Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices

Wei Wang* Martin J. Wainwright†,* Kannan Ramchandran*
{wangwei, wainwrig, kannanr}@eecs.berkeley.edu

Department of Electrical Engineering and Computer Sciences*, and
Department of Statistics†
University of California, Berkeley

Technical Report
Department of Statistics, UC Berkeley
May 2008

Abstract

We study the information-theoretic limits of exactly recovering the support 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 (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 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 $\gamma(n, p, k) \in (0, 1]$ corresponds to the fraction of non-zero 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 effect, a minor effect, or a dramatic effect on the information-theoretic limits of the subset recovery problem.

Keywords: Sparsity recovery; sparse random matrices; subset selection; compressive sensing; signal denoising; sparse approximation; information-theoretic bounds; Fano's inequality.

1 Introduction

The problem of estimating a k-sparse vector $\beta \in \mathbb{R}^p$ based on a set of n noisy linear observations is of broad interest, arising in subset selection in regression, graphical model selection, group testing, signal denoising, sparse approximation, and compressive sensing. A large body of recent work (e.g., [6, 9, 10, 5, 4, 14, 21, 22, 23, 13, 7, 25, 26, 19]) has analyzed the use of ℓ_1-relaxation methods for estimating high-dimensional sparse signals, and established conditions (on signal sparsity and the choice of measurement matrices) under which they succeed 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 of sparsity recovery. As we will see, the standard Gaussian measurement ensemble is an optimal choice in terms of minimizing the number of observations required for recovery. However, this choice produces highly dense measurement matrices, which may lead to prohibitively high computational complexity and storage requirements. Sparse matrices can reduce this complexity, and also lower communication cost and latency in distributed network and streaming applications. On the other hand, such measurement sparsity, though beneficial from the computational standpoint, may reduce statistical efficiency by requiring more observations to decode. Therefore, an important issue 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 [24], this analysis provides a sharp characterization of necessary and 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 non-zeros 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 1.1, and then discuss our contributions and some connections to related work in Section 1.2. Section 2 provides precise statements of our main results, as well as a discussion of their consequences. Section 3 provides proofs of the necessary conditions for various classes of measurement matrices, while proofs of more technical lemmas are given in the appendices. Finally, we conclude and discuss open problems in Section 4.

1.1 Problem formulation

There are a variety of problem formulations in the growing body of work on compressive sensing and related areas. The signal model may be exactly sparse, approximately sparse, or compressible (i.e. that the signal is approximately sparse in some orthonormal basis). The most common signal model is a deterministic one, although Bayesian formulations are also possible. In addition, the observation model can be either noiseless or noisy, and the measurement matrix can be random or deterministic. Furthermore, the signal recovery can be perfect or approximate, assessed by various error metrics (e.g., ℓ_q-norms, prediction error, subset recovery).

In this paper, we consider a deterministic signal model, in which $\beta \in \mathbb{R}^p$ is a fixed but unknown vector with exactly k non-zero entries. We refer to k as the *signal sparsity* and p as the *signal dimension*, and define the support set of β as

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

Note that there are $N = \binom{p}{k}$ possible support sets, corresponding to the N possible k-dimensional

subspaces in which β can lie. We are given a vector of n noisy observations $Y \in \mathbb{R}^n$, of the form

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

where $X \in \mathbb{R}^{n \times p}$ is the measurement matrix, and $W \sim N(0, \sigma^2 I_{n \times n})$ is additive Gaussian noise. Our results apply to various classes of dense and γ-sparsified measurement matrices, which will be defined concretely in Section 2. Throughout this paper, we assume without loss of generality that $\sigma^2 = 1$, since any scaling of σ can be accounted for in the scaling of β.

Our goal is to perform exact recovery of the support set S, which corresponds to a standard model selection error criterion. More precisely, we measure the error between the estimate $\widehat{\beta}$ and the true signal β using the $\{0,1\}$-valued loss function:

$$\rho(\widehat{\beta}, \beta) := \mathbb{I}\left[\{\widehat{\beta}_i \neq 0, \ \forall i \in S\} \cap \{\widehat{\beta}_j = 0, \ \forall j \notin S\}\right]. \tag{3}$$

The results of this paper apply to arbitrary decoders. Any decoder is a mapping g from the observations Y to an estimated subset $\widehat{S} = g(Y)$. Let $\mathbb{P}[g(Y) \neq S \mid S]$ be the conditional probability of error given that the true support is S. Assuming that β has support S chosen uniformly at random over the N possible subsets of size k, the average probability of error is given by

$$p_{err} \quad = \quad \frac{1}{\binom{p}{k}} \sum_{S} \mathbb{P}[g(Y) \neq S \mid S]. \tag{4}$$

We say that sparsity recovery is asymptotically reliable if $p_{err} \to 0$ as $n \to \infty$. Since we are trying to recover the support exactly from noisy measurements, our results necessarily involve the minimum value of β on its support,

$$\beta_{min} \quad := \quad \min_{i \in S} |\beta_i|. \tag{5}$$

In particular, our results apply to decoders that operate over the signal class

$$\mathcal{C}(\beta_{min}) \quad := \quad \{\beta \in \mathbb{R}^p \mid |\beta_i| \geq \beta_{min} \ \forall i \in S\}. \tag{6}$$

With this set-up, our goal is to find necessary conditions on the parameters $(n, p, k, \beta_{min}, \gamma)$ 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. As our analysis shows, the appropriate notion of rate for this problem is $R = \frac{\log \binom{p}{k}}{n}$.

1.2 Our contributions

One body of past work [12, 18, 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

$$\text{SNR} \quad := \quad \frac{\mathbb{E}[\|X\beta\|_2^2]}{\mathbb{E}[\|W\|_2^2]} \quad = \quad \|\beta\|_2^2. \tag{7}$$

(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 (7), though appropriate for ℓ_2-distortion, is not suitable for the support recovery problem. Although the minimum value is related to this power-based measure by the inequality $k\beta_{min}^2 \leq \text{SNR}$, for the ensemble of signals $\mathcal{C}(\beta_{min})$ defined in equation (6), the ℓ_2-based SNR measure (7) 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 $\beta_{min} \to 0$ at an arbitrarily rapid rate—even as the power-based SNR (7) becomes arbitrarily large.

The paper [24] was the first to consider the information-theoretic limits of exact subset recovery using dense Gaussian measurement ensembles, explicitly identifying the minimum value β_{min} as the key parameter. This analysis yielded necessary and sufficient conditions on general quadruples (n, p, k, β_{min}) for asymptotically reliable recovery. Subsequent work [16, 2] has extended this type of analysis to the criterion of partial support recovery. In this paper, we consider only exact support recovery, but provide results for general dense measurement ensembles, thereby extending previous results. In conjunction with known sufficient conditions [24], 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 ($n = \Theta(p)$). Moreover, for the special case of the standard Gaussian ensemble, Theorem 1 also recovers some results independently obtained in concurrent work by Reeves [16], and Fletcher et al. [11].

We then consider the effect of measurement sparsity, which we assess in terms of the fraction $\gamma \in (0, 1]$ of non-zeros per row of the the measurement matrix X. Some past work in compressive sensing has proposed computationally efficient recovery methods based on sparse measurement matrices, including work inspired by expander graphs and coding theory [26, 19], sparse random projections for Johnson-Lindenstrauss embeddings [25], and sketching and group testing [13, 7]. All of this work deals with the noiseless observation model, in contrast to the noisy observation model (2) considered here. The paper [1] provides results on sparse measurements for noisy problems and distortion-type error metrics, using a Bayesian signal model and power-based SNR that is not appropriate for the subset recovery problem. Also, some concurrent work [15] provides sufficient conditions for support recovery using the Lasso (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 this paper, 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 (see equation (8)), for general scalings of the parameters $(n, p, k, \beta_{min}, \gamma)$. This analysis reveals three regimes of interest, corresponding to whether measurement sparsity has no 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.

