Information-Theoretic Limits on Sparse Signal Recovery: Dense versus Sparse Measurement Matrices

Wei Wang, *Member, IEEE*, Martin J. Wainwright, *Member, IEEE*, and Kannan Ramchandran, *Fellow, IEEE*

***Abstract*—We study the information-theoretic limits of exactly recovering the support set of a sparse signal, using noisy projections defined by various classes of measurement matrices. Our analysis is high-dimensional in nature, in which the number of observations n, the ambient signal dimension p, and the signal sparsity k are all allowed to tend to infinity in a general manner. This paper makes two novel contributions. First, we provide sharper necessary conditions for exact support recovery using general (including non-Gaussian) dense measurement matrices. Combined with previously known sufficient conditions, this result yields sharp characterizations of when the optimal decoder can recover a signal for various scalings of the signal sparsity k and sample size n, including the important special case of linear sparsity ($k = \Theta(p)$) using a linear scaling of observations ($n = \Theta(p)$). Our second contribution is to prove necessary conditions on the number of observations n required for asymptotically reliable recovery using a class of γ-sparsified measurement matrices, where the measurement sparsity parameter $\gamma(n, p, k) \in (0, 1]$ corresponds to the fraction of nonzero entries per row. Our analysis allows general scaling of the quadruplet (n, p, k, γ), and reveals three different regimes, corresponding to whether measurement sparsity has no asymptotic effect, a minor effect, or a dramatic effect on the information-theoretic limits of the subset recovery problem.**

***Index Terms*—ℓ_1-Relaxation, compressed sensing, Fano's method, high-dimensional statistical inference, information-theoretic bounds, sparse approximation, sparse random matrices, sparsity recovery, subset selection, support recovery.**

I. Introduction

SPARSITY recovery refers to the problem of estimating the support of a p-dimensional but k-sparse vector $\beta \in \mathbb{R}^p$, based on a set of n noisy linear observations. The sparsity recovery problem is of broad interest, arising in subset selection in regression [19], graphical model selection [18], group testing, signal denoising [8], sparse approximation [20], and compressive sensing [11], [7]. A large body of recent work (e.g., [6]–[8], [11], [12], [18], [27], [28], and [30]) has analyzed the performance of computationally tractable methods, in particular based on ℓ_1 or other convex relaxations, for estimating high-dimensional sparse signals. Such results have established conditions, on signal sparsity and the choice of measurement matrices, under which a given recovery method succeeds with high probability.

Of complementary interest are the information-theoretic limits of the sparsity recovery problem, which apply to the performance of any procedure regardless of its computational complexity. Such analysis has two purposes: first, to demonstrate where known polynomial-time methods achieve the information-theoretic bounds, and second, to reveal situations in which current methods are sub-optimal. An interesting question which arises in this context is the effect of the choice of measurement matrix on the information-theoretic limits. As we will see, the standard Gaussian measurement ensemble achieves an optimal scaling of the number of observations required for recovery. However, this choice produces highly dense matrices, which may lead to prohibitively high computational complexity and storage requirements.[1] In contrast, sparse measurement matrices directly reduce encoding and storage costs, and can also lead to fast decoding algorithms by exploiting problem structure (see Section I-B for a brief overview of the growing literature in this area). In addition, measurement sparsity can be used to lower communication cost and latency in distributed sensor network and streaming applications. On the other hand, measurement sparsity can potentially reduce statistical efficiency by requiring more observations to recover the signal. Intuitively, the nonzeros in the signal may rarely align with the nonzeros in a sparse measurement matrix.[2] Therefore, an important question is to characterize the trade-off between measurement sparsity and statistical efficiency.

With this motivation, this paper makes two contributions. First, we derive sharper necessary conditions for exact support recovery, applicable to a general class of dense measurement matrices (including non-Gaussian ensembles). In conjunction with the sufficient conditions from previous work [29], this analysis provides a sharp characterization of necessary and

Manuscript received August 01, 2008; revised September 01, 2009. Current version published May 19, 2010. The work of W. Wang and K. Ramchandran was supported by NSF Grant CCF-0635114. The work of M. J. Wainwright was supported by NSF Grants CAREER-CCF-0545862 and DMS-0605165. This work was posted in May 2008 as Technical Report 754 in the Department of Statistics, University of California, Berkeley, and was posted as arXiv:0806.0604 [math.ST]. The material in this work was presented in part at the IEEE International Symposium on Information Theory, Toronto, ON, Canada, July 2008.

W. Wang and K. Ramchandran are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: wangwei@eecs.berkeley.edu; kannanr@eecs.berkeley.edu).

M. J. Wainwright is with the Department of Statistics and the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: wainwrig@eecs.berkeley.edu).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2046199

[1]For example, ℓ_1-recovery methods based on linear programming have complexity $O(p^3)$ in the signal dimension p.

[2]Note, however, that misalignments between the measurements and the signal still reveal *some* information about the locations of the nonzeros in the signal.

sufficient conditions for various sparsity regimes. Our second contribution is to address the effect of measurement sparsity, meaning the fraction $\gamma \in (0, 1]$ of nonzeros per row in the matrices used to collect measurements. We derive lower bounds on the number of observations required for exact sparsity recovery, as a function of the signal dimension p, signal sparsity k, and measurement sparsity γ. This analysis highlights a trade-off between the statistical efficiency of a measurement ensemble and the computational complexity associated with storing and manipulating it.

The remainder of the paper is organized as follows. We first define our problem formulation in Section I-A, and then discuss our contributions and some connections to related work in Section I-B. Section II provides precise statements of our main results, as well as a discussion of their consequences. Section III provides proofs of the necessary conditions for various classes of measurement matrices, while proofs of more technical lemmas are given in Appendices A–F. Finally, we conclude and discuss open problems in Section IV.

A. Problem Formulation

Let $\beta \in \mathbb{R}^p$ be a fixed but unknown vector, with the support set of β defined as

$$S(\beta) := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}. \tag{1}$$

We refer to $k := |S(\beta)|$ as the *signal sparsity*, and p as the *signal dimension*. Suppose we are given a vector of n noisy observations $Y \in \mathbb{R}^n$, of the form

$$Y = X\beta + W \tag{2}$$

where $X \in \mathbb{R}^{n \times p}$ is the known measurement matrix, and $W \sim N\left(0, \sigma^2 I_{n \times n}\right)$ is additive Gaussian noise. Our goal is to perform exact recovery of the underlying sparsity pattern $S(\beta)$, which we refer to as the sparsity recovery problem. The focus of this paper is to find conditions on the model parameters (n, p, k) that are necessary for any method to successfully recover the support set $S(\beta)$. Our results apply to various classes of dense and γ-sparsified measurement matrices, which will be defined in Section II.

1) Classes of Signals: The difficulty of sparsity recovery from noisy measurements naturally depends on the minimum value of β on its support, defined by the function

$$\lambda^*(\beta) := \min_{i \in S(\beta)} |\beta_i|. \tag{3}$$

In this paper, we study the class of signals parameterized by a lower bound λ on the minimum value

$$\mathcal{C}_{p,k}(\lambda) := \{\beta \in \mathbb{R}^p \mid |S(\beta)| = k, \lambda^*(\beta) \geq \lambda\}. \tag{4}$$

The associated class of sparsity patterns $\mathcal{C}_{p,k}$ is the collection of all $N = \binom{p}{k}$ possible subsets of size k. We assume without loss of generality that the noise variance $\sigma^2 = 1$, since any scaling of σ can be accounted for in the scaling of β.

2) Decoders and Error Criterion: Suppose that nature chooses some vector β from the signal class $\mathcal{C}_{p,k}(\lambda)$. The statistician observes n samples $Y = X\beta + W \in \mathbb{R}^n$ and tries to infer the underlying sparsity pattern $S(\beta)$. The results of this paper apply to arbitrary decoders. A decoder is a mapping $g : \mathbb{R}^n \to \mathcal{C}_{p,k}$ from the observations Y to an estimated subset $\widehat{S} = g(Y)$. We measure the error between the estimate $\widehat{S}$ and the true support $S(\beta)$ using the $\{0, 1\}$-valued loss function $\mathbb{I}[g(Y) \neq S(\beta)]$, which corresponds to a standard model selection error criterion. The probability of incorrect subset selection is then the associated 0–1 risk $\mathbb{P}[g(Y) \neq S \mid S(\beta) = S]$, where the probability is taken over the measurement noise W and the choice of random measurement matrix X. We define the maximal probability of error over the class $\mathcal{C}_{p,k}(\lambda)$ as

$$\omega(g) := \max_{\beta \in \mathcal{C}_{p,k}(\lambda)} \mathbb{P}[g(Y) \neq S \mid S(\beta) = S]. \tag{5}$$

We say that sparsity recovery is asymptotically reliable over the signal class $\mathcal{C}_{p,k}(\lambda)$ if $\omega(g) \to 0$ as $n \to \infty$.

With this setup, our goal is to find necessary conditions on the parameters $(n, p, k, \lambda, \gamma)$ that any decoder, regardless of its computational complexity, must satisfy for asymptotically reliable recovery to be possible. We are interested in lower bounds on the number of measurements n in general settings where both the signal sparsity k and the measurement sparsity γ are allowed to scale with the signal dimension p.

