Necessary and Sufficient Conditions for Sparsity Pattern Recovery

Necessary and Sufficient Conditions for Sparsity Pattern Recovery

Alyson K. Fletcher, *Member, IEEE*, Sundeep Rangan, and Vivek K Goyal, *Senior Member, IEEE*

***Abstract*—The paper considers the problem of detecting the sparsity pattern of a k-sparse vector in $\mathbb{R}^n$ from m random noisy measurements. A new necessary condition on the number of measurements for asymptotically reliable detection with maximum-likelihood (ML) estimation and Gaussian measurement matrices is derived. This necessary condition for ML detection is compared against a sufficient condition for simple maximum correlation (MC) or thresholding algorithms. The analysis shows that the gap between thresholding and ML can be described by a simple expression in terms of the total signal-to-noise ratio (SNR), with the gap growing with increasing SNR. Thresholding is also compared against the more sophisticated Lasso and orthogonal matching pursuit (OMP) methods. At high SNRs, it is shown that the gap between Lasso and OMP over thresholding is described by the range of powers of the nonzero component values of the unknown signals. Specifically, the key benefit of Lasso and OMP over thresholding is the ability of Lasso and OMP to detect signals with relatively small components.**

***Index Terms*—Compressed sensing, convex optimization, Lasso, maximum-likelihood (ML) estimation, orthogonal matching pursuit (OMP), random matrices, random projections, sparse approximation, subset selection, thresholding.**

I. Introduction

A common problem in signal processing is to estimate an unknown sparse vector $x \in \mathbb{R}^n$ from linear observations of the form $y = Ax + d$. Here, the measurement matrix $A \in \mathbb{R}^{m \times n}$ is known and $d \in \mathbb{R}^m$ is an additive noise vector with a known distribution. The vector $x \in \mathbb{R}^n$ is said to be *sparse* in that it is known *a priori* to have a relatively small number of nonzero components, but the locations of those components are not known and must be detected as part of the signal estimation.

Sparse signal estimation arises in a number of applications, notably subset selection in linear regression [1]. In this case, determining the locations of the nonzero components of x corresponds to finding a small subset of features which linearly influence the observed data y. In digital communication over channel A with additive noise d, the locations of nonzero components in x could convey information [2], [3].

In this paper, we are concerned with establishing necessary and sufficient conditions for the recovery of the *positions* of the nonzero entries of x, which we call the *sparsity pattern*. The ability to detect the sparsity pattern of a vector from noisy measurements depends on a number of factors, including the signal dimension, level of sparsity, number of measurements, and noise level. The broad goal of this paper is to explain the influence of these factors on the performances of various detection algorithms.

A. Related Work

Sparsity pattern recovery (or more simply, sparsity recovery) has received considerable attention in a variety of guises. Most transparent from our formulation is the connection to sparse approximation. In a typical sparse approximation problem, one is given data $y \in \mathbb{R}^m$, dictionary[1] $A \in \mathbb{R}^{m \times n}$, and tolerance $\epsilon > 0$. The aim is to find $\hat{x}$ with the fewest number of nonzero entries among those satisfying $\|A\hat{x} - y\| \leq \epsilon$. This problem is NP-hard [5] but greedy heuristics (matching pursuit [4] and its variants) and convex relaxations (basis pursuit [6], Lasso [7] and others) can be effective under certain conditions on A and y [8]–[10]. In our formulation, y without additive Gaussian noise would have an exact sparse approximation with k terms.

More recently, the concept of "sensing" sparse or compressible x through multiplication by a suitable random matrix A has been termed compressed sensing [11]–[13]. This has popularized the study of sparse approximation with respect to random dictionaries, which was considered also in [14], [15].

The principal results in compressed sensing bound the ℓ^2 error of a reconstruction computed from Ax relative to the error of an optimal k-term nonlinear approximation of x with respect to a suitable fixed basis. These results show that for m only moderately larger than k, these ℓ^2 errors are similar. For the case where the k-term representation of x is exact, these results thus establish exact recovery of x. For example, if A has independent and identically distributed (i.i.d.) Gaussian entries and x has k nonzero entries, then

$$m \asymp 2k \log(n/m) \tag{1}$$

dictates the minimum scaling at which basis pursuit

$$\arg\min_{\hat{x} \,:\, y = A\hat{x}} \|\hat{x}\|_1$$

succeeds at recovering x exactly from $y = Ax$ with high probability [16].

[1]The term seems to have originated in [4] and may apply to A or the columns of A as a set.

Manuscript received April 11, 2008; revised February 03, 2009. Current version published November 20, 2009. This work was supported in part by a University of California President's Postdoctoral Fellowship, the National Science Foundation (NSF) under CAREER Grant CCF-643836, and the Centre Bernoulli at École Polytechnique Fédérale de Lausanne. The material in this paper was presented in part at the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, December 2008.

A. K. Fletcher is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94709 USA (e-mail: alyson@eecs.berkeley.edu).

S. Rangan is with Qualcomm Technologies, Bedminster, NJ 07302 USA (e-mail: srangan@qualcomm.com, sdrangan@yahoo.com).

V. K. Goyal is with the Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: vgoyal@mit.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2009.2032726

TABLE I
SUMMARY OF RESULTS ON MEASUREMENT SCALING FOR RELIABLE SPARSITY RECOVERY (SEE BODY FOR DEFINITIONS AND TECHNICAL LIMITATIONS)

	finite $\mathsf{SNR} \cdot \mathsf{MAR}$ (or: $x_{\min} = O(1/\sqrt{k})$)	$\mathsf{SNR} \cdot \mathsf{MAR} \to \infty$ (or: $x_{\min} = \omega(1/\sqrt{k})$)
Necessary for ML	$m > \frac{2}{\mathsf{SNR}\cdot\mathsf{MAR}} k \log(n-k) + k - 1$ Theorem 1 or: $m > \frac{2}{x_{\min}^2} \log(n-k) + k - 1$	$m > k$ (elementary)
Necessary and sufficient for lasso	known partially (see footnote 3)	$m > 2k \log(n-k) + k + 1$ Wainwright [17], [18]
Sufficient for orthogonal matching pursuit	unknown (expression above is necessary)	$m > 2k \log(n-k)$ Fletcher and Rangan [29]
Sufficient for thresholding (16)	$m > \frac{8(1+\mathsf{SNR})}{\mathsf{SNR}\cdot\mathsf{MAR}} k \log(n-k)$ Theorem 2 or: $m > \frac{8(1+\|x\|^2)}{x_{\min}^2} \log(n-k)$	$m > \frac{8}{\mathsf{MAR}} k \log(n-k)$ from Theorem 2 or: $m > \frac{8\|x\|^2}{x_{\min}^2} \log(n-k)$

Extension of the exact recovery of the *sparsity pattern* to the noisy case is our object of study. One key previous result in this area, reviewed in more detail in Section II-B, is that the scaling

$$m \asymp 2k \log(n-k) + k + 1 \tag{2}$$

is necessary and sufficient for the Lasso technique to succeed in recovering the sparsity pattern, under certain assumptions on the signal-to-noise ratio (SNR) [17], [18]. While the scalings (1) and (2) are superficially similar, the presence of noise can greatly increase the number of measurements required. For example, if $k = \Theta(n)$ then $m = \Theta(k)$ measurements is sufficient in the noiseless case but $m = \Theta(k \log k)$ measurements are needed in the noisy case.

It should be stressed that this work is concerned only with sparsity pattern detection—not with bounding or estimating the ℓ^2 error when estimating x in noise. Recovering the sparsity pattern certainly results in well-controlled ℓ^2 error, but it is not necessary for low ℓ^2 error. Two important works in this regard are the Dantzig selector of Candès and Tao [19] and the risk minimization estimator of Haupt and Nowak [20]. Mean-squared error and other formulations are beyond the scope of this paper.

B. Preview of Main Results

The condition (2) applies to Lasso, which is a computationally tractable but suboptimal estimation method. A natural question then is: What are the limits on sparsity recovery if computational constraints are removed? We address this in our first main result, Theorem 1 in Section III, which considers maximum-likelihood (ML) estimation. This result is a necessary condition for ML to asymptotically recover the sparsity pattern correctly when A has i.i.d. Gaussian entries. It shows that ML requires a scaling of the number of measurements that differs from the Lasso scaling law (2) by a simple factor that depends only on the SNR and what we call the minimum-to-average ratio (MAR) of the component magnitudes. This expression shows that, at high SNRs, there is a potentially large gap in performance between what is achievable with ML detection and current practical algorithms such as Lasso and orthogonal matching pursuit (OMP). Finding alternative practical algorithms that can close this gap is an open research area.

Previous necessary conditions had been based on information-theoretic capacity arguments in [21], [22] and a use of Fano's inequality in [23]. More recent publications with necessary conditions include [24]–[27]. As described in Section III, our new necessary condition is stronger than the previous results in certain important regimes.

In contrast to removing all computational strictures, it is also interesting to understand what performance can be achieved by algorithms even simpler than Lasso and OMP. To this end, we consider a computationally trivial maximum correlation (MC) algorithm and a closely related thresholding estimator that has been recently studied in [28]. Similar to Lasso and OMP, thresholding may also perform significantly worse than ML at high SNRs. In fact, we provide a precise bound on this performance gap and show that thresholding may require as many as $4(1 + \mathsf{SNR})$ more measurements than ML.

However, at high SNRs, the gap between thresholding and other practical methods such as Lasso and OMP is not as large. In particular, the gap does not grow with SNR. We show that the gap between thresholding on the one hand and Lasso and OMP on the other hand is instead described precisely by the MAR. In particular, the Lasso and OMP algorithms perform significantly better than thresholding when the spread of nonzero component magnitudes is large and the estimator must detect relatively small values. On the other hand, when the spread is bounded, their performances (in terms of number of measurements for success) can be matched within a constant factor by a computationally trivial algorithm.

Table I previews our main results in the context of previous results for Lasso and OMP. The measurement model and parameters MAR and SNR are defined in Section II. Arbitrarily small constants have been omitted to simplify the table entries.

C. Organization of the Paper

The setting is formalized in Section II. In particular, we define our concepts of SNR and MAR; our results clarify the roles of these quantities in the sparsity recovery problem. Necessary conditions for success of any algorithm are considered in Section III. There we present a new necessary condition and compare it to previous results and numerical experiments. Section IV introduces and analyzes a very simple thresholding

algorithm. Conclusions are given in Section V, and proofs appear in the Appendices.