2 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 non-zero 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.). 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

$$X_{ij} = \begin{cases} N(0, \frac{1}{\gamma}) & \text{w.p. } \gamma \\ 0 & \text{w.p. } 1-\gamma \end{cases}. \tag{8}$$

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

2.1 Tighter bounds on dense ensembles

We begin by noting an analogy to the Gaussian channel coding problem that yields a straightforward but loose set of necessary conditions. Support recovery can be viewed as a channel coding problem, in which there are $N = \binom{p}{k}$ possible support sets of β, corresponding to messages to be sent over a Gaussian channel with noise variance 1. The effective code rate is then $R = \frac{\log \binom{p}{k}}{n}$. If each support set S is encoded as the codeword $c(S) = X\beta$, where X has i.i.d. Gaussian entries, then by standard Gaussian channel capacity results, we immediately obtain a lower bound on the number of observations n necessary for asymptotically reliable recovery,

$$n \quad > \quad \frac{\log \binom{p}{k}}{\frac{1}{2} \log \left(1 + \|\beta\|_2^2\right)}. \tag{9}$$

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 $k > 1$. The following result provides sharper conditions on subset recovery.

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 variance one. Then a necessary condition for asymptotically reliable recovery over the signal class $\mathcal{C}(\beta_{min})$ is*

$$n \quad > \quad \max\left\{ f_1(p, k, \beta_{min}), \quad f_2(p, k, \beta_{min}), \quad k - 1 \right\}, \tag{10}$$

where

$$f_1(p,k,\beta_{min}) \;:=\; \frac{\log\binom{p}{k}-1}{\frac{1}{2}\log\left(1+k\beta_{min}^2(1-\frac{k}{p})\right)} \tag{11a}$$

$$f_2(p,k,\beta_{min}) \;:=\; \frac{\log(p-k+1)-1}{\frac{1}{2}\log\left(1+\beta_{min}^2(1-\frac{1}{p-k+1})\right)}. \tag{11b}$$

The proof of Theorem 1, given in Section 3, uses Fano's inequality to bound the probability of error of any recovery method. 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 $X_{ij} \in \{-1,+1\}$). It generalizes and strengthens earlier results on subset recovery [24]. Note that $\|\beta\|_2^2 \geq k\beta_{min}^2$ (with equality in the case when $|\beta_i| = \beta_{min}$ for all indices $i \in S$), so that this bound is strictly tighter than the intuitive bound (9). Moreover, by fixing the value of β at $(k-1)$ indices to β_{min} and allowing the last component of β to tend to infinity, we can drive the power $\|\beta\|_2^2$ to infinity, while still having the minimum enter the lower bound.

The necessary conditions in Theorem 1 can be compared against the sufficient conditions in Wainwright [24] for exact support recovery using the standard Gaussian ensemble, as shown in Table 1. We obtain tight necessary and sufficient conditions in the regime of linear signal sparsity (meaning $k/p = \alpha$ for some $\alpha \in (0,1)$), under various scalings of the minimum value β_{min}. We also obtain tight matching conditions in the regime of sublinear signal sparsity (in which $k/p \to 0$), when $k\beta_{min}^2 = \Theta(1)$. There remains a slight gap, however, in the sublinear sparsity regime when $k\beta_{min}^2 \to \infty$ (see bottom two rows in Table 1). Moreover, these information-theoretic bounds can be compared to the recovery threshold of ℓ_1-constrained quadratic programming, known as the Lasso [23]. This comparison reveals that whenever $k\beta_{min}^2 = \Theta(1)$ (in both the linear and sublinear sparsity regimes), then $\Theta(k\log(p-k))$ observations are necessary and sufficient for sparsity recovery, and hence the Lasso method is information-theoretically optimal. In contrast, when $k\beta_{min}^2 \to \infty$ and $k/p = \alpha$, there is a gap between the performance of the Lasso and the information-theoretic bounds.

Theorem 1 has some consequences related to results proved in concurrent work. Reeves and Gastpar [16] have shown that in the regime of linear sparsity $k/p = \alpha > 0$, if any decoder is given only a linear fraction sample size (meaning that $n = \Theta(p)$), then in order to recover the support exactly, one must have $k\beta_{min}^2 \to +\infty$. This result is one corollary of Theorem 1, since if $\beta_{min}^2 = \Theta(1/k)$, 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 other concurrent work, Fletcher et al. [11] used direct methods to show that for the special case of the standard Gaussian ensemble, the number of observations must satisfy $n > \Omega\left(\frac{\log(p-k)}{\beta_{min}^2}\right)$. This bound is a consequence of our lower bound $f_2(p,k,\beta_{min})$; moreover, Theorem 1 implies the same lower bound for general (non-Gaussian) ensembles as well.

In the regime of linear sparsity, Wainwright [24] showed, by direct analysis of the optimal decoder, that the scaling $\beta_{min}^2 = \Omega(\log(k)/k)$ 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:

	Necessary conditions (Theorem 1)	Sufficient conditions (Wainwright [24])
$k = \Theta(p)$ $\beta_{min}^2 = \Theta(\frac{1}{k})$	$\Theta(p \log p)$	$\Theta(p \log p)$
$k = \Theta(p)$ $\beta_{min}^2 = \Theta(\frac{\log k}{k})$	$\Theta(p)$	$\Theta(p)$
$k = \Theta(p)$ $\beta_{min}^2 = \Theta(1)$	$\Theta(p)$	$\Theta(p)$
$k = o(p)$ $\beta_{min}^2 = \Theta(\frac{1}{k})$	$\Theta(k \log(p-k))$	$\Theta(k \log(p-k))$
$k = o(p)$ $\beta_{min}^2 = \Theta(\frac{\log k}{k})$	$\max\left\{\Theta\left(\frac{k \log \frac{p}{k}}{\log \log k}\right), \Theta\left(\frac{k \log(p-k)}{\log k}\right)\right\}$	$\Theta\left(k \log \frac{p}{k}\right)$
$k = o(p)$ $\beta_{min}^2 = \Theta(1)$	$\max\left\{\Theta\left(\frac{k \log \frac{p}{k}}{\log k}\right), \Theta(k)\right\}$	$\Theta\left(k \log \frac{p}{k}\right)$

Table 1. Tight necessary and sufficient conditions on the number of observations n required for exact support recovery are obtained in several regimes of interest.