B. Past Work and Our Contributions

One body of past work [14], [24], [1] has focused on the information-theoretic limits of sparse estimation under ℓ_2 and other distortion metrics, using power-based SNR measures of the form

$$\mathrm{SNR} := \frac{\mathbb{E}\left[\|X\beta\|_2^2\right]}{\mathbb{E}\left[\|W\|_2^2\right]} = \|\beta\|_2^2. \tag{6}$$

(Note that the second equality assumes that the noise variance $\sigma^2 = 1$, and that the measurement matrix is standardized, with each element X_{ij} having zero mean and variance one). It is important to note that the power-based SNR (6), though appropriate for ℓ_2-distortion, is not the key parameter for the support recovery problem. Although the minimum value is related to this power-based measure by the inequality $k\lambda^2 \leq \mathrm{SNR}$, for the ensemble of signals $\mathcal{C}_{p,k}(\lambda)$ defined in (4), the ℓ_2-based SNR (6) can be made arbitrarily large while still having one coefficient β_i equal to the minimum value (assuming that $k > 1$). Consequently, as our results show, it is possible to generate problem instances for which support recovery is arbitrarily difficult—in particular, by sending $\lambda \to 0$ at an arbitrarily rapid rate—even as the power-based SNR (6) becomes arbitrarily large.

The paper [29] was the first to consider the information-theoretic limits of exact subset recovery using standard Gaussian measurement ensembles, explicitly identifying the minimum value λ as the key parameter. This analysis yielded necessary and sufficient conditions on general quadruples (n, p, k, λ) for asymptotically reliable recovery. Subsequent work on the problem has yielded sharper conditions for standard Gaussian ensembles [22], [3], [13], [2], and extended this type of analysis to the criterion of partial support recovery [3], [22]. In this paper (initially posted as [32]), we consider only exact support recovery, but provide results for general dense measurement ensembles, including non-Gaussian matrices. In conjunction

with known sufficient conditions [29], one consequence of our first main result (Theorem 1, below) is a set of sharp necessary and sufficient conditions for the optimal decoder to recover the support of a signal with linear sparsity ($k = \Theta(p)$), using only a linear fraction of observations ($n = \Theta(p)$). As we discuss at more length in Section II-A, for the special case of the standard Gaussian ensemble, Theorem 1 also recovers some results independently obtained in past work by Reeves [22], and concurrent work by Fletcher *et al.* [13] and Aeron *et al.* [2].

In addition, this paper addresses the effect of measurement sparsity, which we assess in terms of the fraction $\gamma \in (0,1]$ of nonzeros per row of the the measurement matrix X. In the noiseless setting, a growing body of work has examined computationally efficient recovery methods based on sparse measurement matrices, including work inspired by expander graphs and coding theory [25], [33], [4], as well as dimension-reducing embeddings and sketching [9], [15], [31]. In addition, some results have been shown to be stable in the ℓ_2 or ℓ_1 norm in the presence of noise [9], [4]; note, however, that ℓ_2/ℓ_1 stability does not guarantee exact recovery of the support set. In the noisy setting, the paper [1] provides results for sparse measurements and distortion-type error metrics (using a power-based SNR), as opposed to the subset recovery metric of interest here. For the noisy observation model (2), some concurrent work [21] provides sufficient conditions for support recovery using the Lasso (i.e., ℓ_1-constrained quadratic programming) for appropriately sparsified ensembles. These results can be viewed as complementary to the information-theoretic analysis of this paper, in which we characterize the inherent trade-off between measurement sparsity and statistical efficiency. More specifically, our second main result (Theorem 2, below) provides necessary conditions for exact support recovery using γ-sparsified Gaussian measurement matrices [defined in (7)], for general scalings of the parameters $(n, p, k, \lambda, \gamma)$. This analysis reveals three regimes of interest, corresponding to whether measurement sparsity has no asymptotic effect, a small effect, or a significant effect on the number of measurements necessary for recovery. Thus, there exist regimes in which measurement sparsity fundamentally alters the ability of any method to decode.

II. Main Results and Consequences

In this section, we state our main results, and discuss some of their consequences. Our analysis applies to random ensembles of measurement matrices $X \in \mathbb{R}^{n \times p}$, where each entry X_{ij} is drawn i.i.d. from some underlying distribution. The most commonly studied random ensemble is the standard Gaussian case, in which each $X_{ij} \sim N(0,1)$. Note that this choice generates a highly dense measurement matrix X, with np nonzero entries. Our first result (Theorem 1) applies to more general ensembles that satisfy the moment conditions $\mathbb{E}[X_{ij}] = 0$ and $\mathrm{var}(X_{ij}) = 1$, which allows for a variety of non-Gaussian distributions (e.g., uniform, Bernoulli, etc.).[3] In addition, we also derive results (Theorem 2) for γ-sparsified matrices X, in which each entry X_{ij} is i.i.d. drawn according to

$$X_{ij} = \begin{cases} N\left(0, \frac{1}{\gamma}\right), & \text{w.p. } \gamma \\ 0, & \text{w.p. } 1-\gamma. \end{cases} \tag{7}$$

Note that when $\gamma = 1$, the distribution in (7) is exactly the standard Gaussian ensemble. We refer to the sparsification parameter $\gamma \in (0,1]$ as the *measurement sparsity*. Our analysis allows this parameter to vary as a function of (n, p, k).

A. Tighter Bounds on Dense Ensembles

We begin by stating a set of necessary conditions on (n, p, k, λ) for asymptotically reliable recovery with any method, which apply to general ensembles of zero-mean and unit-variance measurement matrices. In addition to the standard Gaussian ensemble ($X_{ij} \sim N(0,1)$), this result also covers matrices from other common ensembles (e.g., Bernoulli $X_{ij} \in \{-1, +1\}$). Furthermore, our analysis can be extended to matrices with independent rows drawn from any distribution with zero mean and covariance matrix Σ (see Appendix F).

Theorem 1 (General Ensembles): Let the measurement matrix $X \in \mathbb{R}^{n \times p}$ be drawn with i.i.d. elements from any distribution with zero mean and unit variance. Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}_{p,k}(\lambda)$ is

$$n > \max\{f_1(p,k,\lambda), \ldots, f_k(p,k,\lambda), k\} \tag{8}$$

where

$$f_m(p,k,\lambda) := \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log\left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right)} \tag{9}$$

for $m = 1, \ldots, k$.

The proof of Theorem 1, given in Section III, uses Fano's method [10], [16], [17], [34], [35] to bound the probability of error in a restricted ensemble, which can then be viewed as a type of channel coding problem. Moreover, the proof constructs a family of restricted ensembles that sweeps the range of possible overlaps between subsets, and tries to capture the difficulty of distinguishing between subsets at various distances.

We now consider some consequences of the necessary conditions in Theorem 1 under two scalings of the signal sparsity: the regime of linear signal sparsity, in which $k/p = \alpha$ for some $\alpha \in (0,1)$, and the regime of sublinear signal sparsity, meaning $k/p \to 0$. In particular, the necessary conditions in Theorem 1 can be compared against the sufficient conditions in Wainwright [29] for exact support recovery using the standard Gaussian ensemble, as shown in Table I. This comparison reveals that Theorem 1 generalizes and strengthens earlier results on necessary conditions for subset recovery [29]. We obtain tight scalings of the necessary and sufficient conditions in the regime of linear signal sparsity (meaning $k/p = \alpha$), under various scalings of the minimum value λ (shown in the first three rows of Table I).

[3] In fact, our results can be further generalized to ensembles of matrices which have independent rows drawn from any distribution with zero mean and covariance matrix Σ (see Theorem 3 in Appendix F).

TABLE I
TIGHT SCALINGS OF THE NECESSARY AND SUFFICIENT CONDITIONS ON THE NUMBER OF OBSERVATIONS n REQUIRED FOR EXACT SUPPORT RECOVERY ARE OBTAINED IN SEVERAL REGIMES OF INTEREST

	Necessary conditions (Theorem 1)	Sufficient conditions (Wainwright [29])
$k=\Theta(p)$ $\lambda^2=\Theta(\frac{1}{k})$	$\Theta(p\log p)$	$\Theta(p\log p)$
$k=\Theta(p)$ $\lambda^2=\Theta(\frac{\log k}{k})$	$\Theta(p)$	$\Theta(p)$
$k=\Theta(p)$ $\lambda^2=\Theta(1)$	$\Theta(p)$	$\Theta(p)$
$k=o(p)$ $\lambda^2=\Theta(\frac{1}{k})$	$\Theta(k\log(p-k))$	$\Theta(k\log(p-k))$
$k=o(p)$ $\lambda^2=\Theta(\frac{\log k}{k})$	$\max\left\{\Theta\left(\frac{k\log\frac{p}{k}}{\log\log k}\right),\Theta\left(\frac{k\log(p-k)}{\log k}\right)\right\}$	$\Theta\left(k\log\frac{p}{k}\right)$
$k=o(p)$ $\lambda^2=\Theta(1)$	$\max\left\{\Theta\left(\frac{k\log\frac{p}{k}}{\log k}\right),\Theta(k)\right\}$	$\Theta\left(k\log\frac{p}{k}\right)$

We also obtain tight scaling conditions in the regime of sublinear signal sparsity (in which $k/p\to 0$), when $k\lambda^2=\Theta(1)$ (as shown in row 4 of Table I). There remains a slight gap, however, in the sublinear sparsity regime when $k\lambda^2\to\infty$ (see bottom two rows in Table I).

In the regime of linear sparsity, Wainwright [29] showed, by direct analysis of the optimal decoder, that the scaling $\lambda^2=\Omega(\log(k)/k)$ is sufficient for exact support recovery using a linear fraction $n=\Theta(p)$ of observations. Combined with the necessary condition in Theorem 1, we obtain the following corollary that provides a sharp characterization of the linear-linear regime.

Corollary 1: Consider the regime of linear sparsity, meaning that $k/p=\alpha\in(0,1)$, and suppose that a linear fraction $n=\Theta(p)$ of observations are made. Then the optimal decoder can recover the support exactly if and only if $\lambda^2=\Omega(\log k/k)$.