II. Problem Statement

Consider estimating a k-sparse vector $x \in \mathbb{R}^n$ through a vector of observations

$$y = Ax + d \tag{3}$$

where $A \in \mathbb{R}^{m \times n}$ is a random matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries. The vector $d \in \mathbb{R}^m$ represents additive noise and also has i.i.d. $\mathcal{N}(0, 1/m)$ components. Denote the sparsity pattern of x (positions of nonzero entries) by the set I_{true}, which is a k-element subset of the set of indices $\{1, 2, \ldots, n\}$. Estimates of the sparsity pattern will be denoted by $\hat{I}$ with subscripts indicating the type of estimator. We seek conditions under which there exists an estimator such that $\hat{I} = I_{\text{true}}$ with high probability.

The success or failure of any detection algorithm will depend on the unknown deterministic vector x and the realizations of the measurement matrix A and noise d. Of course, the analysis handles A and d probabilistically. In addition, we would like to reduce the dependence on x to one or two scalar parameters so that our results are easy to interpret. A necessary condition for ML can be given using only the magnitude of the smallest nonzero entry of x[2]

$$x_{\min} = \min_{j \in I_{\text{true}}} |x_j|.$$

For other results and to compare algorithms, we need more than only $x_{\min}$. We thus parameterize the dependence on x through two quantities: the SNR, and what we will call the *minimum-to-average ratio* (MAR).

The SNR is defined by

$$\mathsf{SNR} = \frac{\boldsymbol{E}[\|Ax\|^2]}{\boldsymbol{E}[\|d\|^2]}. \tag{4}$$

Since we are considering x as an unknown deterministic vector, and the matrix A and vector d have i.i.d. $\mathcal{N}(0, 1/m)$ components, it is easily verified that

$$\mathsf{SNR} = \|x\|^2. \tag{5}$$

The MAR of x is defined as

$$\mathsf{MAR} = \frac{\min_{j \in I_{\text{true}}} |x_j|^2}{\|x\|^2/k}. \tag{6}$$

Since $\|x\|^2/k$ is the average of $\{|x_j|^2 \mid j \in I_{\text{true}}\}$, $\mathsf{MAR} \in (0, 1]$ with the upper limit occurring when all the nonzero entries of x have the same magnitude. MAR can be interpreted as a reciprocal of dynamic range.

Another quantity of interest is the minimum component SNR, defined as

$$\mathsf{SNR}_{\min} = \frac{\min_{j \in I_{\text{true}}} \boldsymbol{E}\|x_j a_j\|^2}{\boldsymbol{E}\|d\|^2} \tag{7}$$

[2]The magnitude of the smallest nonzero entry of the sparse vector was first highlighted as a key parameter in sparsity pattern recovery in the work of Wainwright [17], [18], [23].

where a_j is the jth column of A. The quantity $\mathsf{SNR}_{\min}$ has a natural interpretation: The numerator $\min \boldsymbol{E}\|x_j a_j\|^2$ is the signal energy due to the smallest nonzero component in x, while the denominator $\boldsymbol{E}\|d\|^2$ is the total noise energy. The ratio $\mathsf{SNR}_{\min}$ thus represents the contribution to the SNR from the smallest nonzero component of the unknown vector x. Our choice of variances for the elements of A and d yields $\mathsf{SNR}_{\min} = x_{\min}^2$. Also, observe that (5) and (6) show

$$\mathsf{SNR}_{\min} = x_{\min}^2 = \frac{1}{k}\mathsf{SNR} \cdot \mathsf{MAR}. \tag{8}$$

A. Normalizations

Other works use a variety of normalizations, e.g., the entries of A have variance $1/n$ in [13], [25]; the entries of A have unit variance and the variance of d is a variable σ^2 in [17], [18], [23], [26], [27]; and our scaling of A and a noise variance of σ^2 are used in [20]. This necessitates great care in comparing results.

We have expressed all our results in terms of SNR, MAR, and $\mathsf{SNR}_{\min}$ as defined above. All of these quantities are *dimensionless*; if either A and d or x and d are scaled together, these ratios will not change. Thus, the results can be applied to *any* scaling of A, d and x, provided that the quantities SNR, MAR, and $\mathsf{SNR}_{\min}$ are computed via their ratio definitions.

To aid some readers in interpreting the results and comparing to other results in the literature, some expressions are given in equivalent forms using $x_{\min}$ and $\|x\|$. It should be noted, however, that any condition on $x_{\min}$ and $\|x\|$ has an implicit dependence on the normalizations of A and d.

B. Review of Lasso and OMP Performance

As discussed above, one common approach for detecting the sparsity pattern of x is the so-called *Lasso* method of [7], also known as basis pursuit denoising [6]. The Lasso method first finds an estimate $\hat{x}$ of x via the optimization

$$\hat{x} = \arg\min_x \left(\|y - Ax\|_2^2 + \mu \|x\|_1 \right) \tag{9}$$

where $\mu > 0$ is an algorithm parameter. The Lasso estimate is essentially a least-square minimization with an additional regularization term $\|x\|_1$, which encourages the solution $\hat{x}$ to be sparse. The sparsity pattern of $\hat{x}$ can then be used as an estimate of the sparsity pattern of x. We will denote this estimate as $\hat{I}_{\text{Lasso}}$

$$\hat{I}_{\text{Lasso}} = \{ j : \hat{x}_j \neq 0 \}.$$

Necessary and sufficient conditions for sparsity recovery via Lasso were determined by Wainwright [17], [18]. He showed that

$$m > 2k \log(n - k) + k + 1 \tag{10}$$

is necessary for Lasso to asymptotically detect the correct sparsity pattern for any x at *any* SNR level. Conversely, if the minimum component SNR scales as

$$\frac{m}{\log(n - k)} \mathsf{SNR}_{\min} \to \infty \tag{11}$$

then the condition (10) is also sufficient; i.e., there exists a sequence of threshold levels $\{\mu_n\}$ that guarantees asymptotically reliable recovery.[3] Using (8), the condition (11) is equivalent to

$$\frac{m}{k\log(n-k)}\mathsf{SNR}\cdot\mathsf{MAR}\to\infty.$$

Therefore, for the measurement scaling given precisely by (10), we need that $\mathsf{SNR}\cdot\mathsf{MAR}\to\infty$. The condition (10) is included in Table I.

Another common approach to sparsity pattern detection is the greedy OMP algorithm [30]–[32]. This was analyzed by Tropp and Gilbert [33] in a setting with no noise. More recently, Fletcher and Rangan [29] improved this analysis, lowering the number of measurements sufficient for recovery while also allowing noise satisfying (11) and bounded uncertainty in *a priori* knowledge of k. They show that, when A has Gaussian entries, a *sufficient* condition for asymptotic reliable recovery is

$$m>2k\log(n-k), \tag{12}$$

similar to the condition for Lasso. This condition is also supported by numerical experiments reported in [33]. The sufficient condition (12) appears in Table I.

III. Necessary Condition for Sparsity Recovery

We first consider sparsity recovery without being concerned with computational complexity of the estimation algorithm. Since our formulation is non-Bayesian, we consider the ML estimate, which is optimal when there is no prior on x other than it being k-sparse.

The vector $x\in\mathbb{R}^n$ is k-sparse, so the vector Ax belongs to one of $\binom{n}{k}$ subspaces spanned by k of the n columns of A. Estimation of the sparsity pattern is the selection of one of these subspaces, and since the noise d is Gaussian, the ML estimate minimizes the Euclidean norm of the residual between y and its projection to the subspace. More simply, the ML estimate chooses the subspace closest to y.

Mathematically, the ML estimator can be described as follows. Given a subset $J\subseteq\{1,2,\ldots,n\}$, let P_Jy denote the orthogonal projection of the vector y onto the subspace spanned by the vectors $\{a_j\mid j\in J\}$. The ML estimate of the sparsity pattern is

$$\hat{I}_{\mathrm{ML}}=\arg\max_{J:|J|=k}\|P_Jy\|^2$$

where $|J|$ denotes the cardinality of J. That is, the ML estimate is the set of k indices such that the subspace spanned by the corresponding columns of A contain the maximum signal energy of y.

ML estimation for sparsity recovery was first examined by Wainwright [23]. He showed in [23, Theorem 1] that there exists a constant $C>0$ such that the condition

$$m>C\max\left\{\frac{1}{\mathsf{SNR}_{\min}}\log(n-k),k\log(n/k)\right\} \tag{13a}$$

$$=C\max\left\{\frac{1}{x_{\min}^2}\log(n-k),k\log(n/k)\right\} \tag{13b}$$

$$=C\max\left\{\frac{k}{\mathsf{SNR}\cdot\mathsf{MAR}}\log(n-k),k\log(n/k)\right\} \tag{13c}$$

is *sufficient* for ML to asymptotically reliably recover the sparsity pattern. Note that the equality in (13c) is a consequence of (8). Our first theorem provides a corresponding necessary condition.

Theorem 1: Let $k=k(n)$, $m=m(n)$, $\mathsf{SNR}=\mathsf{SNR}(n)$, and $\mathsf{MAR}=\mathsf{MAR}(n)$ be deterministic sequences in n such that $\lim_{n\to\infty}k(n)=\infty$ and

$$m(n)<\frac{2(1-\delta)}{\mathsf{SNR}_{\min}}\log(n-k)+k-1 \tag{14a}$$

$$=\frac{2(1-\delta)}{x_{\min}^2}\log(n-k)+k-1 \tag{14b}$$

$$=\frac{2(1-\delta)}{\mathsf{SNR}\cdot\mathsf{MAR}}k\log(n-k)+k-1 \tag{14c}$$

for some $\delta>0$. Then even the ML estimator asymptotically cannot detect the sparsity pattern, i.e.,

$$\lim_{n\to\infty}\Pr\left(\hat{I}_{\mathrm{ML}}=I_{\mathrm{true}}\right)=0.$$

Proof: See Appendix B. □

The theorem provides a simple lower bound on the minimum number of measurements required to recover the sparsity pattern in terms of k, n, and the minimum component SNR, $\mathsf{SNR}_{\min}$. Note again that the equality in (14c) is due to (8).