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 $\beta_{min}^2 = \Omega(\log k/k)$.*

2.2 Effect of measurement sparsity

We now turn to the effect of measurement sparsity on recovery, considering in particular the γ-sparsified ensemble (8). Even though the average signal-to-noise ratio of our channel remains the same (since $\mathrm{var}(X_{ij}) = 1$ for all choices of γ by construction), the Gaussian channel coding bound (9) is clearly not tight for sparse X, even in the case of $k = 1$. The loss in statistical efficiency is due to the fact that we are constraining our codebook to have a sparse structure, which may be far from a capacity-achieving code. Theorem 1 applies to any ensemble in which the components are zero-mean and unit variance. However, if we apply it to the γ-sparsified ensemble, it yields lower bounds that are independent of γ. Intuitively, it is clear that the procedure of γ-sparsification should cause deterioration in support recovery. Indeed, the following result provides refined bounds that capture the effects of γ-sparsification. Let $\phi(\mu, \sigma^2)$ denote the Gaussian density with mean μ and variance σ^2, and define the following two mixture distributions:

$$\overline{\psi}_1 := \sum_{l=0}^{k} \binom{k}{\ell} \gamma^\ell (1-\gamma)^{k-\ell} \, \phi\left(0, 1 + \frac{\ell \beta_{min}^2}{\gamma}\right) \tag{12}$$

$$\overline{\psi}_2 := \gamma \, \phi\left(0, 1 + \frac{\beta_{min}^2}{\gamma}\right) + (1-\gamma) \, \phi(0,1). \tag{13}$$

Furthermore, let $H(\cdot)$ denote the 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* (8)*. Then a necessary condition for asymptotically*

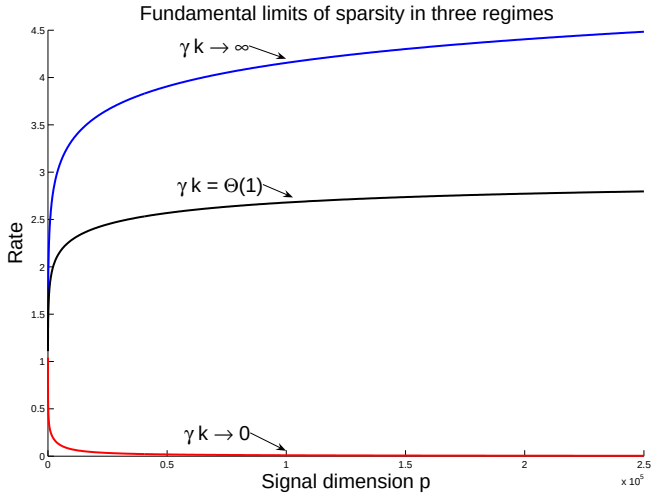

Figure 1. The rate $R = \frac{\log \binom{p}{k}}{n}$ is plotted using equation (14) in three regimes, depending on how the quantity γk scales, where $\gamma \in [0, 1]$ denotes the measurement sparsification parameter and k denotes the signal sparsity.

reliable recovery over the signal class $\mathcal{C}(\beta_{min})$ *is*

$$n \quad > \quad \max\left\{g_1(p, k, \beta_{min}, \gamma), \quad g_2(p, k, \beta_{min}, \gamma), \quad k - 1\right\}, \tag{14}$$

where

$$g_1(p, k, \beta_{min}, \gamma) \quad := \quad \frac{\log \binom{p}{k} - 1}{H(\overline{\psi}_1) - \frac{1}{2}\log(2\pi \mathrm{e})} \tag{15a}$$

$$g_2(p, k, \beta_{min}, \gamma) \quad := \quad \frac{\log(p - k + 1) - 1}{H(\overline{\psi}_2) - \frac{1}{2}\log(2\pi \mathrm{e})}. \tag{15b}$$

The proof of Theorem 2, given in Section 3, again uses Fano's inequality, but explicitly analyzes the effect of measurement sparsification on the distribution of the observations. The necessary condition in Theorem 2 is plotted in Figure 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 thresholds at which measurement sparsity begins to degrade the performance of any decoder, Corollary 2 below further bounds the necessary conditions in Theorem 2 in three cases. For any scalar γ, let $H_{binary}(\gamma)$ denote the entropy of a Ber(γ) variate.

Corollary 2 (Three regimes)**.** *The necessary conditions in Theorem 2 can be simplified as follows.*

(a) In general,

$$g_1(p, k, \beta_{min}, \gamma) \quad \geq \quad \frac{\log \binom{p}{k} - 1}{\frac{1}{2}\log\left(1 + k\beta_{min}^2\right)}, \tag{16a}$$

$$g_2(p, k, \beta_{min}, \gamma) \quad \geq \quad \frac{\log(p - k + 1) - 1}{\frac{1}{2}\log\left(1 + \beta_{min}^2\right)}. \tag{16b}$$

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

Table 2. Necessary conditions on the number of observations n required for exact support recovery is shown in different regimes of the parameters $(p, k, \beta_{min}, \gamma)$.

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

$$g_1(p, k, \beta_{min}, \gamma) \geq \frac{\log \binom{p}{k} - 1}{\frac{1}{2}\tau \log\left(1 + \frac{k\beta_{min}^2}{\tau}\right) + C}, \tag{17a}$$

$$g_2(p, k, \beta_{min}, \gamma) \geq \frac{\log(p-k+1) - 1}{\frac{1}{2}(\frac{\tau}{k}) \log\left(1 + \frac{k\beta_{min}^2}{\tau}\right) + H_{binary}(\frac{\tau}{k})}, \tag{17b}$$

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

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

$$g_1(p, k, \beta_{min}, \gamma) \geq \frac{\log \binom{p}{k} - 1}{\frac{1}{2}\gamma k \log\left(1 + \frac{\beta_{min}^2}{\gamma}\right) + k H_{binary}(\gamma)}, \tag{18a}$$

$$g_2(p, k, \beta_{min}, \gamma) \geq \frac{\log(p-k+1) - 1}{\frac{1}{2}\gamma \log\left(1 + \frac{\beta_{min}^2}{\gamma}\right) + H_{binary}(\gamma)}. \tag{18b}$$

Corollary 2 reveals three regimes of behavior, defined by the scaling of the measurement sparsity γ and the signal sparsity k. If $\gamma k \to \infty$ as $p \to \infty$, then the recovery threshold (16) 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 recovery threshold (18) changes fundamentally compared to the dense case. Finally, if $\gamma k = \Theta(1)$, then the recovery threshold (17) transitions between the two extremes. Using the bounds in Corollary 2, the necessary conditions in Theorem 2 are shown in Table 2 under different scalings of the parameters $(n, p, k, \beta_{min}, \gamma)$. In particular, if $\gamma = o(\frac{1}{k \log k})$ and the minimum value β_{min}^2 does not increase with k, then the denominator $\gamma k \log \frac{1}{\gamma}$ goes to zero. Hence, the number of measurements that any decoder needs in order to recover reliably increases dramatically in this regime.

3 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, β_{min}) (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 inequality [8] to lower bound the probability of error. In order to establish the two types of necessary conditions (e.g, $f_1(p, k, \beta_{min})$ versus $f_2(p, k, \beta_{min})$), we consider two classes of restricted 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 Figure 2a. We note that although the first restricted ensemble is a harder problem, applying Fano to the second restricted ensemble yields a tighter analysis in some regimes. In all cases, we assume that the support S of the unknown vector $\beta \in \mathbb{R}^p$ is chosen randomly and uniformly over all $\binom{p}{k}$ possible support sets. 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.

Restricted ensemble A: In the first restricted problem, also exploited in previous work [24], we assume that while the support set S is unknown, the decoder knows *a priori* that $\beta_j = \beta_{min}$ for all $j \in S$. In other words, the decoder knows the value of β on its support, but it does not know the locations of the non-zeros. Conditioned on the event that S is the true underlying support of β, the observation vector $Y \in \mathbb{R}^n$ can then be written as

$$Y \quad := \quad \sum_{j \in S} X_j \beta_{min} + W. \tag{19}$$

If a decoder can recover the support of any p-dimensional k-sparse vector β, then it must be able to recover a k-sparse vector that is constant on its support. Furthermore, having knowledge of the value β_{min} at the decoder cannot increase the probability of error. Finally, we assume that $\beta_j = \beta_{min}$ for all $j \in S$ to construct the most difficult possible instance within our ensemble. Thus, we can apply Fano's inequality to lower bound the probability of error in the restricted problem, and so obtain a lower bound on the probability of error for the general problem. This procedure yields the lower bounds $f_1(p, k, \beta_{min})$ and $g_1(p, k, \beta_{min}, \gamma)$ in Theorems 1 and 2 respectively.