Theorem 1 has some consequences related to results proved in recent and concurrent work. Reeves and Gastpar [22] have shown that in the regime of linear sparsity $k/p=\alpha>0$, and for standard Gaussian measurements, if any decoder is given only a linear fraction sample size (meaning that $n=\Theta(p)$), then one must have $k\lambda^2\to+\infty$ in order to recover the support exactly. This result is one corollary of Theorem 1, since if $\lambda^2=\Theta(1/k)$, then we have

$$n>\frac{\log(p-k+1)-1}{\frac{1}{2}\log(1+\Theta(1/k))}=\Omega(k\log(p-k))\gg\Theta(p)$$

so that the scaling $n=\Theta(p)$ is precluded. In concurrent work, Fletcher *et al.* [13] used direct methods to show that for the special case of the standard Gaussian ensemble, the number of observations must satisfy $n>\Omega\left(\frac{\log(p-k)}{\lambda^2}\right)$. The qualitative form of this bound follows from our lower bound $f_1(p,k,\lambda)$, which holds for standard Gaussian ensembles as well as more general (non-Gaussian) ensembles. However, we note that the direct methods used by Fletcher *et al.* [13] yield better control of the constant prefactors for the standard Gaussian ensemble. Similarly, concurrent work by Aeron *et al.* [2] showed that in the regime of linear sparsity (i.e., $k=\Theta(p)$) and for standard Gaussian measurements, the number of observations must satisfy $n>\Omega\left(\frac{\log p}{\lambda^2}\right)$. This result also follows as a consequence of our lower bound $f_1(p,k,\lambda)$.

The results in Theorem 1 can also be compared to an intuitive bound based on classical channel capacity results, as pointed out previously by various researchers (e.g., [24] and [3]). Consider a restricted problem, in which the values associated with each possible sparsity pattern on β are fixed and known at the decoder. Then support recovery can be viewed as a type of channel coding problem, in which the $N=\binom{p}{k}$ possible support sets of β correspond to messages to be sent over a Gaussian channel. Suppose each support set S is encoded as the codeword $X\beta$, where X has i.i.d. Gaussian entries. The effective code rate is then $R=\frac{\log\binom{p}{k}}{n}$, and by standard Gaussian channel capacity results, we have the lower bound

$$n>\frac{\log\binom{p}{k}}{\frac{1}{2}\log\left(1+||\beta||_2^2\right)}. \tag{10}$$

This bound is tight for $k=1$ and Gaussian measurements, but loose in general. As Theorem 1 clarifies, there are additional elements in the support recovery problem that distinguish it from a standard Gaussian coding problem: first, the signal power $||\beta||_2^2$ does not capture the inherent problem difficulty for $k>1$, and second, there is overlap between support sets for $k>1$. Note that $||\beta||_2^2\geq k\lambda^2$ (with equality in the case when $|\beta_j|=\lambda$ for all indices $j\in S$), so that Theorem 1 is strictly tighter than the intuitive bound (10). Moreover, by fixing the value of β at $(k-1)$ indices to λ and allowing the last component of β to tend to infinity, we can drive the power $||\beta||_2^2$ to infinity, while still having a non-trivial lower bound in Theorem 1.

B. Effect of Measurement Sparsity

We now turn to the effect of measurement sparsity on subset recovery, considering in particular the γ-sparsified ensemble (7). Since each X_{ij} has zero mean and unit variance for all choices of γ by construction, Theorem 1 applies to the γ-sparsified Gaussian ensemble (7); however, it yields necessary conditions that are independent of γ. Intuitively, it is clear that the procedure of γ-sparsification should cause deterioration in support

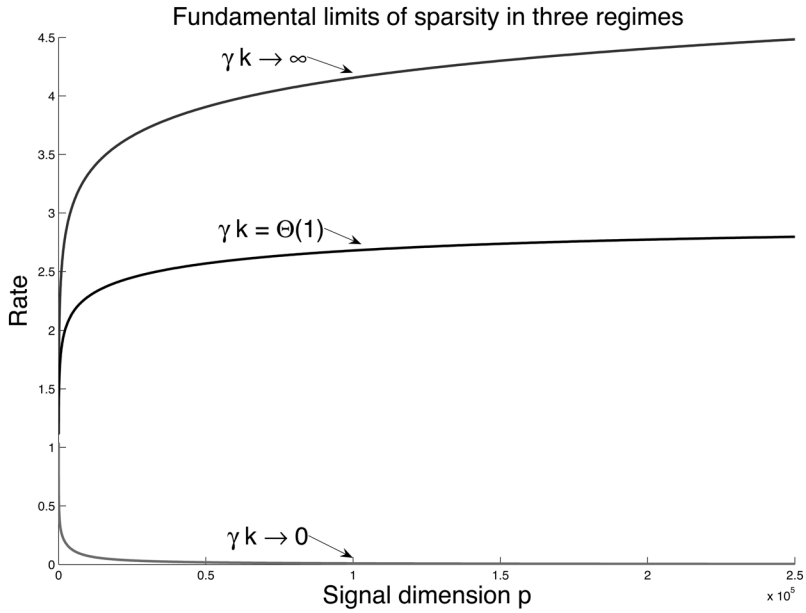

Fig. 1. Rate $R = \frac{\log \binom{p}{k}}{n}$, defined as the logarithm of the number of possible subsets the decoder can reliably estimate based on n observations, is plotted using (12) in three regimes, depending on how the quantity γk scales. In particular, γk corresponds to the average number of nonzeros in β that align with the nonzeros in each row of the measurement matrix.

recovery. Indeed, the following result provides more refined bounds that capture the effects of γ-sparsification. We first state a set of necessary conditions on $(n, p, k, \lambda, \gamma)$ in general form, and subsequently bound these conditions in different regimes of sparsity. Let $\phi(\mu, \sigma^2)$ denote the Gaussian density with mean μ and variance σ^2, and define the family of mixture distributions $\{\overline{\psi}_m\}_{m=1,\ldots,k}$ with

$$\overline{\psi}_m := \sum_{\ell=0}^{m} \binom{m}{\ell} \gamma^\ell (1-\gamma)^{m-\ell} \phi\left(0, 1 + \frac{\ell\lambda^2}{\gamma}\right). \tag{11}$$

Furthermore, let $h(\cdot)$ denote the differential entropy functional. With this notation, we have the following result.

Theorem 2 (Sparse Ensembles): Let the measurement matrix $X \in \mathbb{R}^{n\times p}$ be drawn with i.i.d. elements from the γ-sparsified Gaussian ensemble (7). Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}_{p,k}(\lambda)$ is

$$n > \max\{g_1(p, k, \lambda, \gamma), \ldots, g_k(p, k, \lambda, \gamma), k\} \tag{12}$$

where

$$g_m(p, k, \lambda, \gamma) := \frac{\log\binom{p-k+m}{m} - 1}{h(\overline{\psi}_m) - \frac{1}{2}\log(2\pi\mathrm{e})} \tag{13}$$

for $m = 1, \ldots, k$.

The proof of Theorem 2, given in Section III, again uses Fano's inequality, but explicitly analyzes the effect of measurement sparsification on the entropy of the observations. The necessary condition in Theorem 2 is plotted in Fig. 1, showing distinct regimes of behavior depending on how the quantity γk scales, where $\gamma \in (0, 1]$ is the measurement sparsification parameter and k is the signal sparsity index. In order to characterize the regimes in which measurement sparsity begins to degrade the recovery performance of any decoder, Corollary 2 below further bounds the necessary conditions in Theorem 2 in three cases.

Corollary 2 (Three Regimes): For any scalar γ, let $H_{\text{binary}}(\gamma)$ denote the entropy of a $\text{Bernoulli}(\gamma)$ variate. The necessary conditions in Theorem 2 can be simplified as follows.

(a) If $\gamma m \to \infty$, then

$$g_m(p, k, \lambda, \gamma) \geq \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log(1 + m\lambda^2)}. \tag{14a}$$

(b) If $\gamma m = \tau$ for some constant τ, then

$$g_m(p, k, \lambda, \gamma) \geq \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\tau\log\left(1 + \frac{m\lambda^2}{\tau}\right) + C} \tag{14b}$$

where $C = \frac{1}{2}\log\left(2\pi\mathrm{e}\left(\tau + \frac{1}{12}\right)\right)$ is a constant.