Remarks:

1) The theorem strengthens an earlier necessary condition in [24] which showed that there exists a $C>0$ such that
$$m=\frac{C}{\mathsf{SNR}}k\log(n-k)$$
is necessary for asymptotic reliable recovery. Theorem 1 strengthens the result to reflect the dependence on MAR and make the constant explicit.
2) When the first of the two terms in the maximum in the sufficient condition (13) dominates, the sufficient condition matches the necessary condition (14) within a constant factor. The fact that the two conditions match is not surprising since the proofs of the two use similar methods: The necessary condition is proven by considering the tail probability of all error events with a single incorrect vector. The sufficient condition is proven by bounding the sum probability of all error events.
However, the first term in (13c), $C\log(n-k)/\mathsf{SNR}_{\min}$, is not always dominant. For example, if $\mathsf{SNR}\cdot\mathsf{MAR}\to\infty$ or $k=o(n)$, then the second term $Ck\log(n/k)$ may be larger. In this case, there is a gap between the necessary and sufficient conditions. The exact scalings for reliable sparsity recovery with ML detection in these regimes are not known.
3) The bound (14) strengthens earlier results in the regime where $\mathsf{SNR}\cdot\mathsf{MAR}$ is bounded and the sparsity ratio k/n

[3]Sufficient conditions under weaker conditions on $\mathsf{SNR}_{\min}$ are given in [18]. Interpreting these is more subtle: the scaling of $\mathsf{SNR}_{\min}$ with n determines the sequences of regularization parameters $\{\mu_n\}$ for which asymptotic almost sure success is achieved, and $\{\mu_n\}$ affects the sufficient number of measurements.

is fixed.[4] To see this point, we can compare (14) against previous information-theoretic bounds in [21]–[23], [26], [27]. As one example, consider the bound in [21], which uses a simple capacity argument to show that

$$m \geq \frac{2}{\log_2(1+\mathsf{SNR})} \log_2 \binom{n}{k} \tag{15}$$

is necessary for sparsity pattern detection. When k/n and the SNR are both fixed, m can satisfy (15) while growing only linearly with k. The other capacity-based bounds have the same property.

In contrast, Theorem 1 shows that with k/n fixed and $\mathsf{SNR}\cdot\mathsf{MAR}$ bounded, $m = \Omega(k\log(n-k))$ is necessary for reliable sparsity recovery. That is, the number of measurements must grow *superlinearly* in k in the linear sparsity regime with bounded SNR.

4) In the regime of sublinear sparsity (where $k = o(n)$) or in the regime where $\mathsf{SNR}\cdot\mathsf{MAR} \to \infty$, information-theoretic bounds such as (15) may be stronger than (14) depending on the precise scaling of SNR, MAR, and other terms.
5) For a set of sparse signals specified by a given $x_{\min}$, sparsity pattern recovery is most difficult when the magnitudes of all nonzero entries is $x_{\min}$. (This makes $\mathsf{MAR} = 1$.) Thus, minimax rate conditions for this set are determined by analysis of the case with all nonzero entries of x equal to $x_{\min}$. In [23, Theorem 2] Wainwright gives a necessary condition of

$$m > \frac{k\log(n/k)}{4\,\mathsf{SNR}}$$

for asymptotic reliable sparsity pattern recovery of such signals.[5] When $\log(n/k)$ and $\log(n-k)$ are of the same order, Wainwright's earlier necessary condition is within a constant factor of that of Theorem 1. While Wainwright correctly makes the point that $x_{\min}$ is the most important quantity for sparsity pattern recovery, one should be careful to understand that all the nonzero entries of x affect detection performance; the remaining nonzero entries disappear from Wainwright's analysis because he gives a minimax result.
6) Results more similar to Theorem 1—based on direct analyses of error events rather than information-theoretic arguments—appeared in [24], [25]. The previous results showed that with fixed SNR as defined here, sparsity recovery with $m = \Theta(k)$ must fail. The more refined analysis in this paper gives the additional $\log(n-k)$ factor and the precise dependence on $\mathsf{SNR}\cdot\mathsf{MAR}$.
7) Theorem 1 is not contradicted by the relevant sufficient condition of [26], [27]. That sufficient condition holds for scaling that gives linear sparsity and $\mathsf{SNR}\cdot\mathsf{MAR} = \Omega(\sqrt{n\log n})$. For $\mathsf{SNR}\cdot\mathsf{MAR} = \sqrt{n\log n}$, Theorem 1 shows that fewer than $m \asymp 2\sqrt{k\log k}$ measurements will cause ML decoding to fail, while [27, Theorem 3.1] shows that a typicality based decoder will succeed with $m = \Theta(k)$ measurements.
8) Our definition of the ML estimator requires that the number of nonzero components k must be known *a priori*. Of course, in many practical applications, k may not be known. If k must be estimated, we would expect the number of measurements to be even higher, so the lower bound (14) should apply in this case as well.
9) The condition (14) can be rearranged to be a necessary condition on a parameter other than m. For example, rearranging to obtain a necessary condition on SNR gives an improvement over [34, Corollary 4.1] by about a factor of 4 in the restricted scenario of $m = \Theta(n)$ considered therein.
10) After the posting of [35], Theorem 1 was generalized to i.i.d. non-Gaussian distributions for the entries of A in [36]. An additional contribution of [36] is the study of a specific distribution for A obtained by multiplying i.i.d. Gaussian entries by i.i.d. Bernoulli variables.

Comparison to Lasso and OMP: Comparing the necessary condition for ML in (14c) against the sufficient conditions (10) and (12) for Lasso and OMP, we see that ML may require dramatically fewer measurements than Lasso or OMP. Specifically, ML may require a factor of $O(1/\mathsf{SNR}\cdot\mathsf{MAR})$ fewer measurements. This gap grows as $\mathsf{SNR}\cdot\mathsf{MAR} \to \infty$.

Of course, since the ML scaling (14) is only necessary, the actual performance gap between ML and Lasso may be smaller. However, we do know that in the noiseless case, for any fixed k-sparse x, it is sufficient to have $m = k+1$ for ML to succeed with probability 1 over the choice of A. This is a scaling at which Lasso fails, even in the noiseless case [17], [18]. It is an open question to characterize the precise gap and to determine if there are practical algorithms that close it.

Numerical Validation: The theorem is asymptotic, so it cannot be confirmed computationally. Even qualitative support is hard to obtain because of the high complexity of ML detection. Nevertheless, we present some evidence obtained through Monte Carlo simulation.

Fig. 1 shows the probability of success of ML detection for $n = 20$ as k, m, SNR, and MAR are varied, with each point representing at least 500 independent trials. Each subpanel gives simulation results for $k \in \{1,2,\ldots,5\}$ and $m \in \{1,2,\ldots,40\}$ for one $(\mathsf{SNR},\mathsf{MAR})$ pair. Signals with $\mathsf{MAR} < 1$ are created by having one small nonzero component and $k-1$ equal, larger nonzero components. Overlaid on the intensity plots is a white curve representing (14).[6]

Taking any one column of one subpanel from bottom to top shows that as m is increased, there is a transition from ML failing to ML succeeding. One can see that (14) follows the failure–success transition qualitatively. In particular, the empirical dependence on SNR and MAR approximately follows (14c). Empirically, for the (small) value of $n = 20$, it seems that with $\mathsf{SNR}\cdot\mathsf{MAR}$ held fixed, sparsity recovery becomes easier as SNR increases (and MAR decreases).

Less extensive Monte Carlo simulations for $n = 40$ are reported in Fig. 2. The results are qualitatively similar. As might be expected, the transition from low to high probability of successful recovery as a function of m appears more sharp at $n = 40$ than at $n = 20$.

[4]We will sometimes use *linear sparsity* to mean that k/n is fixed.

[5]While [23, Theorem 2] is stated without specifying a leading constant, the constant 4 can be extracted from its proof.

[6]Color versions of Figs. 1, 2 and 4 are available online in [35].

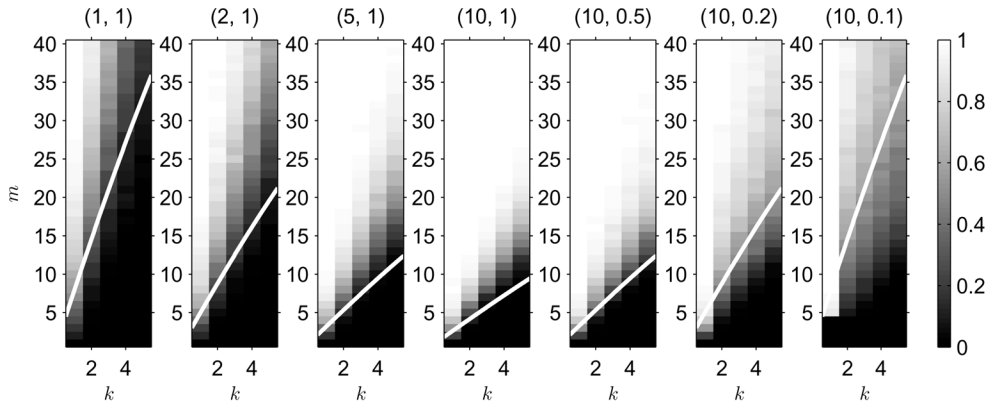

Fig. 1. Simulated success probability of ML detection for $n = 20$ and many values of k, m, SNR, and MAR. Each subgraph gives simulation results for $k \in \{1, 2, \ldots, 5\}$ and $m \in \{1, 2, \ldots, 40\}$ for one (SNR, MAR) pair. Each subgraph heading gives (SNR, MAR). Each point represents at least 500 independent trials. Overlaid on the intensity plots is a white curve representing (14).

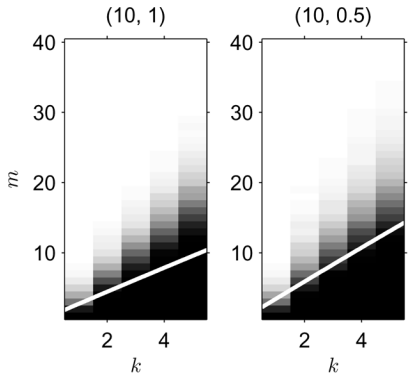

Fig. 2. Simulated success probability of ML detection for $n = 40$; $\mathsf{SNR} = 10$; $\mathsf{MAR} = 1$ (left) or $\mathsf{MAR} = 0.5$ (right); and many values of k and m. Each subgraph gives simulation results for $k \in \{1, 2, \ldots, 5\}$ and $m \in \{1, 2, \ldots, 40\}$, with each point representing at least 1000 independent trials. Overlaid on the intensity plots (with scale as in Fig. 1) is a white curve representing (14).

IV. Sufficient Condition With Thresholding

The ML estimator analyzed in the previous section becomes computationally infeasible quickly as problem size increases. We now consider a sparsity recovery algorithm even simpler than Lasso and OMP. It is not meant as a competitive alternative. Rather, it serves to illustrate the precise benefits of Lasso and OMP.

As before, let a_j be the jth column of the random matrix A. Define the *thresholding estimate* as

$$\hat{I}_{\text{thresh}} = \{j : \rho(j) > \mu\} \tag{16}$$

where $\rho(j)$ is the correlation

$$\rho(j) = |a_j' y|^2 / \|a_j\|^2 \tag{17}$$

and $\mu > 0$ is a threshold level. This algorithm simply correlates the observed signal y with all the frame vectors a_j and selects the indices j where the correlation energy exceeds a certain level μ.