Restricted ensemble B: The second restricted ensemble is designed to capture the confusable effects of the relatively small number $(p - k + 1)$ of very close-by subsets (see Figure 2b). This restricted ensemble is defined as follows. Suppose that the decoder is given the locations of all but the smallest non-zero value of the vector β, as well as the values of β on its support. More precisely, let $j^\star$ denote the unknown location of the smallest non-zero value of β, which we assume achieves the minimum (i.e., $\beta_{j^\star} = \beta_{min}$), and let $T = S \setminus \{j^\star\}$. Given knowledge of $(T, \beta_T, \beta_{min})$, 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

$$\widetilde{Y} \quad := \quad X_{j^\star} \beta_{min} + W. \tag{20}$$

By re-ordering indices as need be, we may assume without loss of generality that $T = \{p - k + 2, \ldots, p\}$, so that $j^\star \in \{1, \ldots, p - k + 1\}$. The remaining sub-problem is to determine, given the

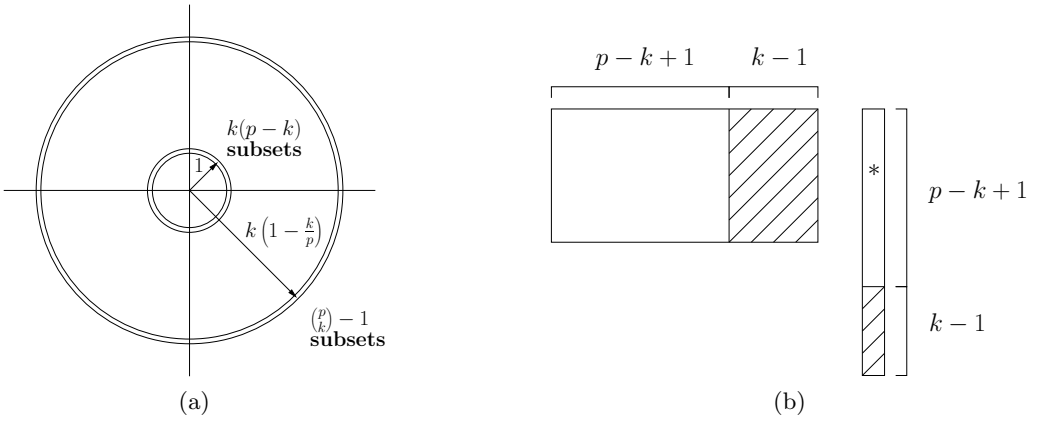

Figure 2. Illustration of restricted ensembles. (a) In restricted ensemble A, the decoder must distinguish between $\binom{p}{k}$ support sets with an average overlap of size $\frac{k^2}{p}$, whereas in restricted ensemble B, it must decode amongst a subset of the $k(p-k)+1$ supports with overlap $k-1$.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

observations $\widetilde{Y}$, the location of the single non-zero. Note that when we assume that the support of β is uniformly chosen over all $\binom{p}{k}$ possible subsets of size k, then given T, the location of the remaining non-zero is uniformly distributed over $\{1,\ldots,p-k+1\}$.

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. Let $\widetilde{\beta} \in \mathbb{R}^{p-k+1}$ be a vector with exactly one non-zero. We can augment $\widetilde{\beta}$ with $k-1$ non-zeros at the end to obtain a p-dimensional vector. If a decoder can recover the support of any p-dimensional k-sparse vector β, then it can recover the support of the augmented $\widetilde{\beta}$, and hence the support of $\widetilde{\beta}$. Similarly, providing the decoder with side information about the non-zero values of β cannot increase the probability of error. As before, we can apply Fano's inequality to lower bound the probability of error in this restricted problem, thereby obtaining the lower bounds $f_2(p,k,\beta_{min})$ and $g_2(p,k,\beta_{min},\gamma)$ in Theorems 1 and 2 respectively.

3.1 Proof of Theorem 1

In this section, we derive the necessary conditions $f_1(p,k,\beta_{min})$ and $f_2(p,k,\beta_{min})$ in Theorem 1 for the general class of measurement matrices, by applying Fano's inequality to bound the probability of decoding error in restricted problems A and B, respectively.

3.1.1 Applying Fano to restricted ensemble A

We first perform our analysis of the error probability for a particular instance of the random measurement matrix X, and subsequently average over the ensemble of matrices. Let Ω denote a random subset chosen uniformly at random over all $\binom{p}{k}$ subsets $S \subset \{1,\ldots,p\}$ of size k. The probability of decoding error, for a given X, can be lower bounded by Fano's inequality as

$$p_{err}(X) \;\geq\; \frac{H(\Omega|Y)-1}{\log\binom{p}{k}} \;=\; 1-\frac{I(\Omega;Y)+1}{\log\binom{p}{k}}$$

where we have used the fact that $H(\Omega|Y) = H(\Omega) - I(\Omega;Y) = \log\binom{p}{k} - I(\Omega;Y)$. Thus the problem is reduced to upper bounding the mutual information $I(\Omega;Y)$ between the random subset Ω and the noisy observations Y. Since both X and β_{min} are known and fixed, the mutual information can be expanded as

$$I(\Omega;Y) \;=\; H(Y) - H(Y|\Omega) \;=\; H(Y) - H(W).$$

We first bound the entropy of the observation vector $H(Y)$, using the fact that differential entropy is maximized by the Gaussian distribution with a matched variance. More specifically, for a given X, let $\Lambda(X)$ denote the covariance matrix of Y conditioned on X. (Hence entry $\Lambda_{ii}(X)$ on the diagonal represents the variance of Y_i.) With this notation, the entropy of Y can be bounded as

$$\begin{aligned} H(Y) &\leq \sum_{i=1}^{n} H(Y_i) \\ &\leq \sum_{i=1}^{n} \frac{1}{2}\log(2\pi \mathrm{e}\ \Lambda_{ii}(X)). \end{aligned}$$

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 mutual information,

$$I(\Omega;Y) \;\leq\; \sum_{i=1}^{n} \frac{1}{2}\log(\Lambda_{ii}(X)).$$

With this bound on the mutual information, we now average the probability of error over the ensemble of measurement matrices X. Exploiting the concavity of the logarithm and applying Jensen's inequality, the average probability of error can be bounded as

$$\mathbb{E}_X[p_{err}(X)] \;\geq\; 1 - \frac{\sum_{i=1}^{n} \frac{1}{2}\log(\mathbb{E}_X[\Lambda_{ii}(X)]) + 1}{\log\binom{p}{k}}. \tag{21}$$

It remains to compute the expectation $\mathbb{E}_X[\Lambda_{ii}(X)]$, over the ensemble of matrices 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. X_{ij} with zero-mean and unit variance, the average covariance is given by*

$$\mathbb{E}_X[\Lambda(X)] \;=\; \left(1 + k\beta_{min}^2\left(1 - \frac{k}{p}\right)\right) I_{n\times n}. \tag{22}$$

Finally, combining Lemma 1 with equation (21), we obtain that the average probability of error is bounded away from zero if

$$n \;<\; \frac{\log\binom{p}{k} - 1}{\frac{1}{2}\log\left(k\beta_{min}^2\left(1 - \frac{k}{p}\right) + 1\right)},$$

as claimed.