(c) If $\gamma m \to 0$, then

$$g_m(p, k, \lambda, \gamma) \geq \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\gamma m\log\left(1 + \frac{\lambda^2}{\gamma}\right) + mH_{\text{binary}}(\gamma)}. \tag{14c}$$

Corollary 2 reveals three regimes of behavior, defined by the scaling of the measurement sparsity γ and the signal sparsity k. Intuitively, γk is the average number of nonzeros in β that align with the nonzeros in each row of the measurement matrix. If $\gamma k \to \infty$ as $p \to \infty$, then the recovery threshold (14a) is of the same order as the threshold for dense measurement ensembles. In this regime, sparsifying the measurement ensemble has no asymptotic effect on performance. In sharp contrast, if $\gamma k \to 0$ sufficiently fast as $p \to \infty$, then the denominator in (14c) goes to zero, and the recovery threshold changes fundamentally compared to the dense case. Hence, the number of measurements that any decoder needs in order to reliably recover increases

TABLE II
NECESSARY CONDITIONS ON THE NUMBER OF OBSERVATIONS n REQUIRED FOR EXACT SUPPORT RECOVERY IS SHOWN IN DIFFERENT REGIMES OF THE PARAMETERS (p, k, λ, γ)

Necessary conditions (Theorem 2)	$k = o(p)$	$k = \Theta(p)$
$\lambda^2 = \Theta(\frac{1}{k})$ $\gamma = o(\frac{1}{k \log k})$	$\Theta\left(\frac{k \log(p-k)}{\gamma k \log \frac{1}{\gamma}}\right)$	$\Theta\left(\frac{p \log p}{\gamma p \log \frac{1}{\gamma}}\right)$
$\lambda^2 = \Theta(\frac{1}{k})$ $\gamma = \Omega(\frac{1}{k \log k})$	$\Theta(k \log(p-k))$	$\Theta(p \log p)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = o(\frac{1}{k \log k})$	$\Theta\left(\frac{k \log(p-k)}{\gamma k \log \frac{1}{\gamma}}\right)$	$\Theta\left(\frac{p \log p}{\gamma p \log \frac{1}{\gamma}}\right)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = \Theta(\frac{1}{k \log k})$	$\Theta(k \log(p-k))$	$\Theta(p \log p)$
$\lambda^2 = \Theta(\frac{\log k}{k})$ $\gamma = \Omega(\frac{1}{k})$	$\max\left\{\Theta\left(\frac{k \log \frac{p}{k}}{\log \log k}\right), \Theta\left(\frac{k \log(p-k)}{\log k}\right)\right\}$	$\Theta(p)$

dramatically in this regime. Finally, if $\gamma k = \Theta(1)$, then the recovery threshold (14b) transitions between the two extremes. Using the bounds in Corollary 2, the necessary conditions in Theorem 2 are shown in Table II under different scalings of the parameters $(n, p, k, \lambda, \gamma)$. In particular, if $\gamma = o\left(\frac{1}{k \log k}\right)$ and the minimum value λ^2 does not increase with k, then the denominator $\gamma k \log \frac{1}{\gamma}$ goes to zero.

III. PROOFS OF OUR MAIN RESULTS

In this section, we provide the proofs of Theorems 1 and 2. Establishing necessary conditions for exact sparsity recovery amounts to finding conditions on (n, p, k, λ) (and possibly γ) under which the probability of error of any recovery method stays bounded away from zero as $n \to \infty$. At a high-level, our general approach is quite simple: we consider restricted problems in which the decoder has been given some additional side information, and then apply Fano's method [10], [16], [17], [35], [34] to lower bound the probability of error. In order to establish the collection of necessary conditions (e.g., $\{f_1(p, k, \lambda), \ldots, f_k(p, k, \lambda)\}$), we construct a family of restricted ensembles which sweeps the range of possible overlaps between support sets. At the extremes of this family are two classes of ensembles: one which captures the bulk effect of having many competing subsets at large distances, and the other which captures the effect of a smaller number of subsets at very close distances [this is illustrated in Fig. 2(a)]. Accordingly, we consider the family of ensembles $\left\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\right\}_{m=1,\ldots,k}$, where the mth restricted ensemble is defined as follows.

Throughout the remainder of the paper, we use the notation $X_j \in \mathbb{R}^n$ to denote column j of the matrix X, and $X_U \in \mathbb{R}^{n \times |U|}$ to denote the submatrix containing columns indexed by set U. Similarly, let $\beta_U \in \mathbb{R}^{|U|}$ denote the subvector of β corresponding to the index set U. In addition, let $H(\cdot)$ and $h(\cdot)$ denote the entropy and differential entropy functionals, respectively.

A. Restricted Ensemble $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$

Suppose that the decoder is given the locations of all but the m smallest nonzero values of the vector β, as well as the values of β on its support. More precisely, let S represent the true underlying support of β and let T denote the set of revealed indices, which has size $|T| = k - m$. Let $U = S \setminus T$ denote the set of unknown locations, and assume that $\beta_j = \lambda$ for all $j \in U$. Given knowledge of (T, β_T, λ), the decoder may simply subtract $X_T \beta_T = \sum_{j \in T} X_j \beta_j$ from Y, so that it is left with the modified n-vector of observations

$$\widetilde{Y} := \sum_{j \in U} X_j \lambda + W. \tag{15}$$

By re-ordering indices as need be, we may assume without loss of generality that $T = \{p - k + m + 1, \ldots, p\}$, so that $U \subset \{1, \ldots, p-k+m\}$. The remaining sub-problem is to determine, given the observations $\widetilde{Y}$, the locations of the m nonzeros in U.[4]

We will now argue that analyzing the probability of error of this restricted problem gives us a lower bound on the probability of error in the original problem. Consider the restricted signal class $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ defined as

$$\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda) := \left\{ \widetilde{\beta} \in \mathbb{R}^{p-k+m} \,\middle|\, |U(\widetilde{\beta})| = m \right. \\ \left. \widetilde{\beta}_j = \lambda \quad \forall j \in U(\widetilde{\beta}) \right\} \tag{16}$$

where we denote the support set of vector $\widetilde{\beta}$ as $U(\widetilde{\beta}) := \left\{ j \mid \widetilde{\beta}_j \neq 0 \right\}$. For any $\widetilde{\beta} \in \widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$, we can concatenate $\widetilde{\beta}$ with a vector v of $k-m$ nonzeros (with $\min_j |v_j| \geq \lambda$) at the end to obtain a p-dimensional vector. If a decoder can recover the support of any p-dimensional k-sparse vector $\beta \in \mathcal{C}_{p,k}(\lambda)$, then it can recover the support of the augmented $\widetilde{\beta}$ and, hence, the support of $\widetilde{\beta}$. Furthermore, providing the decoder with the nonzero values of β cannot increase the probability of error. Thus, we can apply Fano's inequality to lower bound the

[4]Note that if we assume the support of β is uniformly chosen over all $\binom{p}{k}$ possible subsets of size k, then given T, the remaining subset U is uniformly distributed over the $\binom{p-k+m}{m}$ possible subsets of size m.

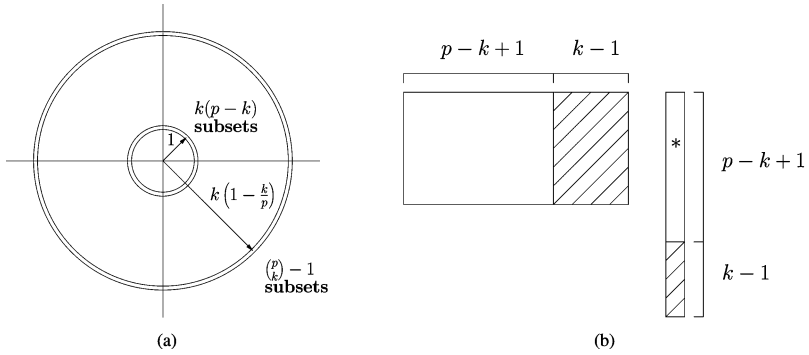

Fig. 2. Illustration of restricted ensembles. (a) In restricted ensemble $\widetilde{\mathcal{C}}_{p,k}(\lambda)$, the decoder must distinguish between $\binom{p}{k}$ support sets with an average overlap of size $\frac{k^2}{p}$, whereas in restricted ensemble $\widetilde{\mathcal{C}}_{p-k+1,1}(\lambda)$, it must decode amongst a subset of the $k(p-k)+1$ supports with overlap $k-1$. (b) In restricted ensemble $\widetilde{\mathcal{C}}_{p-k+1,1}(\lambda)$, the decoder is given the locations of the $k-1$ largest nonzeros, and it must estimate the location of the smallest nonzero from the $p-k+1$ remaining possible indices.

probability of error in the restricted problem, and so obtain a lower bound on the probability of error for the general problem.

B. Applying Fano to Restricted Ensembles

Consider the class of signals $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ defined in (16), which consists of $M = \binom{p-k+m}{m}$ models $\left\{\widetilde{\beta}^{(1)}, \ldots, \widetilde{\beta}^{(M)}\right\}$ corresponding to the M possible subsets $U \subset \{1, \ldots, p-k+m\}$ of size k. Suppose that a model index θ is chosen uniformly at random from $\{1, \ldots, M\}$, and we sample n observations $\widetilde{Y} \in \mathbb{R}^n$ via the measurement matrix $\widetilde{X} \in \mathbb{R}^{n \times (p-k+m)}$. For any decoding function $f : \mathbb{R}^n \to \{1, \ldots, M\}$, the average probability of error is defined as

$$p_{err}(f) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}\left[f(\widetilde{Y}) \neq i \mid \theta = i\right]$$

while the maximal probability of error over the class $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ is defined as

$$\omega(f) = \max_{i=1,\ldots,M} \mathbb{P}\left[f(\widetilde{Y}) \neq i \mid \theta = i\right].$$

We first apply Fano's lemma [10] to bound the error probability over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ for a particular instance of the random measurement matrix $\widetilde{X}$, and subsequently average over the ensemble of matrices. Thus, by Fano's inequality, the average probability of error and hence the maximal probability of error, is lower bounded as

$$p_{err}(f) \geq \frac{H\left(\theta \middle| \widetilde{Y}, \widetilde{X}\right) - 1}{\log M} = 1 - \frac{I\left(\theta; \widetilde{Y} \middle| \widetilde{X}\right) + 1}{\log M}. \quad (17)$$

Consequently, the problem of establishing necessary conditions for asymptotically reliable recovery is reduced to obtaining upper bounds on the conditional mutual information $I\left(\theta; \widetilde{Y} \middle| \widetilde{X}\right)$.

C. Proof of Theorem 1

In this section, we derive the necessary conditions stated in Theorem 1 for the general class of measurement matrices, by applying Fano's inequality to bound the probability of decoding error in each of the k restricted ensembles in the family $\left\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\right\}_{m=1,\ldots,k}$.