A closely related algorithm is to compute the *maximum correlation* (MC) estimate

$$\hat{I}_{\text{MC}} = \{j : \rho(j) \text{ is one of the } k \text{ largest values}\}. \tag{18}$$

This has slightly higher complexity because it requires the sorting of $\{\rho(j)\}_{j=1}^n$, but it also has better performance in principle. It is straightforward to show that $\hat{I}_{\text{MC}} = I_{\text{true}}$ if and only if there exists a threshold μ such that $\hat{I}_{\text{thresh}} = I_{\text{true}}$. Using MC instead of thresholding requires *a priori* knowledge of k but eliminates the need to pick a "correct" threshold value. Our sufficient condition is for the weaker thresholding estimate and thus applies also to the MC estimate.

A variant of these algorithms, called the "one-step greedy algorithm (OSGA)," was proposed by Duarte *et al.* [37] for detecting jointly sparse signals. For our application, the most closely related previous study was by Rauhut *et al.* [28]. They proved a sufficient condition for asymptotic reliability of the thresholding estimate when there is no noise. The following theorem tightens the previous result in the noiseless case and generalizes to the noisy case.

Theorem 2: Let $k = k(n)$, $m = m(n)$, $\mathsf{SNR} = \mathsf{SNR}(n)$, and $\mathsf{MAR} = \mathsf{MAR}(n)$ be deterministic sequences in n such that $\lim_{n\to\infty} k = \infty$ and

$$m > \frac{2(1+\delta)(1+\mathsf{SNR})}{\mathsf{SNR} \cdot \mathsf{MAR}} k L(k,n) \tag{19}$$

for some $\delta > 0$, where

$$L(n,k) = \left(\sqrt{\log(n-k)} + \sqrt{\log(k)}\right)^2. \tag{20}$$

Then there exists a sequence of threshold levels $\mu = \mu(n)$ such that thresholding asymptotically detects the sparsity pattern, i.e.,

$$\lim_{n\to\infty} \Pr\left(\hat{I}_{\text{thresh}} = I_{\text{true}}\right) = 1.$$

Proof: See Appendix C. □

Remarks:

1) The factor $L(n,k)$ in (20) can be bounded to express (19) in a form more easily comparable to (14c). If $k/n \le 1/2$, then $\log(k) \le \log(n-k)$, so $L(n,k) \le 4\log(n-k)$. Using this bound with (19) shows that the more restrictive condition

$$m > \frac{8(1+\delta)(1+\mathsf{SNR})}{\mathsf{SNR} \cdot \mathsf{MAR}} k \log(n-k) \tag{21}$$

is sufficient for asymptotic reliability of thresholding when $k/n < 1/2$. This expression is shown in Table I, where the infinitesimal δ quantity has been omitted for simplicity. From the expression

$$\frac{L(n,k)}{\log(n-k)} = \left(1+\sqrt{\frac{\log(k)}{\log(n-k)}}\right)^2$$

we can draw further conclusions: (a) $L(n,k)/\log(n-k) \geq 1$; (b) when $k = \Theta(n)$, $\lim_{n\to\infty} L(n,k)/\log(n-k) = 4$, so (21) is asymptotically equivalent to (19); and (c) when $k = o(n)$, the simpler form (21) is pessimistic. Fig. 3 plots the ratio $L(n,k)/\log(n-k)$ as a function of n for a few sparsity scalings $k(n)$.

2) Comparing (14c) and (19), we see that thresholding requires a factor of at most $L(n,k)(1+\mathsf{SNR})/\log(n-k)$ more measurements than ML estimation. This factor is upper-bounded by $4(1+\mathsf{SNR})$ and can be as small as $(1+\mathsf{SNR})$.
The factor $1+\mathsf{SNR}$ has a natural interpretation: The lower bound for ML estimation in (14c) is proven by essentially assuming that, when detecting each component of the unknown vector x, the $n-1$ other components are known. Thus, the detection only sees interference from the additive noise d. In contrast, thresholding treats the other $n-1$ vectors as noise, resulting in a total increase in effective noise by a factor of $1+\mathsf{SNR}$. To compensate for this increase in effective noise, the number of measurements m must be scaled proportionally.
We can think of this additional noise as *self-noise*, by which we mean the interference caused from different components of the signal x interfering with one another in the observed signal y through the measurement matrix A. This self-noise is distinct from the additive noise d.
3) The gap between thresholding and ML can be large at high SNRs. As one extreme, consider the case where $\mathsf{SNR} \to \infty$. For ML estimation, the lower bound on the number of measurements required by ML decreases to $k-1$ as $\mathsf{SNR} \to \infty$.[7] In contrast, with thresholding, increasing the SNR has diminishing returns: as $\mathsf{SNR} \to \infty$, the bound on the number of measurements in (19) approaches

$$m > \frac{2}{\mathsf{MAR}} k L(n,k) \approx \frac{8}{\mathsf{MAR}} k \log(n-k) \qquad (22)$$

where the approximation holds for linear sparsity. Thus, even with $\mathsf{SNR} \to \infty$, the minimum number of measurements is not improved from the scaling $m = O(k\log(n-k))$.
By the discussion in Remark 2, we can think of this problem as a self-noise limit: As the additive noise is reduced, thresholding becomes limited by signal-dependent noise. This self-noise limit is also exhibited by more sophisticated methods such as Lasso. For example, as discussed earlier, the analysis of [17] shows that when $\mathsf{SNR}\cdot\mathsf{MAR} \to \infty$, Lasso requires

$$m > 2k\log(n-k) + k + 1 \qquad (23)$$

[7]Of course, at least $k+1$ measurements are necessary.

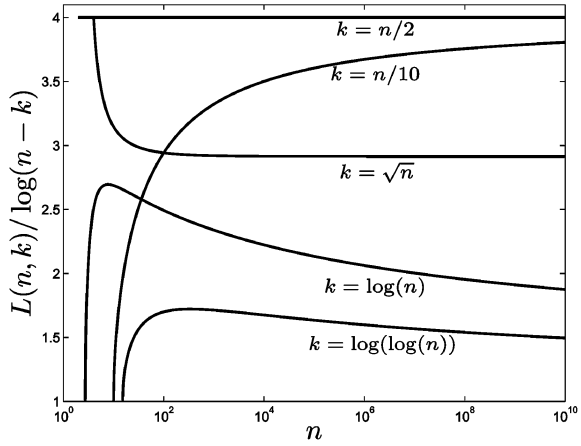

Fig. 3. Plots of $L(n,k)/\log(n-k)$ as a function of k for a few sparsity regimes $k(n)$. When k/n is constant, the ratio approaches 4. When $k = n^\alpha$ for $\alpha \in (0,1)$, the ratio approaches $(1+\sqrt{\alpha})^2$. When k is asymptotically smaller than any power of α, the ratio approaches 1.

for reliable recovery. Therefore, like thresholding, Lasso does not achieve a scaling better than $m = O(k\log(n-k))$, even at infinite SNR.
4) There is an important advantage of Lasso over thresholding. Comparing (22) and (23), we see that, at high SNR, thresholding requires a factor of up to $4/\mathsf{MAR}$ more measurements than Lasso. This factor is largest when MAR is small, which occurs when there are relatively small nonzero components. This gap reveals the key benefit of Lasso over thresholding: its ability to detect small coefficients, even when they are much below the average energy.
At high SNRs, the gap of $4/\mathsf{MAR}$ can be arbitrarily large. As an extreme example, suppose the unknown vector x has $k-1$ components with $|x_i| = 1$ and one component with $|x_i|^2 = k^{-\alpha}$ for some $\alpha < 1$. Then, $\mathsf{SNR} \approx k$ and $\mathsf{MAR} \approx k^{-\alpha}$, where the approximation is valid for large k. Now if m satisfies (10), then it can be verified that (11) is also satisfied. Therefore, the scaling (10) will be sufficient for Lasso. In comparison, thresholding could need as much as $4/\mathsf{MAR} \approx 4k^\alpha$ more measurements, which grows to infinity with k. So, at high SNRs, Lasso can significantly outperform thresholding because we require the estimator to recover all the nonzero components, even the very small ones.
5) The high SNR limit (22) matches the sufficient condition in [28] for the noise-free case, except that the constant in (22) is tighter.
6) We have emphasized dimensionless quantities so that the normalizations of A and d are immaterial. An equivalent to (19) that depends on the normalizations defined herein is

$$m > \frac{2(1+\delta)(1+\|x\|^2)}{x_{\min}^2} L(n,k).$$

Threshold Selection: Typically, the threshold level μ in (16) is set by trading off the false alarm and missed detection probabilities. Optimal selection depends on a variety of factors including the noise, component magnitudes, and the statistics on

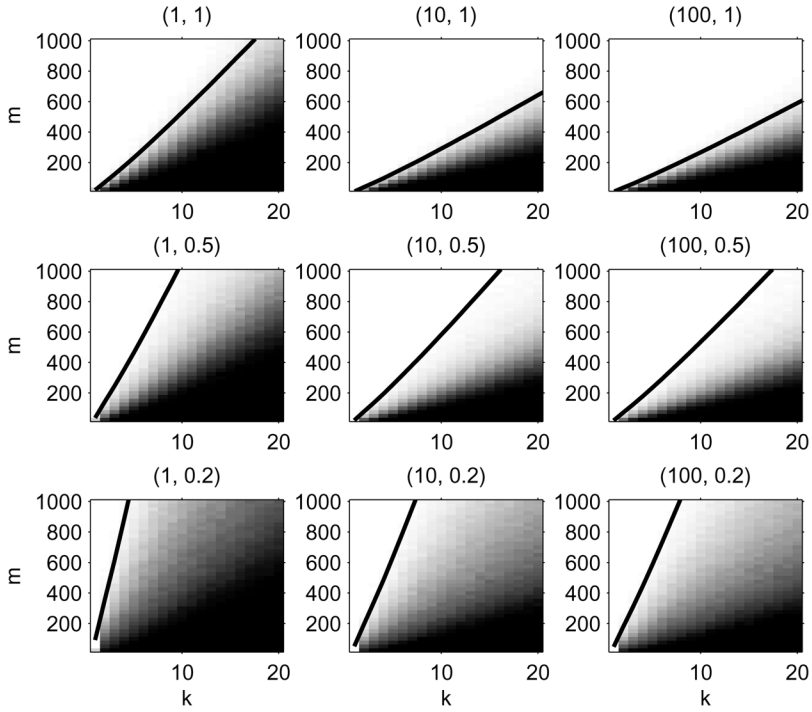

Fig. 4. Simulated success probability of thresholding detection for $n = 100$ and many values of k, m, SNR, and MAR. Each subgraph gives simulation results for $k \in \{1, 2, \ldots, 20\}$ and $m \in \{25, 50, \ldots, 1000\}$ for one (SNR, MAR) pair. Each subgraph heading gives (SNR, MAR), so $\mathsf{SNR} = 1, 10, 100$ for the three columns and $\mathsf{MAR} = 1, 0.5, 0.2$ for the three rows. Each point represents 1000 independent trials. Overlaid on the intensity plots (with scale as in Fig. 1) is a black curve representing (19).

the number of nonzero components. The proof of Theorem 2 (see Appendix C) sets the threshold level to

$$\mu = F(n,k) = 2(1+\epsilon)(1+\mathsf{SNR})\frac{\log(n-k)}{m}$$

where ϵ depends on δ. Thus, the threshold selection explicitly requires knowledge of k.