3.1.2 Applying Fano to restricted ensemble B

The analysis of restricted ensemble B is completely analogous to the proof for restricted ensemble A, so we will only outline the key steps below. Let Ω denote a random variable with uniform distribution over the indices $\{1, \ldots, p-k+1\}$. The probability of decoding error, for a given measurement matrix X, can be lower bounded by Fano's inequality as

$$p_{err}(X) \quad \geq \quad 1 - \frac{I(\Omega; \widetilde{Y}) + 1}{\log(p-k+1)}$$

As before, the key problem of bounding the mutual information $I(\Omega; \widetilde{Y})$ between the random index Ω and the modified observation vector $\widetilde{Y}$, can be reduced to bounding the entropy $H(\widetilde{Y})$. For each fixed X, let $\Lambda(X)$ denote the covariance matrix of $\widetilde{Y}$. Since the differential entropy of $\widetilde{Y}_i$ is upper bounded by the entropy of a Gaussian distribution with variance $\Lambda_{ii}(X)$, we obtain the following bound on the mutual information

$$\begin{aligned} I(\Omega; \widetilde{Y}) &= H(\widetilde{Y}) - \frac{n}{2}\log(2\pi \mathrm{e}) \\ &\leq \sum_{i=1}^{n} \frac{1}{2}\log(\Lambda_{ii}(X)). \end{aligned}$$

Applying Jensen's inequality, we can then bound the average probability of error, averaged over the ensemble of measurement matrices X, as

$$\mathbb{E}_X[p_{err}(X)] \quad \geq \quad 1 - \frac{\sum_{i=1}^{n} \frac{1}{2}\log(\mathbb{E}_X[\Lambda_{ii}(X)]) + 1}{\log(p-k+1)}. \tag{23}$$

The proof of Lemma 2 below follows the same steps as the derivation of Lemma 1, and is omitted.

Lemma 2. *Given i.i.d. X_{ij} with zero-mean and unit variance, the average covariance is given by*

$$\mathbb{E}_X[\Lambda(X)] \quad = \quad \left(1 + \beta_{min}^2 \left(1 - \frac{1}{p-k+1}\right)\right) I_{n \times n}. \tag{24}$$

Finally, combining Lemma 2 with the Fano bound (23), we obtain that the average probability of error is bounded away from zero if

$$n \quad < \quad \frac{\log(p-k+1) - 1}{\frac{1}{2}\log(1 + \beta_{min}^2(1 - \frac{1}{p-k+1}))}$$

as claimed.

3.2 Proof of Theorem 2

This section contains proofs of the necessary conditions in Theorem 2 for the γ-sparsified Gaussian measurement ensemble (8). We proceed as before, applying Fano's inequality to restricted problems A and B, in order to derive the conditions $g_1(p, k, \beta_{min}, \gamma)$ and $g_2(p, k, \beta_{min}, \gamma)$, respectively.

3.2.1 Analyzing restricted ensemble A

In analyzing the probability of error in restricted ensemble A, the initial steps proceed as in the proof of Theorem 1, first bounding the probability of error for a fixed instance of the measurement matrix X, and later averaging over the γ-sparsified Gaussian ensemble (8). Let Ω denote a random subset uniformly distributed over the $\binom{p}{k}$ possible subsets $S \subset \{1, \ldots, p\}$ of size k. As before, the probability of decoding error, for each fixed X, can be lower bounded by Fano's inequality as

$$p_{err}(X) \;\geq\; 1 - \frac{I(\Omega; Y) + 1}{\log \binom{p}{k}}.$$

We can similarly bound the mutual information

$$\begin{aligned} I(\Omega; Y) &= H(Y) - H(W) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \frac{n}{2}\log(2\pi \mathrm{e}), \end{aligned}$$

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

From this point, the key subproblem is to compute the entropy of $Y_i = \sum_{j\in S} X_{ij}\beta_{min} + W_i$. To characterize the limiting behavior of the random variable Y_i, note that Y_i is distributed according to the density defined as

$$\psi_1(y, i; X) \;=\; \frac{1}{\binom{p}{k}} \sum_{S} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \beta_{min} \sum_{j\in S} X_{ij})^2\right).$$

For each fixed matrix X, this density is a mixture of Gaussians with unit variances and means that depend on the values of $\{X_{i1}, \ldots, X_{ip}\}$, summed over subsets $S \subset \{1, \ldots, p\}$ with $|S| = k$. At a high-level, our immediate goal is to characterize the entropy $H(\psi_1)$.

Note that as X varies over the ensemble (8), the sequence $\{\psi_1(\cdot\,; X)\}_p$, 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 3. *Let X be drawn with i.i.d. entries from the γ-sparsified Gaussian ensemble* (8)*. For each fixed y and for all $i = 1, \ldots, n$, $\mathbb{E}_X[\psi_1(y, i; X)] = \overline{\psi}_1(y)$, where*

$$\overline{\psi}_1(y) \;=\; \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi(1 + \frac{L\beta_{min}^2}{\gamma})}} \exp\left(-\frac{y^2}{2(1 + \frac{L\beta_{min}^2}{\gamma})}\right)\right] \tag{25}$$

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

For certain scalings, we can use concentration results for U-statistics [20] to prove that ψ_1 converges uniformly to $\overline{\psi}_1$, and from there that $H(\psi_1) \xrightarrow{p} H(\overline{\psi}_1)$. In general, however, we always have an upper bound, which is sufficient for our purposes. Indeed, since differential entropy $H(\psi_1)$ is a concave function of ψ_1, by Jensen's inequality and Lemma 3, we have

$$\mathbb{E}_X[H(\psi_1)] \;\leq\; H\left(\mathbb{E}_X[\psi_1]\right) \;=\; H(\overline{\psi}_1).$$

With these ingredients, we conclude that the average error probability of any decoder, averaged over the sparsified Gaussian measurement ensemble, is lower bounded by

$$\begin{aligned}
\mathbb{E}_X[p_{err}(X)] &\geq 1 - \frac{\sum_{i=1}^n \mathbb{E}_X[H(Y_i)] - \frac{n}{2}\log(2\pi \mathrm{e}) + 1}{\log \binom{p}{k}} \\
&= 1 - \frac{\sum_{i=1}^n \mathbb{E}_X[H(\psi_1)] - \frac{n}{2}\log(2\pi \mathrm{e}) + 1}{\log \binom{p}{k}} \\
&\geq 1 - \frac{nH(\overline{\psi}_1) - \frac{n}{2}\log(2\pi \mathrm{e}) + 1}{\log \binom{p}{k}}.
\end{aligned}$$

Therefore, the probability of decoding error is bounded away from zero if

$$n < \frac{\log \binom{p}{k} - 1}{H(\overline{\psi}_1) - \frac{1}{2}\log(2\pi \mathrm{e})},$$

as claimed.

3.2.2 Analyzing restricted ensemble B

The analysis of restricted ensemble B mirrors exactly the derivation of restricted ensemble A. Hence we only outline the key steps in this section. Letting $\Omega \sim \mathrm{Uni}\{1, \ldots, p-k+1\}$, we again apply Fano's inequality to restricted problem B, using the sparse measurement ensemble (8):

$$p_{err}(X) \geq 1 - \frac{I(\Omega; \widetilde{Y}) + 1}{\log(p-k+1)}.$$

In order to upper bound $I(\Omega; \widetilde{Y})$, we need to upper bound the entropy $H(\widetilde{Y})$. The sequence of densities associated with $\widetilde{Y}_i$ becomes

$$\psi_2(y, i; X) = \frac{1}{p-k+1} \sum_{j=1}^{p-k+1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \beta_{min} X_{ij})^2\right).$$

Lemma 4 below characterizes the average pointwise behavior of these densities, and follows from the proof of Lemma 3, with S taken to be subsets of the indices $\{1, \ldots, p-k+1\}$ of size $|S| = 1$.