We begin by performing our analysis of the error probability over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$ for any $m \in \{1, \ldots, k\}$. Let $\widetilde{X} \in \mathbb{R}^{n \times (p-k+m)}$ be a matrix with independent, zero-mean and unit-variance entries. Conditioned on the event that U is the true underlying support of $\widetilde{\beta}$, the vector of n observations can be written as

$$\widetilde{Y} := \widetilde{X}_U \widetilde{\beta}_U + W = \lambda \sum_{j \in U} \widetilde{X}_j + W.$$

Accordingly, the conditional mutual information in (17) can be expanded as

$$\begin{aligned} I\left(\theta; \widetilde{Y} \middle| \widetilde{X}\right) &= h\left(\widetilde{Y} \middle| \widetilde{X}\right) - h\left(\widetilde{Y} \middle| \theta, \widetilde{X}\right) \\ &= h\left(\widetilde{Y} \middle| \widetilde{X}\right) - h(W). \end{aligned}$$

We bound the first term using the fact that the differential entropy of the observation vector $\widetilde{Y}$ for a particular instance of matrix $\widetilde{X}$ is maximized by the Gaussian distribution with a matched variance. More specifically, for a fixed $\widetilde{X}$, the distribution of $\widetilde{Y}$ is a Gaussian mixture with density $\psi\left(y|\widetilde{X}\right) = \frac{1}{\binom{p-k+m}{m}} \sum_U \phi\left(\widetilde{X}_U \widetilde{\beta}_U, I\right)$, where we are using ϕ to denote the density of a Gaussian random vector with mean $\widetilde{X}_U \widetilde{\beta}_U$ and covariance I. Let $\Lambda\left(\widetilde{X}\right)$ denote the covariance matrix of $\widetilde{Y}$ conditioned on $\widetilde{X}$ (hence, entry $\Lambda_{ii}\left(\widetilde{X}\right)$ on the diagonal represents the variance of $\widetilde{Y}_i$ given $\widetilde{X}$). With this notation, the entropy associated with the marginal density $\psi(y_i|\widetilde{X})$ is upper bounded by $\frac{1}{2}\log\left(2\pi \mathrm{e}\, \Lambda_{ii}\left(\widetilde{X}\right)\right)$. When $\widetilde{X}$ is randomly chosen, the conditional entropy of $\widetilde{Y}$ given $\widetilde{X}$ (averaged over the choice of $\widetilde{X}$) can be bounded as

$$\begin{aligned} h\left(\widetilde{Y} \middle| \widetilde{X}\right) &\leq \sum_{i=1}^{n} h\left(\widetilde{Y}_i \middle| \widetilde{X}\right) \\ &\leq \sum_{i=1}^{n} \mathbb{E}_{\widetilde{X}}\left[\frac{1}{2}\log\left(2\pi \mathrm{e} \Lambda_{ii}\left(\widetilde{X}\right)\right)\right]. \end{aligned}$$

The conditional entropy can be further bounded by exploiting the concavity of the logarithm and applying Jensen's inequality, as

$$h\left(\widetilde{Y}\middle|\widetilde{X}\right) \leq \sum_{i=1}^{n} \frac{1}{2}\log\left(2\pi \mathrm{e}\, \mathbb{E}_{\widetilde{X}}\left[\Lambda_{ii}\left(\widetilde{X}\right)\right]\right).$$

Next, the entropy of the Gaussian noise vector $W \sim N(0, I_{n\times n})$ can be computed as $h(W) = \frac{n}{2}\log(2\pi \mathrm{e})$. Combining these two terms, we then obtain the following bound on the conditional mutual information:

$$I\left(\theta;\widetilde{Y}\middle|\widetilde{X}\right) \leq \sum_{i=1}^{n} \frac{1}{2}\log\left(\mathbb{E}_{\widetilde{X}}\left[\Lambda_{ii}\left(\widetilde{X}\right)\right]\right).$$

It remains to compute the expectation $\mathbb{E}_{\widetilde{X}}\left[\Lambda_{ii}\left(\widetilde{X}\right)\right]$, over the ensemble of matrices $\widetilde{X}$ drawn with i.i.d. entries from any distribution with zero mean and unit variance. The proof of the following lemma involves some relatively straightforward but lengthy calculation, and is given in Appendix A.

Lemma 1: Given i.i.d. $\widetilde{X}_{ij}$ with zero mean and unit variance, the averaged covariance matrix of $\widetilde{Y}$ given $\widetilde{X}$ is

$$\mathbb{E}_{\widetilde{X}}\left[\Lambda\left(\widetilde{X}\right)\right] = \left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right) I_{n\times n}. \quad (18)$$

Finally, combining Lemma 1 with equation (17), we obtain that the average probability of error is bounded away from zero if

$$n < \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log\left(1 + m\lambda^2\left(1 - \frac{m}{p-k+m}\right)\right)}$$

as claimed.

D. Proof of Theorem 2

This section contains proofs of the necessary conditions in Theorem 2 for the γ-sparsified Gaussian measurement ensemble (7). We proceed as before, applying Fano's inequality to each restricted class in the family $\left\{\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)\right\}_{m=1,\ldots,k}$, in order to derive the corresponding k conditions in Theorem 2.

In analyzing the probability of error over $\widetilde{\mathcal{C}}_{p-k+m,m}(\lambda)$, the initial steps proceed as in the proof of Theorem 1, by expanding the conditional mutual information in equation (17) as

$$\begin{aligned} I\left(\theta;\widetilde{Y}\middle|\widetilde{X}\right) &= h\left(\widetilde{Y}\middle|\widetilde{X}\right) - h(W) \\ &\leq \sum_{i=1}^{n} h\left(\widetilde{Y}_i\middle|\widetilde{X}\right) - \frac{n}{2}\log(2\pi \mathrm{e}) \end{aligned}$$

using the Gaussian entropy for $W \sim N(0, I_{n\times n})$.

From this point, the key subproblem is to compute the conditional entropy of $\widetilde{Y}_i = \lambda \sum_{j\in U(\widetilde{\beta})} \widetilde{X}_{ij} + W_i$, when the support of $\widetilde{\beta}$ is uniformly chosen over all $\binom{p-k+m}{m}$ possible subsets of size m. To characterize the limiting behavior of the random variable $\widetilde{Y}_i$, note that for a fixed matrix $\widetilde{X}$, each $\widetilde{Y}_i$ is distributed according to the density defined as

$$\psi_m\left(y_i\middle|\widetilde{X}\right) = \frac{1}{\binom{p-k+m}{m}} \sum_{U} \frac{1}{\sqrt{2\pi}} \times \exp\left(-\frac{1}{2}\left(y_i - \lambda \sum_{j\in U} \widetilde{X}_{ij}\right)^2\right).$$

This density is a mixture of Gaussians with unit variances and means that depend on the values of $\{\widetilde{X}_{i1}, \ldots, \widetilde{X}_{i(p-k+m)}\}$, summed over subsets $U \subset \{1, \ldots, p-k+m\}$ with $|U| = m$. At a high-level, our immediate goal is to characterize the entropy $h(\psi_m)$.

Note that as $\widetilde{X}$ varies over the sparse ensemble (7), the sequence $\left\{\psi_m\left(y_i\middle|\widetilde{X}\right)\right\}_p$, indexed by the signal dimension p, is actually a sequence of random densities. As an intermediate step, the following lemma characterizes the average pointwise behavior of this random sequence of densities, and is proven in Appendix B.

Lemma 2: Let $\widetilde{X}$ be drawn with i.i.d. entries from the γ-sparsified Gaussian ensemble (7). For any fixed y_i and m, $\mathbb{E}_{\widetilde{X}}\left[\psi_m\left(y_i\middle|\widetilde{X}\right)\right] = \overline{\psi}_m(y_i)$, where

$$\overline{\psi}_m(y_i) = \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\left(1 + \frac{L\lambda^2}{\gamma}\right)}} \exp\left(-\frac{y_i^2}{2\left(1 + \frac{L\lambda^2}{\gamma}\right)}\right)\right] \quad (19)$$

is a mixture of Gaussians with binomial weights $L \sim \mathrm{Binomial}(m, \gamma)$.

For certain scalings, we can use concentration results for U-statistics [26] to prove that ψ_m converges uniformly to $\overline{\psi}_m$, and from there that $h(\psi_m) \xrightarrow{p} h(\overline{\psi}_m)$. In general, however, we always have an upper bound, which is sufficient for our purposes. Indeed, since differential entropy $h(\psi_m)$ is a concave function of ψ_m, by Jensen's inequality and Lemma 2, we have

$$\mathbb{E}_{\widetilde{X}}[h(\psi_m)] \leq h\left(\mathbb{E}_{\widetilde{X}}[\psi_m]\right) = h(\overline{\psi}_m).$$

With these ingredients, we conclude that the conditional mutual information in (17) is upper bounded by

$$\begin{aligned} I\left(\theta;\widetilde{Y}\middle|\widetilde{X}\right) &\leq \sum_{i=1}^{n} h\left(\widetilde{Y}_i\middle|\widetilde{X}\right) - \frac{n}{2}\log(2\pi \mathrm{e}) \\ &= \sum_{i=1}^{n} \mathbb{E}_{\widetilde{X}}[h(\psi_m)] - \frac{n}{2}\log(2\pi \mathrm{e}) \\ &\leq n h(\overline{\psi}_m) - \frac{n}{2}\log(2\pi \mathrm{e}) \end{aligned}$$

where the last inequality uses the fact that the entropies $h(\overline{\psi}_m)$ associated with the densities $\overline{\psi}_m(y_i)$ are the same for all i. Therefore, the probability of decoding error, averaged over the

sparsified Gaussian measurement ensemble, is bounded away from zero if

$$n < \frac{\log \binom{p-k+m}{m} - 1}{H(\overline{\psi}_m) - \frac{1}{2}\log(2\pi \mathrm{e})}$$

as claimed.