If k is not known, but has a known statistical distribution, one can use the threshold, $\mu = F(n, \overline{k})$, where $\overline{k}$ is the expected value of k. A straightforward modification to the proof of Theorem 2 shows that if k has a distribution such that

$$\lim_{n\to\infty} \frac{F(n,k)}{F(n,\overline{k})} = 1$$

almost surely, then the threshold $\mu = F(n, \overline{k})$ will work as well.

Of course, if k is actually known *a priori*, one can do slightly better than thresholding with the MC algorithm, obviating the need for threshold selection.

Numerical Validation: Thresholding is extremely simple and can thus be simulated easily for large problem sizes. Fig. 4 reports the results of a large number Monte Carlo simulations of the thresholding method with $n = 100$. The sufficient condition predicted by (19) matches well to the parameter combinations where the probability of success drops below about 0.995. To avoid the issue of threshold selection, we have used the maximum correlation estimator (18) instead of (16).

V. Conclusion

We have considered the problem of detecting the sparsity pattern of a sparse vector from noisy random linear measurements. Necessary and sufficient scaling laws for the number of measurements to recover the sparsity pattern for different detection algorithms were derived. The analysis reveals the effect of two key factors: the total SNR, as well as the MAR, which is a measure of the spread of component magnitudes. The product of these factors is k times the SNR contribution from the smallest nonzero component; this product often appears.

Our main conclusions are as follows.

- *Necessary and sufficient scaling laws*. As a necessary condition for sparsity pattern detection, we have proven a lower bound on the minimum number of measurements for ML estimation to work. We also derived a sufficient condition for a trivially simple thresholding estimator.
  With fixed SNR and MAR, both the necessary and sufficient scaling laws have the form $m = O(k\log(n-k))$. However, the effect of the SNR and MAR can be dramatic and is what primarily differentiates the performance between different algorithms.
- *Self-noise limits at high SNR*. Thresholding may require as many as $4(1+\mathsf{SNR})$ times more measurements than ML estimation, which is significant at high SNRs. The factor $1+\mathsf{SNR}$ has an interpretation as a self-noise effect.
  As a result there is a self-noise limit: As $\mathsf{SNR} \to \infty$, thresholding achieves a fundamentally worse scaling than ML. Specifically, ML may in principle be able to detect the sparsity pattern with $m = O(k)$ measurements. In contrast, due to the self-noise effect, thresholding requires at least $m = \Omega(k\log(n-k))$. Unfortunately, the more sophisticated Lasso and OMP methods also require $m = \Omega(k\log(n-k))$ scaling.

- *Lasso, OMP and dependence on MAR*. Lasso and OMP, however, have an important advantage over thresholding at high SNRs, which is their ability to deal with a large range in component magnitudes. Specifically, thresholding may require up to $4/\mathsf{MAR}$ times more measurements than Lasso. Thus, when there are nonzero components that are much below the average component energy, thresholding will perform significantly worse. However, when the components of the unknown vector have equal magnitudes ($\mathsf{MAR} = 1$), Lasso and thresholding have asymptotic scaling within a constant factor.

While the results provide insight into the limits of detection, there are clearly many open issues. Most importantly, the well-known "practical" algorithms—Lasso, OMP, and thresholding—all appear to have a scaling of at least $m = \Omega(k \log(n-k))$ measurements as $\mathsf{SNR} \to \infty$. In contrast, ML may be able to achieve a scaling of $m = O(k)$ with sufficient SNR. An open question is whether there is any practical algorithm that can achieve a similar scaling at high SNR.

A second open issue is to determine conditions for partial sparsity recovery. The above results define conditions for recovering all the positions in the sparsity pattern. However, in many practical applications, obtaining some large fraction of these positions would be adequate. Neither the limits of partial sparsity recovery nor the performance of practical algorithms are completely understood, though some results have been reported in [25]–[27], [38].

Appendix A
Deterministic Necessary Condition

The proof of Theorem 1 is based on the following deterministic necessary condition for sparsity recovery. Recall the notation that for any $J \subseteq \{1,2,\ldots,n\}$, P_J denotes the orthogonal projection onto the span of the vectors $\{a_j\}_{j\in J}$. Additionally, let $P_J^{\perp} = I - P_J$ denote the orthogonal projection onto the orthogonal complement of $\mathrm{span}(\{a_j\}_{j\in J})$.

Lemma 1: A necessary condition for ML detection to succeed (i.e., $\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}$) is

$$\text{for all } i \in I_{\mathrm{true}} \text{ and } j \notin I_{\mathrm{true}}, \quad \frac{\left|a_i' P_K^{\perp} y\right|^2}{a_i' P_K^{\perp} a_i} \geq \frac{\left|a_j' P_K^{\perp} y\right|^2}{a_j' P_K^{\perp} a_j} \tag{24}$$

where $K = I_{\mathrm{true}} \setminus \{i\}$.

Proof: Note that $y = P_K y + P_K^{\perp} y$ is an orthogonal decomposition of y into the portions inside and outside the subspace $S = \mathrm{span}(\{a_j\}_{j\in K})$. An approximation of y in subspace S leaves residual $P_K^{\perp} y$. Intuitively, the condition (24) requires that the residual be at least as highly correlated with the remaining "correct" vector a_i as it is with any of the "incorrect" vectors $\{a_j\}_{j\notin I_{\mathrm{true}}}$.

Fix any $i \in I_{\mathrm{true}}$ and $j \notin I_{\mathrm{true}}$, and let

$$J = K \cup \{j\} = (I_{\mathrm{true}} \setminus \{i\}) \cup \{j\}.$$

That is, J is equal to the true sparsity pattern I_{true}, except that a single "correct" index i has been replaced by an "incorrect" index j. If the ML estimator is to select $\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}$ then the energy of the noisy vector y must be larger on the true subspace I_{true} than the incorrect subspace J. Therefore

$$\|P_{I_{\mathrm{true}}} y\|^2 \geq \|P_J y\|^2. \tag{25}$$

Now, a simple application of the matrix inversion lemma shows that since $I_{\mathrm{true}} = K \cup \{i\}$

$$\|P_{I_{\mathrm{true}}} y\|^2 = \|P_K y\|^2 + \frac{\left|a_i' P_K^{\perp} y\right|^2}{a_i' P_K^{\perp} a_i}. \tag{26a}$$

Also, since $J = K \cup \{j\}$

$$\|P_J y\|^2 = \|P_K y\|^2 + \frac{\left|a_j' P_K^{\perp} y\right|^2}{a_j' P_K^{\perp} a_j}. \tag{26b}$$

Substituting (26a)–(26b) into (25) and canceling $\|P_K y\|^2$ shows (24). □

Appendix B
Proof of Theorem 1

To simplify notation, assume without loss of generality that $I_{\mathrm{true}} = \{1,2,\ldots,k\}$. Also, assume that the minimization in (8) occurs at $j = 1$ with

$$|x_1|^2 = \frac{1}{k}\,\mathsf{SNR} \cdot \mathsf{MAR}. \tag{27}$$

Finally, since adding measurements (i.e., increasing m) can only improve the chances that ML detection will work, we can assume that in addition to satisfying (14c), the numbers of measurements satisfy the lower bound

$$m > \epsilon k \log(n-k) + k - 1 \tag{28}$$

for some $\epsilon > 0$. This assumption implies that

$$\lim_{n\to\infty} \frac{\log(n-k)}{m-k+1} = \lim_{n\to\infty} \frac{1}{\epsilon k} = 0. \tag{29}$$

Here and in the remainder of the proof the limits are as m, n, and $k \to \infty$ subject to (14c) and (28). With these requirements on m, we need to show $\lim_{n\to\infty} \Pr(\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}) = 0$.

From Lemma 1 in Appendix A, $\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}$ implies (24). Thus

$$\begin{aligned}
&\Pr\left(\hat{I}_{\mathrm{ML}} = I_{\mathrm{true}}\right) \\
&\quad \leq \Pr\left(\frac{\left|a_i' P_K^{\perp} y\right|^2}{a_i' P_K^{\perp} a_i} \geq \frac{\left|a_j' P_K^{\perp} y\right|^2}{a_j' P_K^{\perp} a_j} \;\forall\, i \in I_{\mathrm{true}}, j \notin I_{\mathrm{true}}\right) \\
&\quad \leq \Pr\left(\frac{\left|a_1' P_K^{\perp} y\right|^2}{a_1' P_K^{\perp} a_1} \geq \frac{\left|a_j' P_K^{\perp} y\right|^2}{a_j' P_K^{\perp} a_j} \;\forall\, j \notin I_{\mathrm{true}}\right) \\
&\quad = \Pr(\Delta^- \geq \Delta^+)
\end{aligned} \tag{30}$$

where

$$\Delta^- = \frac{m}{\log(n-k)} \frac{\left|a_1' P_K^\perp y\right|^2}{a_1' P_K^\perp a_1}$$

$$\Delta^+ = \frac{m}{\log(n-k)} \max_{j \in \{k+1,\ldots,n\}} \frac{\left|a_j' P_K^\perp y\right|^2}{a_j' P_K^\perp a_j}$$

and $K = I_{\text{true}} \setminus \{1\} = \{2,\ldots,k\}$. The $-$ and $+$ superscripts are used to reflect that Δ^- is the energy lost from removing "correct" index 1, and Δ^+ is the energy added from adding the worst "incorrect" index. We will show that

$$\limsup_{n\to\infty} \Delta^- \leq 2(1-\delta) < 2 \leq \liminf_{n\to\infty} \Delta^+ \tag{31}$$

where both limits are in probability. This will show that

$$\lim_{n\to\infty} \Pr\left(\Delta^- \geq \Delta^+\right) = 0$$

which in turn, by (30) will imply that

$$\lim_{n\to\infty} \Pr\left(\hat{I}_{\text{ML}} = I_{\text{true}}\right) = 0$$

and thus prove the theorem. We therefore need to show the two limits in (31). We will consider the two limits separately.