Lemma 4. *Let X be drawn with i.i.d. entries according to (8). For each fixed y and for all $i = 1, \ldots, n$, $\mathbb{E}_X[\psi_2(y, i; X)] = \overline{\psi}_2(y)$, where*

$$\overline{\psi}_2(y) = \mathbb{E}_B\left[\frac{1}{\sqrt{2\pi(1 + \frac{B\beta_{min}^2}{\gamma})}} \exp\left(-\frac{y^2}{2(1 + \frac{B\beta_{min}^2}{\gamma})}\right)\right] \tag{26}$$

is a mixture of Gaussians with Bernoulli weights $B \sim \mathrm{Ber}(\gamma)$.

As before, we can apply Jensen's inequality to obtain the bound

$$\mathbb{E}_X[H(\psi_2)] \leq H(\mathbb{E}_X[\psi_2]) = H(\overline{\psi}_2).$$

The necessary condition then follows by the Fano bound on the probability of error.

3.3 Proof of Corollary 2

In this section, we derive bounds on the expressions $g_1(p, k, \beta_{min}, \gamma)$ and $g_2(p, k, \beta_{min}, \gamma)$ in Theorem 2. We begin by noting that the Gaussian mixture distribution $\overline{\psi}_1$ defined in (12) is a strict generalization of the distribution $\overline{\psi}_2$ defined in (13); moreover, setting the parameter $k = 1$ in $\overline{\psi}_1$ recovers $\overline{\psi}_2$. The variance associated with the mixture distribution $\overline{\psi}_1$ is equal to $\sigma_1^2 = 1 + k\beta_{min}^2$, and so the entropy of $\overline{\psi}_1$ is always bounded by the entropy of a Gaussian distribution with variance σ_1^2, as

$$H(\overline{\psi}_1) \;\; \leq \;\; \frac{1}{2}\log(2\pi \mathrm{e}(1 + k\beta_{min}^2)).$$

Similarly, the mixture distribution $\overline{\psi}_2$ has variance equal to $1+\beta_{min}^2$, so that the entropy associated with $\overline{\psi}_2$ can in general be bounded as

$$H(\overline{\psi}_2) \;\; \leq \;\; \frac{1}{2}\log(2\pi \mathrm{e}(1 + \beta_{min}^2)).$$

This yields the first set of bounds in (16).

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 distribution $\overline{\psi}_1$:

Lemma 5. *For the Gaussian mixture distribution $\overline{\psi}_1$ defined in (12),*

$$H(\overline{\psi}_1) \;\; \leq \;\; \mathbb{E}_L\left[\frac{1}{2}\log\left(2\pi \mathrm{e}\left(1 + \frac{L\beta_{min}^2}{\gamma}\right)\right)\right] + H(L),$$

where $L \sim \mathrm{Bin}(k, \gamma)$.

We can further bound the expression in Lemma 5 in three cases, delineated by the quantity γk. The proof of the following claim in given in Appendix D.

Lemma 6. *Let $E = \mathbb{E}_L\left[\frac{1}{2}\log\left(1 + \frac{L\beta_{min}^2}{\gamma}\right)\right]$, where $L \sim \mathrm{Bin}(k, \gamma)$.*

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

$$\frac{1}{4}\log\left(1 + \frac{k\beta_{min}^2}{3}\right) \;\; \leq \;\; E \;\; \leq \;\; \frac{1}{2}\log\left(1 + k\beta_{min}^2\right).$$

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

$$\frac{1}{2}(1 - \mathrm{e}^{-\tau})\log\left(1 + \frac{k\beta_{min}^2}{\tau}\right) \;\; \leq \;\; E \;\; \leq \;\; \frac{1}{2}\tau\log\left(1 + \frac{k\beta_{min}^2}{\tau}\right).$$

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

$$\frac{1}{4}\gamma k\log\left(1 + \frac{\beta_{min}^2}{\gamma}\right) \;\; \leq \;\; E \;\; \leq \;\; \frac{1}{2}\gamma k\log\left(1 + \frac{\beta_{min}^2}{\gamma}\right).$$

Finally, combining Lemmas 5 and 6 with some simple bounds on the entropy of the binomial variate L (given in Appendix E), we obtain the bounds on $g_1(p, k, \beta_{min}, \gamma)$ in (17) and (18).

We can similarly bound the entropy associated with the Gaussian mixture distribution $\overline{\psi}_2$. Since the density $\overline{\psi}_2$ is a special case of the density $\overline{\psi}_1$ with k set to 1, we can again apply Lemma 5 to obtain

$$\begin{aligned} H(\overline{\psi}_2) &\leq \mathbb{E}_B \left[\frac{1}{2} \log \left(2\pi \mathrm{e} \left(1 + \frac{B\beta_{min}^2}{\gamma} \right) \right) \right] + H(B) \\ &= \frac{\gamma}{2} \log \left(1 + \frac{\beta_{min}^2}{\gamma} \right) + H_{binary}(\gamma) + \frac{1}{2} \log(2\pi \mathrm{e}). \end{aligned}$$

We have thus obtained the bounds on $g_2(p, k, \beta_{min}, \gamma)$ in equations (17) and (18).

4 Discussion

In this paper, we have studied the information-theoretic limits of exact support recovery for general scalings of the parameters $(n, p, k, \beta_{min}, \gamma)$. 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 [24], yields a sharp characterization of recovering signals with linear sparsity with a linear fraction of observations (Corollary 2). 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 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 the standard Gaussian ensemble is an information-theoretically optimal choice for the measurement matrix: no other zero-mean unit variance distribution can reduce the number of observations necessary for recovery, and in fact the standard Gaussian distribution achieves matching sufficient bounds [24]. This fact raises an interesting open question on the design of other, more computationally friendly, measurement matrices which are optimal in the information-theoretic sense.

Acknowledgment

The work of WW and KR was supported by NSF grant CCF-0635114. The work of MJW was supported by NSF grants CAREER-CCF-0545862 and DMS-0605165.

A Proof of Lemma 1

We begin by defining some additional notation. Let $\bar{\beta} \in \mathbb{R}^p$ be a k-sparse vector with $\bar{\beta}_j = \beta_{min}$ for all indices j in the support set S. Recall that Ω denotes a random subset uniformly distributed over all $\binom{p}{k}$ possible subsets $S \subset \{1, \ldots, p\}$ with $|S| = k$. Conditioned on the event that $\Omega = S$, the vector of n observations can then be written as

$$Y \;:=\; X_S \bar{\beta}_S + W \;=\; \beta_{min} \sum_{j \in S} X_j + W.$$

Note that for a given instance of the matrix X, the distribution of Y is a Gaussian mixture with density $f(y) = \frac{1}{\binom{p}{k}} \sum_S \phi(X_S \bar{\beta}_S, I)$, where we are using ϕ to denote the density of a Gaussian random vector with mean $X_S \bar{\beta}_S$ and covariance I. Let $\mu(X) = \mu \in \mathbb{R}^n$ and $\Lambda(X) = \Lambda \in \mathbb{R}^{n \times n}$ be the mean vector and covariance matrix of Y, respectively. The covariance matrix of Y can be computed as $\Lambda = \mathbb{E}\left[YY^T\right] - \mu\mu^T$, where

$$\mu \;=\; \frac{1}{\binom{p}{k}} \sum_S X_S \bar{\beta}_S \;=\; \frac{\beta_{min}}{\binom{p}{k}} \sum_S \sum_{j \in S} X_j$$

and

$$\begin{aligned} \mathbb{E}\left[YY^T\right] &= \mathbb{E}\left[(X\bar{\beta})(X\bar{\beta})^T\right] + \mathbb{E}\left[WW^T\right] \\ &= \frac{1}{\binom{p}{k}} \sum_S (X_S \bar{\beta}_S)(X_S \bar{\beta}_S)^T \;+\; I. \end{aligned}$$