E. Proof of Corollary 2

In this section, we derive bounds on the necessary conditions $g_m(p,k,\lambda,\gamma)$ for $m = 1,\ldots,k$, which are stated in Theorem 2. We begin by applying a simple yet general bound on the entropy of the Gaussian mixture distribution with density $\overline{\psi}_m$ defined in (11). The variance associated with the density $\overline{\psi}_m$ is equal to $\sigma_m^2 = 1 + m\lambda^2$, and so $h(\overline{\psi}_m)$ is bounded by the entropy of a Gaussian distribution with variance σ_m^2, as

$$h(\overline{\psi}_m) \leq \frac{1}{2}\log(2\pi \mathrm{e}(1+m\lambda^2)).$$

This yields the first set of bounds in (14a).

Next, to derive more refined bounds which capture the effects of measurement sparsity, we will make use of the following lemma (which is proven in Appendix C) to bound the entropy associated with the mixture density $\overline{\psi}_m$.

Lemma 3: For the Gaussian mixture distribution with density $\overline{\psi}_m$ defined in (11)

$$h(\overline{\psi}_m) \leq \mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right] + H(L) + \frac{1}{2}\log(2\pi \mathrm{e})$$

where $L \sim \mathrm{Binomial}(m,\gamma)$.

We can further bound the expression in Lemma 3 in three cases, delineated by the quantity γm. The proof of the following claim in given in Appendix D.

Lemma 4: Let $E := \mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right]$, where $L \sim \mathrm{Binomial}(m,\gamma)$.

(a) If $\gamma m > 3$, then

$$E \leq \frac{1}{2}\log\left(1+m\lambda^2\right) \text{ and} \tag{20a}$$

$$E \geq \frac{1}{4}\log\left(1+\frac{m\lambda^2}{3}\right). \tag{20b}$$

(b) If $\gamma m = \tau$ for some constant τ, then

$$E \leq \frac{1}{2}\tau\log\left(1+\frac{m\lambda^2}{\tau}\right) \text{ and} \tag{21a}$$

$$E \geq \frac{1}{2}(1-\mathrm{e}^{-\tau})\log\left(1+\frac{m\lambda^2}{\tau}\right). \tag{21b}$$

(c) If $\gamma m \leq 1$, then

$$E \leq \frac{1}{2}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right) \text{ and} \tag{22a}$$

$$E \geq \frac{1}{4}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right). \tag{22b}$$

Finally, combining Lemmas 3 and 4 with some simple bounds on the entropy of the binomial variate L (summarized in Lemmas 5 and 6 in Appendix E), we obtain the bounds on $g_m(p,k,\lambda,\gamma)$ in (14b) and (14c).

IV. Discussion

In this paper, we have studied the information-theoretic limits of exact support recovery for general scalings of the parameters (n,p,k,λ,γ). Our first result (Theorem 1) applies generally to measurement matrices with zero-mean and unit-variance entries. It strengthens previously known bounds, and combined with known sufficient conditions [29], yields a sharp characterization of recovering signals with linear sparsity with a linear fraction of observations (Corollary 1). Our second result (Theorem 2) applies to γ-sparsified Gaussian measurement ensembles, and reveals three different regimes of measurement sparsity, depending on how significantly they impair statistical efficiency. For linear signal sparsity, Theorem 2 is not a sharp result (by a constant factor in comparison to Theorem 1 in the dense case); however, its tightness for sublinear signal sparsity is an interesting open problem. Finally, Theorem 1 implies that no measurement ensemble with zero-mean and unit-variance entries can further reduce the number of observations necessary for recovery, while the paper [29] shows that the standard Gaussian ensemble can achieve the same scaling. This raises an interesting open question on the design of other, more computationally friendly, measurement matrices which achieve the same information-theoretic bounds.

Appendix

A) Proof of Lemma 1: We begin by defining some additional notation. Recall that for a given instance of the matrix $\widetilde{X}$, the observation vector $\widetilde{Y}$ has a Gaussian mixture distribution with density $\psi\left(y|\,\widetilde{X}\right) = \frac{1}{\binom{p-k+m}{m}}\sum_U \phi\left(\widetilde{X}_U\widetilde{\beta}_U, I\right)$, where ϕ denotes the Gaussian density with mean $\widetilde{X}_U\widetilde{\beta}_U$ and covariance I. Let $\mu\left(\widetilde{X}\right) = \mathbb{E}\left[\widetilde{Y}\middle|\,\widetilde{X}\right] \in \mathbb{R}^n$ and $\Lambda\left(\widetilde{X}\right) = \mathbb{E}\left[\widetilde{Y}\widetilde{Y}^T\middle|\,\widetilde{X}\right] - \mu\left(\widetilde{X}\right)\mu\left(\widetilde{X}\right)^T \in \mathbb{R}^{n\times n}$ be the mean vector and covariance matrix of $\widetilde{Y}$ given $\widetilde{X}$, respectively. Accordingly, we have

$$\mu\left(\widetilde{X}\right) = \frac{1}{\binom{p-k+m}{m}}\sum_U \widetilde{X}_U\widetilde{\beta}_U$$

and

$$\mathbb{E}\left[\widetilde{Y}\widetilde{Y}^T\middle|\,\widetilde{X}\right] = \frac{1}{\binom{p-k+m}{m}}\sum_U \left(\widetilde{X}_U\widetilde{\beta}_U\right)\left(\widetilde{X}_U\widetilde{\beta}_U\right)^T + I.$$

With this notation, we can now compute the expectation of the covariance matrix $\mathbb{E}_{\widetilde{X}}\left[\Lambda\left(\widetilde{X}\right)\right]$, averaged over any distribution

on $\widetilde{X}$ with independent, zero-mean and unit-variance entries. To compute the first term, we have

$$\begin{aligned}\mathbb{E}_{\widetilde{X}}\left[\mathbb{E}\left[\widetilde{Y}\widetilde{Y}^T\middle|\widetilde{X}\right]\right] &= \frac{\lambda^2}{\binom{p-k+m}{m}}\sum_U \mathbb{E}_{\widetilde{X}}\left[\sum_{j\in U}\widetilde{X}_j\widetilde{X}_j^T + \sum_{i\neq j\in U}\widetilde{X}_i\widetilde{X}_j^T\right] + I \\ &= \frac{\lambda^2}{\binom{p-k+m}{m}}\sum_U\sum_{j\in U} I + I \\ &= \left(1+m\lambda^2\right) I\end{aligned}$$

where the second equality uses the fact that $\mathbb{E}_{\widetilde{X}}\left[\widetilde{X}_j\widetilde{X}_j^T\right] = I$, and $\mathbb{E}_{\widetilde{X}}\left[\widetilde{X}_i\widetilde{X}_j^T\right] = 0$ for $i \neq j$. Next, we compute the second term as

$$\begin{aligned}\mathbb{E}_{\widetilde{X}}\left[\mu\left(\widetilde{X}\right)\mu\left(\widetilde{X}\right)^T\right] &= \left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \\ &\quad\times \mathbb{E}_{\widetilde{X}}\left[\sum_{U,V}\sum_{j\in U\cap V}\widetilde{X}_j\widetilde{X}_j^T + \sum_{U,V}\sum_{\substack{i\in U, j\in V\\ i\neq j}}\widetilde{X}_i\widetilde{X}_j^T\right] \\ &= \left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2\sum_{U,V}\sum_{j\in U\cap V} I \\ &= \left(\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2\sum_{U,V}|U\cap V|\right) I.\end{aligned}$$

From here, note that there are $\binom{p-k+m}{m}$ possible subsets U of size m. For each U, a counting argument reveals that there are $\binom{m}{\delta}\binom{p-k}{m-\delta}$ subsets V of size m which have $|U\cap V| = \delta$ overlaps with U. Thus, the scalar multiplicative factor above can be written as

$$\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2\sum_{U,V}|U\cap V| = \frac{\lambda^2}{\binom{p-k+m}{m}}\sum_{\delta=1}^{m}\binom{m}{\delta}\binom{p-k}{m-\delta}\delta.$$

Finally, using a substitution of variables (by setting $\delta' = \delta - 1$) and applying Vandermonde's identity [23], we have

$$\begin{aligned}&\left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2\sum_{U,V}|U\cap V| \\ &\quad= \frac{\lambda^2}{\binom{p-k+m}{m}}m\sum_{\delta'=0}^{m-1}\binom{m-1}{\delta'}\binom{p-k}{m-\delta'-1} \\ &\quad= \frac{\lambda^2}{\binom{p-k+m}{m}}m\binom{p-k+m-1}{m-1} \\ &\quad= \frac{m^2\lambda^2}{p-k+m}.\end{aligned}$$

Combining these terms, we conclude that

$$\mathbb{E}_{\widetilde{X}}\left[\Lambda\left(\widetilde{X}\right)\right] = \left(1+m\lambda^2\left(1-\frac{m}{p-k+m}\right)\right) I.$$

B) Proof of Lemma 2: Consider the following sequences of densities:

$$\begin{aligned}\psi_m\left(y_i|\,\widetilde{X}\right) &= \frac{1}{\binom{p-k+m}{m}} \\ &\quad\times\sum_U\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y_i-\lambda\sum_{j\in U}\widetilde{X}_{ij}\right)^2\right)\end{aligned}$$

and

$$\begin{aligned}&\overline{\psi}_m(y_i) \\ &\quad= \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\left(1+\frac{L\lambda^2}{\gamma}\right)}}\exp\left(-\frac{y_i^2}{2\left(1+\frac{L\lambda^2}{\gamma}\right)}\right)\right]\end{aligned}$$

where $L \sim \text{Binomial}(m,\gamma)$. Our goal is to show that for any fixed y_i, the pointwise average of the stochastic sequence of densities ψ_m over the ensemble of matrices $\widetilde{X}$ satisfies $\mathbb{E}_{\widetilde{X}}\left[\psi_m\left(y_i|\,\widetilde{X}\right)\right] = \overline{\psi}_m(y_i)$.