A. Limit of Δ^+

Let V_K be the $k-1$-dimensional space spanned by the vectors $\{a_j\}_{j\in K}$. For each $j \notin I_{\text{true}}$, let u_j be the unit vector

$$u_j = P_K^\perp a_j / \|P_K^\perp a_j\|.$$

Since a_j has i.i.d. Gaussian components, it is spherically symmetric. Also, if $j \notin K$, a_j is independent of the subspace V_K. Hence, in this case, u_j will be a unit vector uniformly distributed on the unit sphere in $V_K^\perp$. Since $V_K^\perp$ is an $m-k+1$-dimensional subspace, it follows from Lemma 5 (see Appendix D) that if we define

$$z_j = \left|u_j' P_K^\perp y\right|^2 / \left\|P_K^\perp y\right\|^2$$

then z_j follows a $\text{Beta}(1, m-k)$ distribution. See Appendix D for a review of the chi-squared and beta distributions and some simple results on these variables that will be used in the proofs below.

By the definition of u_j

$$\frac{\left|a_j' P_K^\perp y\right|^2}{a_j' P_K^\perp a_j} = \left|u_j' P_K^\perp y\right|^2 = z_j \left\|P_K^\perp y\right\|^2$$

and therefore

$$\Delta^+ = \frac{1}{\log(n-k)} \left\|P_K^\perp y\right\|^2 \max_{j\in\{k+1,\ldots,n\}} z_j. \tag{32}$$

Now the vectors a_j are independent of one another, and for $j \notin I_{\text{true}}$, each a_j is independent of $P_K^\perp y$. Therefore, the variables z_j will be i.i.d. Hence, using Lemma 6 (see Appendix D) and (29)

$$\lim_{n\to\infty} \frac{m-k+1}{\log(n-k)} \max_{j=k+1,\ldots,n} z_j = 2 \tag{33}$$

in probability. Also

$$\begin{aligned}\liminf_{n\to\infty} \frac{m}{m-k+1} \left\|P_K^\perp y\right\|^2 &\overset{(a)}{\geq} \lim_{n\to\infty} \frac{m}{m-k+1} \left\|P_{I_{\text{true}}}^\perp y\right\|^2 \\ &\overset{(b)}{=} \lim_{n\to\infty} \frac{m}{m-k+1} \left\|P_{I_{\text{true}}}^\perp d\right\|^2 \\ &\overset{(c)}{=} \lim_{n\to\infty} \frac{m-k}{m-k+1} = 1 \end{aligned} \tag{34}$$

where (a) follows from the fact that $K \subset I_{\text{true}}$ and hence $\|P_K^\perp y\| \geq \|P_{I_{\text{true}}}^\perp y\|$; (b) is valid since $P_{I_{\text{true}}}^\perp a_j = 0$ for all $j \in I_{\text{true}}$ and, therefore, $P_{I_{\text{true}}}^\perp x = 0$; and (c) follows from the fact that $\sqrt{m} P_{I_{\text{true}}}^\perp d$ is a unit-variance white random vector in an $m-k$-dimensional space, and therefore

$$\lim_{n\to\infty} \frac{m}{m-k} \left\|P_{I_{\text{true}}}^\perp d\right\|^2 = 1$$

a.s. and in probability. Combining (32)–(34) shows that

$$\liminf_{n\to\infty} \Delta^+ \geq 2 \tag{35}$$

where the limit is in probability.

B. Limit of Δ^-

For any $j \in K$, $P_K^\perp a_j = 0$. Therefore

$$P_K^\perp y = P_K^\perp \left(\sum_{j=1}^k a_j x_j + d\right) = x_1 P_K^\perp a_1 + P_K^\perp d.$$

Hence

$$\frac{\left|a_1' P_K^\perp y\right|^2}{a_1' P_K^\perp a_1} = \left|\left\|P_K^\perp a_1\right\| x_1 + v\right|^2 \tag{36}$$

where v is given by

$$v = a_1' P_K^\perp d / \left\|P_K^\perp a_1\right\|.$$

Since $P_K^\perp a_1 / \|P_K^\perp a_1\|$ is a random unit vector independent of d, and d has i.i.d. $\mathcal{N}(0, 1/m)$ components, the scalar v is distributed $\mathcal{N}(0, 1/m)$. Therefore

$$\sqrt{m} v / \log^{1/2}(n-k) \to 0 \tag{37}$$

in probability.

Also, a_1 is a Gaussian vector with variance $1/m$ in each component and $P_K^\perp$ is a projection onto an $(m-k+1)$-dimensional space. Hence

$$\lim_{n\to\infty} \frac{m\left\|P_K^\perp a_1\right\|^2}{m-k+1} = 1 \tag{38}$$

a.s. and in probability. Therefore

$$\begin{aligned}
&\limsup_{n\to\infty} \Delta^- \\
&\quad \overset{(a)}{=} \limsup_{n\to\infty} \left| \frac{\sqrt{m}\left\|P_K^\perp a_1\right\| x_1}{\log^{1/2}(n-k)} + \sqrt{\frac{m}{\log(n-k)}} v \right|^2 \\
&\quad \overset{(b)}{=} \limsup_{n\to\infty} \frac{m\left\|P_K^\perp a_1\right\|^2 |x_1|^2}{\log(n-k)} \\
&\quad \overset{(c)}{=} \limsup_{n\to\infty} \frac{(m-k+1)|x_1|^2}{\log(n-k)} \\
&\quad \overset{(d)}{=} \limsup_{n\to\infty} \frac{(\mathsf{SNR}\cdot\mathsf{MAR})(m-k+1)}{k\log(n-k)} \\
&\quad \overset{(e)}{<} 2(1-\delta),
\end{aligned} \tag{39}$$

where (a) follows from the definition of Δ^- and (36); (b) follows from (37); (c) follows from (38); (d) uses (27); and (e) uses (14a). Note that all the limits are in probability.

Comparing (35) and (39) proves (31), thus completing the proof.

Appendix C
Proof of Theorem 2

We will show that there exists a $\mu > 0$ such that, with high probability

$$\rho(j) > \mu \text{ for all } i \in I_{\text{true}} \tag{40a}$$

$$\rho(j) < \mu \text{ for all } j \notin I_{\text{true}}. \tag{40b}$$

Since $\delta > 0$, we can find an $\epsilon > 0$ such that

$$(1+\delta) \geq (1+\epsilon)^2. \tag{41}$$

Set the threshold level as

$$\mu = 2(1+\epsilon)(1+\mathsf{SNR})\frac{\log(n-k)}{m} \tag{42}$$

and define two probabilities corresponding to the two conditions in (40)

$$p_{\text{MD}} = \Pr\left(\rho(j) < \mu \text{ for some } i \in I_{\text{true}}\right) \tag{43a}$$

$$p_{\text{FA}} = \Pr\left(\rho(j) > \mu \text{ for some } j \notin I_{\text{true}}\right). \tag{43b}$$

The first probability p_{MD} is the probability of missed detection, i.e., the probability that the energy on one of the "true" vectors, a_i with $i \in I_{\text{true}}$, is below the threshold μ. The second probability p_{FA} is the false alarm probability, i.e., the probability that the energy on one of the "incorrect" vectors, a_j with $j \notin I_{\text{true}}$, is above the threshold μ. Since the correlation estimator detects the correct sparsity pattern when there are no missed vectors or false alarms, we have the bound

$$\Pr\left(\hat{I}_{\text{thresh}} \neq I_{\text{true}}\right) \leq p_{\text{MD}} + p_{\text{FA}}.$$

The result will be proven if we can show that p_{MD} and p_{FA} approach zero as m, n, and $k \to \infty$ satisfying (21). We analyze these two probabilities separately. Specifically, we will first see that choosing μ as in (42) ensures $p_{\text{FA}} \to 0$ as long as $n-k \to \infty$, regardless of the number of measurements. Then, we will see that (21) along with (42) ensures $p_{\text{MD}} \to 0$.

A. Limit of p_{FA}

Consider any index $j \notin I_{\text{true}}$, and define the random variable

$$u_j = \frac{m}{(1+\mathsf{SNR})}\rho(j) = \frac{m\left|a_j' y\right|^2}{(1+\mathsf{SNR})\left\|a_j\right\|^2}. \tag{44}$$

Since y is a linear combination of vectors $\{a_i\}_{i\in I_{\text{true}}}$ and the noise vector d, the vector a_j is independent of y. Also, y has independent Gaussian components with a per-component variance of $(1+\mathsf{SNR})/m$. It follows from Lemma 2(b) in Appendix D that each u_j is a chi-squared random variable with one degree of freedom. Since there are $n-k$ indices j not in I_{true}, Lemma 4 in Appendix D shows that

$$\limsup_{n\to\infty} \max_{j\notin I_{\text{true}}} \frac{u_j}{2\log(n-k)} \leq 1 \tag{45}$$

where the limit is in probability.

Therefore

$$\begin{aligned}
&\limsup_{n\to\infty} \max_{j\notin I_{\text{true}}} \frac{1}{\mu}\rho(j) \\
&\quad \overset{(a)}{=} \limsup_{n\to\infty} \max_{j\notin I_{\text{true}}} \frac{m\rho(j)}{2(1+\epsilon)(1+\mathsf{SNR})\log(n-k)} \\
&\quad \overset{(b)}{=} \limsup_{n\to\infty} \max_{j\notin I_{\text{true}}} \frac{u_j}{2(1+,\epsilon)\log(n-k)} \\
&\quad \overset{(c)}{\leq} \frac{1}{1+\epsilon},
\end{aligned}$$

where all limits are in probability and (a) follows from the definition of μ in (42); (b) follows from the definition of u_j in (44); and (c) follows from the limit in (45). Therefore

$$p_{\text{FA}} = \Pr\left(\max_{j\notin I_{\text{true}}} \rho(j) > \mu\right) \to 0.$$

B. Limit of p_{MD}

We first need the technical result that $\lim_{n\to\infty} \log(k)/m = 0$. To prove this, observe that

$$\begin{aligned}
\frac{\log(k)}{m} &\overset{(a)}{\leq} \frac{\log(k)\,\mathsf{SNR}\cdot\mathsf{MAR}}{2kL(n,k)(1+\mathsf{SNR})} \\
&\overset{(b)}{\leq} \frac{\log(k)}{2kL(n,k)} \overset{(c)}{\leq} \frac{1}{2k} \to 0
\end{aligned} \tag{46}$$

where (a) follows from (19); (b) follows from $\mathsf{SNR} \geq 0$ and $\mathsf{MAR} \in (0,1]$; and (c) follows from the definition of $L(n,k)$ in

(20) from which $L(n,k) > \log(k)$. Note that this limit involves entirely deterministic quantities.