With this notation, we can now compute the expectation of the covariance matrix $\mathbb{E}_X\big[\Lambda\big]$, averaged over any distribution on X with independent, zero-mean and unit variance entries. To compute the first term, we have

$$\begin{aligned} \mathbb{E}_X\left[\mathbb{E}\left[YY^T\right]\right] &= \frac{\beta_{min}^2}{\binom{p}{k}} \sum_S \mathbb{E}_X \left[\sum_{j \in S} X_j X_j^T + \sum_{i \neq j \in S} X_i X_j^T \right] \;+\; I \\ &= \frac{\beta_{min}^2}{\binom{p}{k}} \sum_S \sum_{j \in S} I \;+\; I \\ &= \left(1 + k\beta_{min}^2\right) \; I \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}_X\left[\mu\mu^T\right] &= \left(\frac{\beta_{min}}{\binom{p}{k}}\right)^2 \mathbb{E}_X \left[\sum_{S,U} \sum_{j \in S \cap U} X_j X_j^T + \sum_{S,U} \sum_{\substack{i \in S, j \in U \\ i \neq j}} X_i X_j^T \right] \\ &= \left(\frac{\beta_{min}}{\binom{p}{k}}\right)^2 \sum_{S,U} \sum_{j \in S \cap U} I \\ &= \left(\left(\frac{\beta_{min}}{\binom{p}{k}}\right)^2 \sum_{S,U} |S \cap U| \right) \; I. \end{aligned}$$

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

$$\left(\frac{\beta_{min}}{\binom{p}{k}}\right)^2 \sum_{S,U} |S \cap U| \;=\; \frac{\beta_{min}^2}{\binom{p}{k}} \sum_{\lambda=1}^{k} \binom{k}{\lambda}\binom{p-k}{k-\lambda} \lambda.$$

Finally, using a substitution of variables (by setting $\lambda' = \lambda - 1$) and applying Vandermonde's identity [17], we have

$$\begin{aligned}\left(\frac{\beta_{min}}{\binom{p}{k}}\right)^2 \sum_{S,U} |S \cap U| &= \frac{\beta_{min}^2}{\binom{p}{k}}\, k \sum_{\lambda'=0}^{k-1} \binom{k-1}{\lambda'}\binom{p-k}{k-\lambda'-1} \\ &= \frac{\beta_{min}^2}{\binom{p}{k}}\, k \binom{p-1}{k-1} \\ &= \frac{k^2\beta_{min}^2}{p}.\end{aligned}$$

Combining these terms, we conclude that

$$\mathbb{E}_X\big[\Lambda(X)\big] = \left(1 + k\beta_{min}^2\left(1 - \frac{k}{p}\right)\right) I.$$

B Proof of Lemma 3

Consider the following sequences of densities,

$$\psi_1(y, i; X) = \frac{1}{\binom{p}{k}} \sum_S \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \beta_{min} \sum_{j \in S} X_{ij})^2\right)$$

and

$$\overline{\psi}_1(y) = \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi(1 + \frac{L\beta_{min}^2}{\gamma})}} \exp\left(-\frac{y^2}{2(1 + \frac{L\beta_{min}^2}{\gamma})}\right)\right]$$

where $L \sim \mathrm{Bin}(k, \gamma)$. Our goal is to show that for each fixed y and row index i, the pointwise average of the stochastic sequence of densities ψ_1 over the ensemble of matrices X satisfies $\mathbb{E}_X[\psi_1(y, i; X)] = \overline{\psi}_1(y)$. By symmetry, it is sufficient to compute this expectation for the subset $S = \{1, \ldots, k\}$. When each X_{ij} is i.i.d. drawn according to the γ-sparsified ensemble (8), the random variable $Z = (y - \beta_{min} \sum_{j=1}^k X_{ij})$ has a Gaussian mixture distribution which can be described as follows. Denoting the mixture label by $L \sim \mathrm{Bin}(k, \gamma)$, then $Z \sim N\left(y, \frac{\ell\beta_{min}^2}{\gamma}\right)$ if $L = \ell$, for $\ell = 0, \ldots, k$. Thus, conditioned on the mixture label $L = \ell$, the random variable $\tilde{Z} = \frac{\gamma}{\ell\beta_{min}^2}(y - \beta_{min} \sum_{j=1}^k X_{ij})^2$ has a noncentral chi-square distribution with 1 degree of freedom and parameter $\lambda = \frac{\gamma y^2}{\ell\beta_{min}^2}$. Evaluating

$M_{\tilde{Z}}(t) = \mathbb{E}[\mathrm{e}^{t\tilde{Z}}]$, the moment-generating function [3] of $\tilde{Z}$, then gives us the desired quantity,

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\beta_{min}\sum_{j=1}^{k}X_{ij})^2\right)\right] \\
&= \sum_{\ell=0}^{k}\frac{1}{\sqrt{2\pi}}\mathbb{E}_X\left[\exp\left(-\frac{1}{2}(y-\beta_{min}\sum_{j=1}^{k}X_{ij})^2\right)\;\middle|\; L=\ell\right]\mathbb{P}(L=\ell) \\
&= \sum_{\ell=0}^{k}\frac{1}{\sqrt{2\pi}}M_{\tilde{Z}}\left(-\frac{\ell\beta_{min}^2}{2\gamma}\right)\mathbb{P}(L=\ell) \\
&= \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi(1+\frac{L\beta_{min}^2}{\gamma})}}\exp\left(-\frac{y^2}{2(1+\frac{L\beta_{min}^2}{\gamma})}\right)\right]
\end{aligned}$$

as claimed.

C Proof of Lemma 5

Let Z be a random variable distributed according to the density

$$\overline{\psi}_1(y) \;=\; \mathbb{E}_L\left[\frac{1}{\sqrt{2\pi(1+\frac{L\beta_{min}^2}{\gamma})}}\exp\left(-\frac{y^2}{2(1+\frac{L\beta_{min}^2}{\gamma})}\right)\right],$$

where $L \sim Bin(k,\gamma)$. To compute the entropy of Z, we can expand the following mutual information in two ways, $I(Z;L) = H(Z) - H(Z|L) = H(L) - H(L|Z)$, 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

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

Furthermore, we can bound the conditional entropy of L given Z as $0 \le H(L|Z) \le H(L)$. This gives upper and lower bounds on the entropy of Z as

$$H(Z|L) \;\le\; H(Z) \;\le\; H(Z|L) + H(L).$$

D Proof of Lemma 6

We first derive upper and lower bounds in the case when $\gamma k \le 1$. We can rewrite the binomial distribution as

$$p(\ell) \;:=\; \binom{k}{\ell}\gamma^\ell(1-\gamma)^{k-\ell} \;=\; \frac{\gamma k}{\ell}\binom{k-1}{\ell-1}\gamma^{\ell-1}(1-\gamma)^{k-\ell}$$

and hence

$$\begin{aligned} E &= \frac{1}{2}\sum_{\ell=1}^{k} \log\left(1+\frac{\ell\beta_{min}^2}{\gamma}\right) p(\ell) \\ &= \frac{1}{2}\gamma k \sum_{\ell=1}^{k} \frac{\log\left(1+\frac{\ell\beta_{min}^2}{\gamma}\right)}{\ell} \binom{k-1}{\ell-1} \gamma^{\ell-1}(1-\gamma)^{k-\ell}. \end{aligned}$$