By symmetry of the random measurement matrix $\widetilde{X}$, it is sufficient to compute this expectation for the subset $U = \{1,\ldots,m\}$. When each $\widetilde{X}_{ij}$ is i.i.d. drawn according to the γ-sparsified ensemble (7), the random variable $Z := \left(y_i - \lambda\sum_{j=1}^m\widetilde{X}_{ij}\right)$ has a Gaussian mixture distribution which can be described as follows. Denoting the mixture label by L, then $Z \sim N\left(y_i, \frac{\ell\lambda^2}{\gamma}\right)$ if $L = \ell$, for $\ell = 0,\ldots,m$. Moreover, define the modified random variable $\tilde{Z} := \frac{\gamma}{L\lambda^2}\left(y_i - \lambda\sum_{j=1}^m\widetilde{X}_{ij}\right)^2$. Then, conditioned on the mixture label $L = \ell$, the random variable $\tilde{Z}$ has a noncentral chi-square distribution with 1 degree of freedom and parameter $\frac{\gamma y_i^2}{\ell\lambda^2}$. Letting $M_\ell(t) = \mathbb{E}\left[\exp(t\tilde{Z})\,\middle|\,L=\ell\right]$ denote the ℓth moment-generating function of $\tilde{Z}$, we have

$$\begin{aligned}&\mathbb{E}_{\widetilde{X}}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y_i-\lambda\sum_{j=1}^m\widetilde{X}_{ij}\right)^2\right)\right] \\ &\qquad= \sum_{\ell=0}^m\frac{1}{\sqrt{2\pi}}M_\ell\left(-\frac{\ell\lambda^2}{2\gamma}\right)\mathbb{P}(L=\ell).\end{aligned}$$

Evaluating the moment generating function [5] of a noncentral chi-square random variable then gives the desired quantity

$$\begin{aligned}&\mathbb{E}_{\widetilde{X}}\left[\psi_m\left(y_i|\,\widetilde{X}\right)\right] \\ &\qquad= \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi\left(1+\frac{L\lambda^2}{\gamma}\right)}}\exp\left(-\frac{y_i^2}{2\left(1+\frac{L\lambda^2}{\gamma}\right)}\right)\right]\end{aligned}$$

as claimed.

C) Proof of Lemma 3: Let Z be a random variable distributed according to the density (19) with mixture label $L \sim \text{Binomial}(m,\gamma)$. To compute the entropy of Z, we expand the mutual information $I(Z;L)$ and obtain

$$h(Z) = h(Z|L) + H(L) - H(L|Z).$$

The conditional distribution of Z given that $L = \ell$ is Gaussian, and so the conditional entropy of Z given L can be written as

$$h(Z|L) = \mathbb{E}_L\left[\frac{1}{2}\log\left(2\pi\mathrm{e}\left(1+\frac{L\lambda^2}{\gamma}\right)\right)\right].$$

Using the fact that $0 \le H(L|Z) \le H(L)$, we obtain the upper and lower bounds on $h(Z)$

$$h(Z|L) \le h(Z) \le h(Z|L) + H(L)$$

as claimed.

D) Proof of Lemma 4: Let $E := \mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right]$, where $L \sim \mathrm{Binomial}(m,\gamma)$. We first derive a general upper bound on E and then show that this bound is reasonably tight in the case when $\gamma m \le 1$. We can rewrite the binomial probability as

$$\begin{aligned} p(\ell) &:= \binom{m}{\ell}\gamma^\ell(1-\gamma)^{m-\ell} \\ &= \frac{\gamma m}{\ell}\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell} \end{aligned}$$

and hence

$$E = \frac{1}{2}\gamma m\sum_{\ell=1}^{m}\frac{\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)}{\ell}\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell}.$$

Taking the first two terms of the binomial expansion of $\left(1+\frac{\lambda^2}{\gamma}\right)^\ell$ and noting that all the terms are non-negative, we obtain the inequality

$$\left(1+\frac{\lambda^2}{\gamma}\right)^\ell \ge 1+\frac{\ell\lambda^2}{\gamma}$$

and consequently $\log\left(1+\frac{\lambda^2}{\gamma}\right) \ge \frac{1}{\ell}\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)$. Using a change of variables (by setting $\ell' = \ell - 1$) and applying the binomial theorem, we obtain the upper bound

$$\begin{aligned} E &\le \frac{1}{2}\gamma m\sum_{\ell=1}^{m}\log\left(1+\frac{\lambda^2}{\gamma}\right)\binom{m-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{m-\ell} \\ &= \frac{1}{2}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right)\sum_{\ell'=0}^{m-1}\binom{m-1}{\ell'}\gamma^{\ell'}(1-\gamma)^{m-\ell'-1} \\ &= \frac{1}{2}\gamma m\log\left(1+\frac{\lambda^2}{\gamma}\right). \end{aligned}$$

In the case when $\gamma m \le 1$, we can derive a similar lower bound by first bounding E as

$$\begin{aligned} E &\ge \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)\sum_{\ell=1}^{m}p(\ell) \\ &= \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)(1-(1-\gamma)^m). \end{aligned}$$

Now using the fact that $1+x \le \mathrm{e}^x$ for all $x \in \mathbb{R}$, and $\mathrm{e}^{-x} \le 1-\frac{x}{2}$ for $x \in [0,1]$, we have

$$\begin{aligned} E &\ge \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)(1-\mathrm{e}^{-\gamma m}) \\ &\overset{(a)}{\ge} \frac{1}{2}\log\left(1+\frac{\lambda^2}{\gamma}\right)\left(\frac{\gamma m}{2}\right). \end{aligned}$$

This yields the upper and lower bounds in (22).

Next, we examine the case when $\gamma m = \tau$ for some constant τ. The derivation of the upper bound for the $\gamma m \le 1$ case holds when $\gamma m = \tau$ as well. The proof of the lower bound follows the same steps as in the $\gamma m \le 1$ case, except that we stop before applying the last inequality (a). This gives the bounds in (21).

Finally, we derive bounds in the case when $\gamma m > 3$. Since the mean of a $L \sim \mathrm{Binomial}(m,\gamma)$ random variable is γm, by Jensen's inequality the following upper bound always holds:

$$\mathbb{E}_L\left[\frac{1}{2}\log\left(1+\frac{L\lambda^2}{\gamma}\right)\right] \le \frac{1}{2}\log(1+m\lambda^2).$$

To derive a matching lower bound, we use the fact that the median of a $\mathrm{Binomial}(m,\gamma)$ distribution is one of $\{\lfloor\gamma m\rfloor - 1, \lfloor\gamma m\rfloor, \lfloor\gamma m\rfloor + 1\}$. This allows us to bound

$$\begin{aligned} E &\ge \frac{1}{2}\sum_{\ell=\lfloor\gamma m\rfloor-1}^{m}\log\left(1+\frac{\ell\lambda^2}{\gamma}\right)p(\ell) \\ &\ge \frac{1}{2}\log\left(1+\frac{(\lfloor\gamma m\rfloor-1)\lambda^2}{\gamma}\right)\sum_{\ell=\lfloor\gamma m\rfloor-1}^{m}p(\ell) \\ &\ge \frac{1}{4}\log\left(1+\frac{m\lambda^2}{3}\right) \end{aligned}$$

where in the last step we used the fact that $\frac{(\lfloor\gamma m\rfloor-1)\lambda^2}{\gamma} \ge \frac{(\gamma m-2)\lambda^2}{\gamma} \ge \frac{m\lambda^2}{3}$ for $\gamma m > 3$, and $\sum_{\ell=\mathrm{median}}^{m}p(\ell) \ge \frac{1}{2}$. Thus, we obtain the bounds in (20).

E) Bounds on Binomial Entropy:

Lemma 5: Let $L \sim \mathrm{Binomial}(m,\gamma)$, then

$$H(L) \le \frac{1}{2}\log\left(2\pi\mathrm{e}\left(m\gamma(1-\gamma)+\frac{1}{12}\right)\right).$$

Proof: We immediately obtain this bound by applying the differential entropy bound on discrete entropy [10]. As detailed in [10], the proof follows by relating the entropy of the discrete random variable L to the differential entropy of a particular continuous random variable, and then upper bounding the latter by the entropy of a Gaussian random variable. ■

Lemma 6: The entropy of a binomial random variable $L \sim \mathrm{Binomial}(m,\gamma)$ is bounded by

$$H(L) \le mH_{\mathrm{binary}}(\gamma).$$

Proof: We can express the binomial variate as $L = \sum_{i=1}^{m} Z_i$, where $Z_i \sim \text{Bernoulli}(\gamma)$ i.i.d. Since $H(g(Z_1, \ldots, Z_m)) \leq H(Z_1, \ldots, Z_m)$, we have

$$H(L) \leq H(Z_1, \ldots, Z_m) = mH_{\text{binary}}(\gamma).$$

■

Lemma 7: If $\gamma = o\left(\frac{1}{m \log m}\right)$, then $mH_{\text{binary}}(\gamma) \to 0$ as $m \to \infty$.