Now consider any index $i \in I_{\text{true}}$. Observe that

$$a_i' y = \|a_i\|^2 x_i + a_i' e_i,$$

where

$$e_i = y - a_i x_i = \sum_{\ell \in I_{\text{true}}, \ell \neq i} a_\ell x_\ell + d.$$

Therefore

$$\begin{aligned}\frac{1}{\mu}\rho(i) &= \frac{1}{\mu\|a_i\|^2}|a_i' y|^2 \\ &= \left|\sqrt{A_i}\frac{x_i}{|x_i|} + \sqrt{B_i}\frac{(a_i' e_i)}{\|a_i\|\|e_i\|}\right|^2 \\ &\geq \left(\sqrt{A_i} - \sqrt{B_i z_i}\right)^2 \end{aligned} \tag{47}$$

where

$$A_i = \frac{1}{\mu}\|a_i\|^2|x_i|^2, \quad B_i = \frac{1}{\mu}\|e_i\|^2$$
$$z_i = \frac{|a_i' e_i|^2}{\|a_i\|^2\|e_i\|^2}.$$

We bound the terms A_i from below and B_i and z_i from above.

First consider A_i. Starting with the definition of A_i, we have the deterministic inequality

$$\begin{aligned} A_i &= \frac{1}{\mu}\|a_i\|^2|x_i|^2 \\ &\overset{(a)}{=} \frac{m}{2(1+\epsilon)(1+\mathsf{SNR})\log(n-k)}|x_i|^2\|a_i\|^2 \\ &\overset{(b)}{\geq} \frac{m\mathsf{SNR}\cdot\mathsf{MAR}}{2(1+\epsilon)k(1+\mathsf{SNR})\log(n-k)}\|a_i\|^2 \\ &\overset{(c)}{\geq} \frac{(1+\delta)L(n,k)}{(1+\epsilon)\log(n-k)}\|a_i\|^2 \\ &\overset{(d)}{\geq} (1+\epsilon)\frac{L(n,k)}{\log(n-k)}\|a_i\|^2 \end{aligned} \tag{48}$$

where (a) follows from (42); (b) follows from (8); (c) follows from (19); and (d) follows from (41). Also, since each vector a_i is an m-dimensional real vector with i.i.d. $\mathcal{N}(0, 1/m)$ components, Lemma 2(a) in Appendix D shows that $m\|a_i\|^2$ is a chi-squared random variable with m degrees of freedom. Now since there are k elements in I_{true}, condition (46) and Lemma 3 in Appendix D show that

$$\lim_{n\to\infty}\max_{i\in I_{\text{true}}}\|a_i\|^2 = 1$$

where the limit is in probability. Therefore, (48) implies that

$$\liminf_{n\to\infty}\min_{i\in I_{\text{true}}}\frac{\log(n-k)A_i}{L(n,k)} \geq 1+\epsilon \tag{49}$$

where again the limit is in probability.

For the term B_i, observe that each e_i is a Gaussian m-dimensional vector with independent components and total variance

$$\boldsymbol{E}\|e_i\|^2 \leq \boldsymbol{E}\|y\|^2 = 1 + \mathsf{SNR}.$$

Thus, as before, condition (46) and Lemma 3 show that

$$\limsup_{n\to\infty}\max_{i\in I_{\text{true}}}\frac{1}{1+\mathsf{SNR}}\|e_i\|^2 \leq 1$$

where the limit is in probability. Therefore, using the definition of μ in (42)

$$\begin{aligned}&\limsup_{n\to\infty}\max_{i\in I_{\text{true}}}\frac{\log(n-k)}{m}B_i \\ &\quad= \limsup_{n\to\infty}\max_{i\in I_{\text{true}}}\frac{\log(n-k)}{m\mu}\|e_i\|^2 \\ &\quad= \limsup_{n\to\infty}\max_{i\in I_{\text{true}}}\frac{1}{2(1+\epsilon)(1+\mathsf{SNR})}\|e_i\|^2 \\ &\quad\leq \frac{1}{2(1+\epsilon)} < \frac{1+\epsilon}{2}\end{aligned} \tag{50}$$

where the limit is in probability.

Finally, to bound z_i, Lemma 5 in Appendix D shows that z_i follows a $\text{Beta}(1, m-1)$ distribution. Since there are k terms in I_{true}, Lemma 6 and the condition (46) in Appendix D show that

$$\lim_{n\to\infty}\max_{i\in I_{\text{true}}}\frac{m}{2\log(k)}z_i \leq 1 \tag{51}$$

in probability.

Substituting (49)–(51) into (47), we have that

$$\begin{aligned}&\liminf_{n\to\infty}\min_{i\in I_{\text{true}}}\frac{1}{\mu}\rho(i) \\ &\quad\geq (1+\epsilon)\left(\sqrt{\frac{L(n,k)}{\log(n-k)}} - \sqrt{\frac{\log(k)}{\log(n-k)}}\right)^2 \\ &\quad\geq (1+\epsilon)\end{aligned}$$

where the limit is in probability and the last step follows from the definition of $L(n,k)$ in (20). This implies that

$$p_{\text{MD}} = \Pr\left(\min_{i\in I_{\text{true}}}\rho(i) \leq \mu\right) \to 0.$$

Hence, we have shown both $p_{\text{FA}} \to 0$ and $p_{\text{MD}} \to 0$ as $n \to \infty$, and the theorem is proven.

Appendix D
Tail Bounds of Chi-Squared and Beta Random Variables

The proofs of the main results above require a few standard tail bounds for chi-squared and beta random variables. A complete description of chi-squared and beta random variables can be found in [39]. We will omit or just provide some sketches of the proofs of the results in this section since they are all standard.

A random variable u has a *chi-squared* distribution with r degrees of freedom if it can be written as $u = \sum_{i=1}^r z_i^2$, where z_i are i.i.d. $\mathcal{N}(0,1)$. For this work, chi-squared random variables arise in two important instances.

Lemma 2: Suppose x is an r-dimensional real random vector with a Gaussian distribution $\mathcal{N}\left(0, \sigma^2 I_r\right)$. Then

(a) $\|x\|^2/\sigma^2$ is chi-squared with r degrees of freedom; and

(b) if y is any other r-dimensional random vector that is nonzero with probability one and independent of x, then the variable

$$u = \frac{|x'y|^2}{\sigma^2 \|y\|^2}$$

is chi-squared with one degree of freedom.

Proof: Part (a) follows from the fact that the norm $\|x\|^2/\sigma^2$ is a sum of squares of r unit-variance Gaussian random variables, one for each component of x/σ. Part (b) follows from the fact that $x'y/(\|y\|\sigma)$ is a unit-variance Gaussian random variable. □

The following two lemmas provide standard tail bounds.

Lemma 3: Suppose that for each n, $\{u_j^{(n)}\}_{j=1}^n$ is a set of chi-squared random variables with $r = r(n)$ degrees of freedom. The variables may be dependent. If

$$\lim_{n\to\infty} \log(n)/r(n) = 0$$

then

$$\lim_{n\to\infty} \frac{1}{r(n)} \max_{j=1,\ldots,n} u_j^{(n)} = \lim_{n\to\infty} \frac{1}{r(n)} \min_{j=1,\ldots,n} u_j^{(n)} = 1$$

where the limit is in probability.

Proof: A standard tail bound (see, for example [23]), shows that for any $\epsilon > 0$

$$\Pr\left(\frac{u_j^{(n)}}{r(n)} > 1+\epsilon\right) \le \exp(-\epsilon r(n)).$$

So, using the union bound

$$\Pr\left(\max_{j=1,\ldots,n} \frac{u_j^{(n)}}{r(n)} > 1+\epsilon\right) \le n\exp(-\epsilon r(n)) = \exp(-\epsilon r(n) + \log(n)) \to 0$$

where the limit is due to the fact that $\log(n)/r(n) \to 0$. This shows that

$$\limsup_{n\to\infty} \frac{1}{r(n)} \max_{j=1,\ldots,n} u_j^{(n)} \le 1$$

in probability. Similarly, using the tail bound that

$$\Pr\left(\frac{u_j^{(n)}}{r(n)} < 1-\epsilon\right) \le \exp(-\epsilon^2 r(n)/4)$$

one can show that

$$\limsup_{n\to\infty} \frac{1}{r(n)} \min_{j=1,\ldots,n} u_j^{(n)} \ge 1$$

in probability, and this proves the lemma. □

Lemma 4: Suppose that for each n, $\{u_j^{(n)}\}_{j=1}^n$ is a set of chi-squared random variables with one degree of freedom. Then

$$\limsup_{n\to\infty} \max_{j=1,\ldots,n} \frac{u_j^{(n)}}{2\log(n)} \le 1 \tag{52}$$

where the limit is in probability. If the variables are independent, then we have equality in the limit in that

$$\lim_{n\to\infty} \max_{j=1,\ldots,n} \frac{u_j^{(n)}}{2\log(n)} = 1 \tag{53}$$

in probability.

Proof: This uses similar tail bound arguments, so again we will just sketch the proof. Since each $u_j^{(n)}$ is a chi-squared random variable with one degree of freedom, we have the bound (see, for example, [40]), that for any $\nu > 0$

$$\Pr\left(u_j^{(n)} > \nu\right) = \operatorname{erfc}(\sqrt{\nu/2}) \le \frac{\exp(-\nu/2)}{\sqrt{\pi\nu/2}}$$

where erfc is the complementary error function. Combining this with the union bound, we see that for any $\epsilon > 0$

$$\Pr\left(\max_{j=1,\ldots,n} \frac{u_j^{(n)}}{2\log(n)} > 1+\epsilon\right) \le \frac{n\exp(-(1+\epsilon)\log(n))}{\sqrt{\pi(1+\epsilon)\log(n)}} = \frac{1}{n^\epsilon\sqrt{\pi(1+\epsilon)\log(n)}} \to 0.$$

This proves the limit (52).

