1$ solution, with high probability, is guaranteed to equal the maximally sparse solution. However, in cases where this bound is prohibitively tight (which we can now easily determine for Gaussian dictionaries using the methods outlined in Sec. 3), use of $p < 1$ may be preferred for two primary reasons. First, if we have converged to some minimum and $d_0(\boldsymbol{w}) < N$, we know it must be the global minimum (assuming the conditions of Lemma 2 are met). Conversely, if we have converged to a solution with $d_0(\boldsymbol{w}) = N$, then we can always reinitialize and try again, a technique that has proven to be extremely successful in [4] (we note that this option is not possible when $p = 1$ since there is only a single, global minima in this case). Secondly, use of smaller values of p results in faster convergence via interior point methods as also shown in [4].

5. APPENDIX

Proof of Lemma 1: That local minima are only achieved at sparse solutions has been shown in [4]. We will now handle the converse. A vector $\boldsymbol{w}^*$ is a constrained local minimizer of $d_p(\boldsymbol{w})$ (s.t. $\boldsymbol{t} = \Phi\boldsymbol{w}$) if for every vector $\boldsymbol{w}' \in \text{Null}(\Phi)$, there is a $\delta > 0$ such that

$$d_p(\boldsymbol{w}^*) < d_p(\boldsymbol{w}^* + \varepsilon\boldsymbol{w}') \quad \forall \varepsilon \in (0, \delta]. \tag{11}$$

We will now show that all sparse solutions satisfy this condition. We first handle the case where $p > 0$ by defining $g(\varepsilon) \triangleq d_p(\boldsymbol{w}^* + \varepsilon\boldsymbol{w}')$ and then computing the gradient of $g(\varepsilon)$ at a feasible point in the neighborhood of $g(0) = d_p(\boldsymbol{w}^*)$. We then note that at any feasible point $\boldsymbol{w} = \boldsymbol{w}^* + \varepsilon\boldsymbol{w}'$ we have

$$\begin{aligned}\frac{\partial g(\varepsilon)}{\partial \varepsilon} &= \frac{\partial d_p(\boldsymbol{w})}{\partial \boldsymbol{w}}^T \frac{\partial \boldsymbol{w}}{\partial \varepsilon} = \sum_{i=1}^{M} \frac{\partial d_p(\boldsymbol{w})}{\partial (w_i)} w_i' \\ &= \sum_{i=1}^{M} \text{sgn}(w_i^* + \varepsilon w_i') p |w_i^* + \varepsilon w_i'|^{p-1} w_i'. \end{aligned} \tag{12}$$

Since we have assumed we are at a sparse solution, we know that at least $M - N$ entries of $\boldsymbol{w}^*$ are equal to zero. Furthermore, let us assume without loss of generality that the first $M - N$ elements of $\boldsymbol{w}^*$ equal zero. This allows us to reexpress (12) as

$$\begin{aligned}\frac{\partial g(\varepsilon)}{\partial \varepsilon} &= \sum_{i=1}^{M-N} \text{sgn}(w_i') p |\varepsilon w_i'|^{p-1} w_i' + O(1) \\ &= p \sum_{i=1}^{M-N} |w_i'|^p \left(\frac{1}{\varepsilon}\right)^{1-p} + O(1). \end{aligned} \tag{13}$$

At this point we observe that any $\boldsymbol{w}' \in \text{Null}(\Phi)$ must have a nonzero element corresponding to a zero element in $\boldsymbol{w}^*$. This is a direct consequence of the URP assumption. Therefore, at least one $w_i', i \in [1, M - N]$ must be nonzero. As such, with ε sufficiently small, we can ignore terms of order $O(1)$ (since $(1/\varepsilon)^{1-p}$ is unbounded for ε sufficiently small and $p < 1$) and we are left in (13) with a summation that must be positive.