Proof: To find the limit of $mH_{\text{binary}}(\gamma) = m\gamma \log \frac{1}{\gamma} + m(1-\gamma)\log \frac{1}{1-\gamma}$, let $\gamma = \frac{1}{mf(m)}$ for some function f, and assume that $f(m) = \omega(\log m)$. We can expand the first term as

$$m\gamma \log \frac{1}{\gamma} = \frac{1}{f(m)} \log(mf(m)) = \frac{\log m}{f(m)} + \frac{\log f(m)}{f(m)}$$

and so $\lim_{m\to\infty} m\gamma \log \frac{1}{\gamma} = 0$. The second term can also be expanded as

$$\begin{aligned} &- m(1-\gamma)\log(1-\gamma) \\ &= -m \log\left(1 - \frac{1}{mf(m)}\right) + \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right) \\ &= -\log\left(1 - \frac{1}{mf(m)}\right)^m + \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right). \end{aligned}$$

Since $f(m) \to \infty$ as $m \to \infty$, we have the limits

$$\lim_{m\to\infty} \left(1 - \frac{1}{mf(m)}\right)^m = 1 \quad \text{and}$$
$$\lim_{m\to\infty} \left(1 - \frac{1}{mf(m)}\right) = 1$$

which in turn imply that

$$\lim_{m\to\infty} \log\left(1 - \frac{1}{mf(m)}\right)^m = 0 \quad \text{and}$$
$$\lim_{m\to\infty} \frac{1}{f(m)} \log\left(1 - \frac{1}{mf(m)}\right) = 0.$$

■

F) Generalized Measurement Ensembles: In this section, we extend the necessary conditions in Theorem 1 to a generalized class of measurement matrices by relaxing the i.i.d. assumption. The proof of Theorem 3 below exactly mirrors that of Theorem 1, and is omitted. The key modification occurs when constructing the restricted ensembles, because the choice of columns to be removed from the matrix X affects the distribution of the observations. The proof of Lemma 1 can then be easily extended to the generalized measurement ensemble. In order to state the result, we define the functions

$$\alpha_m(p,k,\lambda) := \min_{\substack{T\subseteq\{1,\ldots,p\} \\ |T|=p-k+m}} \left\{ 1 + \frac{\lambda^2}{\binom{p-k+m}{m}} \sum_{\substack{U\subseteq T \\ |U|=m}} \sum_{i,j\in U} \Sigma_{ij} - \left(\frac{\lambda}{\binom{p-k+m}{m}}\right)^2 \sum_{\substack{U,V\subseteq T \\ |U|=|V|=m}} \sum_{\substack{i\in U \\ j\in V}} \Sigma_{ij} \right\} \quad (23)$$

for $m \in \{1, \ldots, k\}$, which sums over all possible subsets of size m of the covariance matrix Σ.

Theorem 3: Let each row of the measurement matrix $X \in \mathbb{R}^{n\times p}$ be drawn i.i.d. from any distribution with zero mean and covariance matrix Σ. Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}_{p,k}(\lambda)$ is

$$n > \max\left\{\bar{f}_1(p,k,\lambda), \ldots, \bar{f}_k(p,k,\lambda), k\right\} \quad (24)$$

where

$$\bar{f}_m(p,k,\lambda) := \frac{\log\binom{p-k+m}{m} - 1}{\frac{1}{2}\log(\alpha_m(p,k,\lambda))} \quad (25)$$

for $m = 1, \ldots, k$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for helpful comments that improved the presentation of this paper.

REFERENCES

[1] S. Aeron, M. Zhao, and V. Saligrama, "Information-theoretic bounds to sensing capacity of sensor networks under fixed snr," presented at the Information Theory Workshop, Sep. 2007.
[2] S. Aeron, M. Zhao, and V. Saligrama, Fundamental Limits on Sensing Capacity for Sensor Networks and Compressed Sensing, 2008, Tech. Rep. arXiv:0804.3439v1 [cs.IT].
[3] M. Akcakaya and V. Tarokh, Shannon Theoretic Limits on Noisy Compressive Sampling 2007, Tech. Rep. arXiv:0711.0366v1 [cs.IT].
[4] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," presented at the Allerton Conf. Communication, Control and Computing, Monticello, IL, Sep. 2008.
[5] L. Birgé, "An alternative point of view on Lepski's method," in *State of the Art in Probability and Statistics*, ser. IMS Lecture Notes. Beachwood, OH: Institute of Mathematical Statistics, 2001, pp. 113–133.
[6] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
[7] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[8] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
[9] G. Cormode and S. Muthukrishnan, Towards an Algorithmic Theory of Compressed Sensing, Rutgers Univ., 2005, Tech. Rep.
[10] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[11] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[12] D. Donoho, M. Elad, and V. M. 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.

[13] A. K. Fletcher, S. Rangan, and V. K. Goyal, Necessary and Sufficient Conditions on Sparsity Pattern Recovery, 2008, Tech. Rep. arXiv:0804.1839v1 [cs.IT].
[14] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *J. Appl. Signal Process.*, vol. 10, pp. 1–19, 2006.
[15] A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin, "Algorithmic linear dimension reduction in the ℓ_1-norm for sparse vectors," presented at the Allerton Conf. Communication, Control and Computing, Monticello, IL, Sep. 2006.
[16] R. Z. Has'minskii, "A lower bound on the risks of nonparametric estimates of densities in the uniform metric," *Theory Prob. Appl.*, vol. 23, pp. 794–798, 1978.
[17] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1981.
[18] N. Meinshausen and P. Buhlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
[19] A. J. Miller, *Subset Selection in Regression*. New York: Chapman-Hall, 1990.
[20] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.
[21] D. Omidiran and M. J. Wainwright, High-Dimensional Subset Recovery in Noise: Sparsified Measurements Without Loss of Statistical Efficiency, Dept. Statistics, Univ. California, Berkeley, 2008, Tech. Rep. 753.
[22] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," presented at the Int. Symp. Information Theory, Toronto, Canada, 2008.
[23] J. Riordan, *Combinatorial Identities*. New York: Wiley, 1968, Wiley Series in Probability and Mathematical Statistics.
[24] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements versus bits: Compressed sensing meets information theory," presented at the Allerton Conf. Control, Communication and Computing, Sep. 2006.
[25] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Sudocodes: Fast measurement and reconstruction of sparse signals," presented at the Int. Symp. Information Theory, Seattle, WA, Jul. 2006.
[26] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, ser. Wiley Series in Probability and Statistics. New York: Wiley, 1980.
[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[28] J. 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.
[29] 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.
[30] 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.
[31] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," presented at the Int. Conf. Information Processing in Sensor Networks, Nashville, TN, Apr. 2007.
[32] W. Wang, M. J. Wainwright, and K. Ramchandran, Information-Theoretic Limits on Sparse Support Recovery: Dense Versus Sparse Measurements, Dept. Statistics, Univ. California, Berkeley, 2008, Tech. Rep. 754.
[33] W. Xu and B. Hassibi, "Efficient compressed sensing with deterministic guarantees using expander graphs," presented at the Information Theory Workshop (ITW), Sep. 2007.
[34] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.
[35] B. Yu, "Assouad, Fano and Le Cam," in *Festschrift for Lucien Le Cam*. Berlin, Germany: Springer-Verlag, 1997, pp. 423–435.

Wei Wang (M'09) received the B.S. degree (with honors) in electrical and computer engineering from Rice University, Houston, TX, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley.

Her research interests include statistical signal processing, high-dimensional statistics, coding and information theory, and large-scale distributed systems. She has been awarded an NSF Graduate Fellowship, a Bell Labs Graduate Research Fellowship, a GAANN Graduate Fellowship, a Hertz Foundation Grant, and the James S. Waters Award (Rice University).

Martin J. Wainwright (M'03) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge.

He is currently an Associate Professor at University of California, Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences. His research interests include statistical signal processing, coding and information theory, statistical machine learning, and high-dimensional statistics.

Prof. Wainwright has been awarded an Alfred P. Sloan Foundation Fellowship, an NSF CAREER Award, the George M. Sprowls Prize for his dissertation research (EECS department, MIT), a Natural Sciences and Engineering Research Council of Canada 1967 Fellowship, an IEEE Signal Processing Society Best Paper Award in 2008, and several outstanding conference paper awards.

Kannan Ramchandran (S'92–M'93–SM'98–F'05) received the Ph.D. degree in electrical engineering from Columbia University, New York, in 1993.

He is a Professor in the Electrical Engineering and Computer Science Department, University of California (UC), Berkeley. He has been at UC Berkeley since 1999. From 1993 to 1999, he was on the faculty of the Electrical and Computer Engineering Department, University of Illinois at Urbana-Champaign (UIUC), Urbana. Prior to that, he was a member of the Technical Staff at AT&T Bell Laboratories from 1984 to 1990. His current research interests include distributed signal processing and coding for networks, video communications and peer-to-peer content delivery, multi-user information theory, security, and multi-scale image processing and modeling.

Prof. Ramchandran has published extensively in his field, holds 12 patents, and has received several awards including an Outstanding Teaching award at Berkeley (2009), an Okawa Foundation Research Prize at Berkeley (2001), a Hank Magnusky Scholar Award at the Univesity of Illinois (1999), two Best Paper awards from the IEEE Signal Processing Society (1997 and 1993), an NSF CAREER Award (1997), an ONR Young Investigator Award (1997), an ARO Young Investigator Award (1996), and the Elaihu I. Jury thesis Award for his doctoral thesis at Columbia University (1993). He has additionally won numerous best conference paper awards in his field, and serves on the technical program committees for the premier conferences in information theory, communications, and signal and image processing.