For the other limit, we use the bound (also found in [40]) that for any $\nu > 0$

$$\Pr\left(u_j^{(n)} < \nu\right) = 1 - \operatorname{erfc}(\sqrt{\nu/2}) \le 1 - \frac{\exp(-\nu/2)}{\sqrt{\pi}\sqrt{2+\nu/2}}.$$

So, if the variables $u_j^{(n)}$, $j = 1, \ldots, n$, are independent

$$\begin{aligned}\Pr\left(\max_{j=1,\ldots,n} \frac{u_j^{(n)}}{2\log(n)} < 1-\epsilon\right) &= \left(1 - \operatorname{erfc}(\sqrt{2(1-\epsilon)\log(n)/2})\right)^n \\ &\le \left(1 - \frac{\exp(-(1-\epsilon)\log(n))}{\sqrt{\pi}\sqrt{2+(1-\epsilon)\log(n)}}\right)^n \\ &= \left(1 - \frac{1}{n^{1-\epsilon}\sqrt{\pi}\sqrt{2+(1-\epsilon)\log(n)}}\right)^n \to 0\end{aligned}$$

where the final limit can be shown for any $\epsilon \in (0,1)$ using L'Hôpital's rule. This shows that

$$\liminf_{n\to\infty} \max_{j=1,\ldots,n} \frac{u_j^{(n)}}{2\log(n)} \ge 1$$

in probability. Combining this with (52) proves (53). □

The next two lemmas concern certain beta-distributed random variables. A real-valued scalar random variable w follows a $\mathrm{Beta}(r,s)$ distribution if it can be written as

$w = u_r/(u_r + v_s)$, where the variables u_r and v_s are independent chi-squared random variables with r and s degrees of freedom, respectively. The importance of the beta distribution is given by the following lemma.

Lemma 5: Suppose x and y are independent random r-dimensional vectors with $x \sim \mathcal{N}\left(0, \sigma^2 I_r\right)$ and y having any distribution that is nonzero with probability one. Then the random variable

$$w = \frac{|x'y|^2}{||x||^2||y||^2}$$

is independent of x and follows a $\mathrm{Beta}(1, r-1)$ distribution.

Proof: This can be proven along the lines of the arguments in [15]. □

The following lemma provides a simple expression for the maxima of certain beta distributed variables.

Lemma 6: For each n, suppose $\{w_j^{(n)}\}_{j=1}^n$ is a set of $\mathrm{Beta}(1, r-1)$ random variables where $r = r(n)$. If

$$\lim_{n\to\infty} \log(n)/r(n) = 0 \tag{54}$$

then

$$\limsup_{n\to\infty} \frac{r(n)}{2\log(n)} \max_{j=1,\ldots,n} w_j^{(n)} \leq 1$$

in probability. If, in addition, the random variables $w_j^{(n)}$ are independent, then

$$\lim_{n\to\infty} \frac{r(n)}{2\log(n)} \max_{j=1,\ldots,n} w_j^{(n)} = 1,$$

in probability.

Proof: We can write $w_j^{(n)} = u_j^{(n)}/(u_j^{(n)} + v_j^{(n)})$ where $u_j^{(n)}$ and $v_j^{(n)}$ are independent chi-squared random variables with one and $r(n) - 1$ degrees of freedom, respectively. Let

$$U_n = \frac{1}{2\log(n)} \max_{j=1,\ldots,n} u_j^{(n)}$$
$$V_n = \frac{1}{r(n)} \min_{j=1,\ldots,n} v_j^{(n)}$$
$$T_n = \max_{j=1,\ldots,n} w_j^{(n)}.$$

The condition (54) and Lemma 3 show that $V_n \to 1$ in probability. Also, Lemma 4 shows that

$$\limsup_{n\to\infty} U_n \leq 1$$

in probability. Using these two limits along with (54) shows that

$$\limsup_{n\to\infty} \frac{r(n)T_n}{2\log(n)} \leq \limsup_{n\to\infty} \frac{U_n}{V_n + U_n \log(n)/r(n)} \leq \frac{1}{1 + (1)(0)} = 1$$

where the limit is in probability. The other parts of the lemma are proven similarly. □

ACKNOWLEDGMENT

The authors wish to thank the reviewers of this manuscript and of [41] for extraordinarily useful criticisms and suggestions. The authors also wish to thank John Sun and Martin Wainwright for comments and Martin Vetterli for his support, wisdom, and encouragement.

REFERENCES

[1] A. Miller, *Subset Selection in Regression*, ser. Monographs on Statistics and Applied Probability, 2nd ed. New York: Chapman & Hall/CRC, 2002.
[2] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On-Off Random Access Channels: A Compressed Sensing Framework," 2009 [Online]. Available: arXiv:0903.1022v1 [cs.IT]
[3] A. K. Fletcher, S. Rangan, and V. K. Goyal, "A sparsity detection framework for on-off random access channels," in *Proc. IEEE Int. Symp. Information Th.*, Seoul, Korea, Jun./Jul. 2009, pp. 169–173.
[4] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[5] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
[7] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[8] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[9] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[10] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[11] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
[12] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[13] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
[14] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Sparse approximation, denoising, and large random frames," in *Proc. Wavelets XI, Part of SPIE Optics & Photonics*, San Diego, CA, Jul./Aug. 2005, vol. 5914, pp. 172–181.
[15] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, Mar. 2006.
[16] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, Jan. 2009.
[17] M. J. Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Recovery of Sparsity," Dept. Statistics, Univ. Calif., Berkeley, , 2006 [Online]. Available: arXiv:math.ST/0605740 v1 30
[18] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[19] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.
[20] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4036–4048, Sep. 2006.
[21] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. 44th Ann. Allerton Conf. Communications, Control and Computing*, Monticello, IL, Sep. 2006.
[22] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Rate-distortion bounds for sparse approximation," in *Proc. IEEE Statistical Signal Processing Workshop*, Madison, WI, Aug. 2007, pp. 254–258.
[23] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
[24] V. K. Goyal, A. K. Fletcher, and S. Rangan, "Compressive sampling and lossy compression," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.
[25] G. Reeves, "Sparse Signal Sampling Using Noisy Linear Projections," Univ. Calif., Berkeley, Dept. Elec. Eng. and Comp. Sci,, Tech. Rep. UCB/EECS-2008-3, 2008.

[26] M. Akçakaya and V. Tarokh, "Shannon Theoretic Limits on Noisy Compressive Sampling" 2007 [Online]. Available: arXiv:0711.0366v1 [cs.IT]
[27] M. Akçakaya and V. Tarokh, "Noisy compressive sampling limits in linear and sublinear regimes," in *Proc. Conf. Information Sciences & Systems*, Princeton, NJ, Mar. 2008, pp. 1–4.
[28] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.
[29] A. K. Fletcher and S. Rangan, "Orthogonal matching pursuit from noisy random measurements: A new analysis," in *Proc. conf. Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2009.
[30] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Contr.*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.
[31] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Signals, Systems, and Computing*, Pacific Grove, CA, Nov. 1993, vol. 1, pp. 40–44.
[32] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decomposition," *Opt. Eng.*, vol. 37, no. 7, pp. 2183–2191, Jul. 1994.
[33] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[34] S. Aeron, M. Zhao, and V. Saligrama, "Fundamental Limits on Sensing Capacity for Sensor Networks and Compressed Sensing," 2008 [Online]. Available: arXiv:0804.3439v3 [cs.IT]
[35] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and Sufficient Conditions on Sparsity Pattern Recovery," 2008 [Online]. Available: arXiv:0804.1839v1 [cs.IT]
[36] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-Theoretic Limits on Sparse Signal Recovery: Dense Versus Sparse Measurement Matrices," 2008 [Online]. Available: arXiv:0806.0604v1 [math.ST]
[37] M. F. Duarte, S. Sarvotham, D. Baron, W. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Conf. Rec. Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Oct./Nov. 2005, pp. 1537–1541.
[38] S. Aeron, M. Zhao, and V. Saligrama, "On Sensing Capacity of Sensor Networks for the Class of Linear Observation, Fixed SNR Models," 2007 [Online]. Available: arXiv:0704.3434v3 [cs.IT]
[39] M. Evans, N. Hastings, and J. B. Peacock, *Statistical Distributions*, 3rd ed. New York: Wiley, 2000.
[40] J. Spanier and K. B. Oldham, *An Atlas of Functions*. Washington: Hemisphere Publishing, 1987.
[41] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Resolution limits of sparse coding in high dimensions," in *Proc. Conf. Neural Information Processing and Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., Vancouver, BC, Canada, Dec. 2008, pp. 449–456.

Alyson K. Fletcher (S'03–M'04) received the B.S. degree in mathematics from the University of Iowa, Iowa City. She received the M.S. degree in electrical engineering in 2002, and the M.A. degree in mathematics, as well as the Ph.D. degree in electrical engineering, from the University of California, Berkeley, both in 2006.

She is currently a President's Postdoctoral Fellow at the University of California, Berkeley. Her research interests include estimation, image processing, statistical signal processing, sparse approximation, wavelets, and control theory.

Dr. Fletcher is a member of SWE, SIAM, and Sigma Xi. In 2005, she received the University of California Eugene L. Lawler Award, the Henry Luce Foundation's Clare Boothe Luce Fellowship, and the Soroptimist Dissertation Fellowship.

Sundeep Rangan received the B.A.Sc. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1997, respectively.

He was then a Postdoctoral Research Fellow at the University of Michigan, Ann Arbor. He joined the Wireless Research Center at Bell Laboratories, Lucent Technologies, in 1998. In 2000, he cofounded, with four others, Flarion Technologies which developed and commercialized an early OFDM cellular wireless data system. Flarion was acquired by Qualcomm in 2006. He is currently a Director of Engineering at Qualcomm Technologies, where he is involved in the development of next generation cellular wireless systems. His research interests include communications, wireless systems, information theory, estimation, and control theory.

Vivek K Goyal (S'92–M'98–SM'03) received the B.S. degree in mathematics and the B.S.E. degree in electrical engineering (both with highest distinction) from the University of Iowa, Iowa City. He received the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1995 and 1998, respectively.

He was a Member of Technical Staff in the Mathematics of Communications Research Department of Bell Laboratories, Lucent Technologies, 1998–2001; and a Senior Research Engineer for Digital Fountain, Inc., 2001–2003. He is currently Esther and Harold E. Edgerton Associate Professor of Electrical Engineering at the Massachusetts Institute of Technology, Cambridge. His research interests include source coding theory, sampling, quantization, and information gathering and dispersal in networks.

Dr. Goyal is a member of Phi Beta Kappa, Tau Beta Pi, Sigma Xi, Eta Kappa Nu and SIAM. In 1998, he received the Eliahu Jury Award of the University of California, Berkeley, awarded to a graduate student or recent alumnus for outstanding achievement in systems, communications, control, or signal processing. He was also awarded the 2002 IEEE Signal Processing Society Magazine Award and an NSF CAREER Award. He served on the IEEE Signal Processing Society's Image and Multiple Dimensional Signal Processing Technical Committee and is a permanent Conference Co-Chair of the SPIE Wavelets conference series.