Consequently, we see that for all $\varepsilon \in (0, \delta]$, $\partial g(\varepsilon)/\partial \varepsilon > 0$. By the Mean Value Theorem, this requires that $g(\delta) > g(0)$ or more explicitly,

$$d_p(\boldsymbol{w}^* + \delta\boldsymbol{w}') > d_p(\boldsymbol{w}^*). \tag{14}$$

Since w' is an arbitrary feasible vector, this completes the proof.

Finally, in the special case of $p = 0$, it is immediately apparent that all sparse solutions must be local minima, since in this case $d_p(\boldsymbol{w}) = \|\boldsymbol{w}\|_0$ exactly. ■

Proof of Lemma 2: Let $\boldsymbol{w}_0'$ be a vector containing the amplitudes of the nonzero entries in $\boldsymbol{w}_0$ and Φ_1 the associated columns of Φ. Now let us suppose that there does exist a second solution $\boldsymbol{w}$ satisfying the conditions given above, with $\boldsymbol{w}'$ and Φ_2 being analogously defined. This implies that for some $\boldsymbol{w}'$,

$$\boldsymbol{t} = \Phi_1 \boldsymbol{w}_0' = \Phi_2 \boldsymbol{w}', \tag{15}$$

or equivalently, that $\boldsymbol{t}$ lies in both the span of Φ_1 and the span of Φ_2, both of which are full column rank by the URP assumption. Let us define this intersection as

$$\mathcal{A} = \text{span}(\Phi_1) \cap \text{span}(\Phi_2), \tag{16}$$

where we know by construction that

$$\begin{aligned}\dim(\mathcal{A}) &= \dim(\text{Null}([\Phi_1\ \Phi_2])) \\ &= \max(0, D + D_0 - N) \\ &< D_0. \end{aligned} \tag{17}$$

Note that the latter inequality follows since $D < N$ by assumption. At this point there are two possibilities. First, if $D \leq N - D_0$, then $\dim(\mathcal{A}) = 0$ and no solution $\boldsymbol{w}'$ (or $\boldsymbol{w}$ with $d(\boldsymbol{w}) = D$) can exist. Conversely, if $D > N - D_0$, the existence of a solution $\boldsymbol{w}'$ requires that $\Phi_1 \boldsymbol{w}_0'$ resides in a $(D + D_0 - N)$-dimensional subspace of the D_0-dimensional space Range(Φ_1). However, we know that with the entries of $\boldsymbol{w}_0'$ independently drawn from a continuous, bounded density function, $\Phi_1 \boldsymbol{w}_0$ also has a continuous and bounded density in Range(Φ_1) and the set $\{\boldsymbol{w}_0' : \Phi_1 \boldsymbol{w}_0' \in \mathcal{A}\}$ is of probability measure zero (see [7] for a discussion of probability measures). Therefore, we know that

$$\text{P}(\boldsymbol{w} \neq \boldsymbol{w}_0 \text{ exists s.t. } d(\boldsymbol{w}) < N) = \text{P}(\Phi_1 \boldsymbol{w}_0' \in \mathcal{A}) = 0, \tag{18}$$

completing the proof. ■

6. REFERENCES

[1] B. D. Rao, "Signal processing with the sparseness constraint," *Proc. ICASSP*, vol. 3, May 1998.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, 1999.

[3] R. M. Leahy and B. D. Jeffs, "On the design of maximally sparse beamforming arrays," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 8, Aug. 1991.

[4] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. on Signal Processing*, vol. 47, no. 1, Jan. 1999.

[5] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, March 2003.

[6] S. F. Cotter and B. D. Rao, "Sparse channel estimation via Matching Pursuit with application to equalization," *IEEE Trans. on Communications*, vol. 50, no. 3, March 2002.

[7] S. I. Resnick, *A Probability Path*, Birkhauser, Boston, 1999.

[8] D. P. Wipf and B. D. Rao, "Bayesian learning for sparse signal reconstruction," *Proc. ICASSP*, vol. 6, April 2003.