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

$$\left(1+\frac{\beta_{min}^2}{\gamma}\right)^{\ell} \geq 1+\frac{\ell\beta_{min}^2}{\gamma}$$

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

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

To derive the lower bound, we will use the fact that $1+x \leq \mathrm{e}^x$ for all $x \in \mathbb{R}$, and $\mathrm{e}^{-x} \leq 1-\frac{x}{2}$ for $x \in [0,1]$.

$$\begin{aligned} E &= \frac{1}{2}\sum_{\ell=1}^{k} \log\left(1+\frac{\ell\beta_{min}^2}{\gamma}\right) p(\ell) \\ &\geq \frac{1}{2}\log\left(1+\frac{\beta_{min}^2}{\gamma}\right) \sum_{\ell=1}^{k} p(\ell) \\ &= \frac{1}{2}\log\left(1+\frac{\beta_{min}^2}{\gamma}\right) (1-(1-\gamma)^k) \\ &\geq \frac{1}{2}\log\left(1+\frac{\beta_{min}^2}{\gamma}\right) (1-\mathrm{e}^{-\gamma k}) \\ &\overset{(a)}{\geq} \frac{1}{2}\log\left(1+\frac{\beta_{min}^2}{\gamma}\right) \left(\frac{\gamma k}{2}\right). \end{aligned}$$

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

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

$$\mathbb{E}_L \left[\frac{1}{2} \log \left(1 + \frac{L\beta_{min}^2}{\gamma} \right) \right] \quad \leq \quad \frac{1}{2} \log(1 + k\beta_{min}^2).$$

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

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

where in the last step we used the fact that $\frac{(\lfloor \gamma k \rfloor - 1)\beta_{min}^2}{\gamma} \geq \frac{(\gamma k - 2)\beta_{min}^2}{\gamma} \geq \frac{k\beta_{min}^2}{3}$ for $\gamma k > 3$, and $\sum_{\ell=\mathrm{median}}^{k} p(\ell) \geq \frac{1}{2}$.

E Bounds on binomial entropy

Lemma 7. *Let $L \sim \mathrm{Bin}(k, \gamma)$. Then*

$$H(L) \quad \leq \quad kH_{binary}(\gamma).$$

Furthermore, if $\gamma = o\left(\frac{1}{k \log k}\right)$, then $kH_{binary}(\gamma) \to 0$ as $k \to \infty$.

Proof. We can express the binomial variate as $L = \sum_{i=1}^{k} Z_i$, where $Z_i \sim \mathrm{Ber}(\gamma)$ i.i.d. Since $H(g(Z_1, \ldots, Z_k)) \leq H(Z_1, \ldots, Z_k)$, we have

$$H(L) \quad \leq \quad H(Z_1, \ldots, Z_k) \quad = \quad kH_{binary}(\gamma).$$

Next we find the limit of $kH_{binary}(\gamma) = k\gamma \log \frac{1}{\gamma} + k(1-\gamma) \log \frac{1}{1-\gamma}$. Let $\gamma = \frac{1}{kf(k)}$, and assume that $f(k) \to \infty$ as $k \to \infty$. Hence the first term can be written as

$$k\gamma \log \frac{1}{\gamma} \quad = \quad \frac{1}{f(k)} \log(kf(k)) \quad = \quad \frac{\log k}{f(k)} + \frac{\log f(k)}{f(k)},$$

and so $k\gamma \log \frac{1}{\gamma} \to 0$ if $f(k) = \omega(\log k)$. The second term can also be expanded as

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

If $f(k) \to \infty$ as $k \to \infty$, then we have the limits

$$\lim_{k\to\infty} \left(1 - \frac{1}{kf(k)}\right)^k \;=\; 1 \qquad \text{and} \qquad \lim_{k\to\infty} \left(1 - \frac{1}{kf(k)}\right) \;=\; 1,$$

which in turn imply that

$$\lim_{k\to\infty} \log \left(1 - \frac{1}{kf(k)}\right)^k \;=\; 0 \qquad \text{and} \qquad \lim_{k\to\infty} \frac{1}{f(k)} \log \left(1 - \frac{1}{kf(k)}\right) \;=\; 0.$$

□

Lemma 8. *Let $L \sim \mathrm{Bin}(k, \gamma)$, then*

$$H(L) \;\leq\; \frac{1}{2} \log(2\pi \mathrm{e}(k\gamma(1-\gamma) + \frac{1}{12})).$$

Proof. We immediately obtain this bound by applying the differential entropy bound on discrete entropy [8]. □

References

[1] S. Aeron, M. Zhao, and S. Venkatesh. Information-theoretic bounds to sensing capacity of sensor networks under fixed snr. In *Information Theory Workshop*, September 2007.

[2] M. Akcakaya and V. Tarokh. Shannon theoretic limits on noisy compressive sampling. Technical Report arXiv:cs.IT:0711.0366, Harvard, November 2007.

[3] L. Birgé. An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics*, number 37 in IMS Lecture Notes, pages 113–133. Institute of Mathematical Statistics, 2001.

[4] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, August 2006.

[5] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.

[6] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[7] G. Cormode and S. Muthukrishnan. Towards an algorithmic theory of compressed sensing. Technical report, Rutgers University, July 2005.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[9] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.

[10] D. Donoho, M. Elad, and V. M. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info Theory*, 52(1):6–18, January 2006.

[11] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions on sparsity pattern recovery. Technical Report arXiv:cs.IT:0804.1839, UC Berkeley, April 2008.

[12] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran. Denoising by sparse approximation: Error bounds based on rate-distortion theory. *Journal on Applied Signal Processing*, 10:1–19, 2006.

[13] A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin. Algorithmic linear dimension reduction in the ℓ_1-norm for sparse vectors. In *Proc. Allerton Conference on Communication, Control and Computing*, Allerton, IL, September 2006.

[14] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 2006. To appear.

[15] D. Omidiran and M. J. Wainwright. High-dimensional subset recovery in noise: Sparsified measurements without loss of statistical efficiency. Technical report, Department of Statistics, UC Berkeley, April 2008. Short version presented at Int. Symp. Info. Theory, July 2008.

[16] G. Reeves. Sparse signal sampling using noisy linear projections. Master's thesis, UC Berkeley, December 2007.

[17] J. Riordan. *Combinatorial Identities*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1968.

[18] S. Sarvotham, D. Baron, and R. G. Baraniuk. Measurements versus bits: Compressed sensing meets information theory. In *Proc. Allerton Conference on Control, Communication and Computing*, September 2006.

[19] S. Sarvotham, D. Baron, and R. G. Baraniuk. Sudocodes: Fast measurement and reconstruction of sparse signals. In *Int. Symposium on Information Theory*, Seattle, WA, July 2006.

[20] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 1980.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[22] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051, March 2006.

[23] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using using ℓ_1-constrained quadratic programs. Technical Report 709, Department of Statistics, UC Berkeley, 2006.

[24] M. J. Wainwright. Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting. Technical Report 725, Department of Statistics, UC Berkeley, January 2007. Presented at International Symposium on Information Theory, June 2007.

[25] W. Wang, M. Garofalakis, and K. Ramchandran. Distributed sparse random projections for refinable approximation. In *Proc. International Conference on Information Processing in Sensor Networks*, Nashville, TN, April 2007.

[26] W. Xu and B. Hassibi. Efficient compressed sensing with deterministic guarantees using expander graphs. In *Information Theory Workshop (ITW)*, September 2007.