

# Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting

Martin Wainwright

Department of Statistics, and  
Department of Electrical Engineering, UC Berkeley,  
Berkeley, CA 94720  
wainwrig@{eecs,stat}.berkeley.edu

**Abstract**—The problem of recovering the sparsity pattern of a fixed but unknown vector  $\beta^* \in \mathbb{R}^p$  based on a set of  $n$  noisy observations arises in a variety of settings, including subset selection in regression, graphical model selection, signal denoising, compressive sensing, and constructive approximation. Of interest are conditions on the model dimension  $p$ , the sparsity index  $s$  (number of non-zero entries in  $\beta^*$ ), and the number of observations  $n$  that are necessary and/or sufficient to ensure asymptotically perfect recovery of the sparsity pattern. This paper focuses on the information-theoretic limits of sparsity recovery: in particular, for a noisy linear observation model based on measurement vectors drawn from the standard Gaussian ensemble, we derive both a set of sufficient conditions for asymptotically perfect recovery using the optimal decoder, as well as a set of necessary conditions that any decoder must satisfy for perfect recovery. This analysis of optimal decoding limits complements our previous work [20] on thresholds for the behavior of  $\ell_1$ -constrained quadratic programming for Gaussian measurement ensembles.

## I. INTRODUCTION

Suppose that we are given a set of  $n$  observations of a fixed but unknown vector  $\beta^* \in \mathbb{R}^p$ . In a variety of settings, it is known *a priori* that the vector  $\beta^*$  is sparse, meaning that its support set  $S$ —corresponding to those indices  $i$  for which  $\beta_i^*$  is non-zero—is relatively small, say with size  $|S| =: s \ll p$ . Sparsity recovery refers to the problem of correctly estimating the support set  $S$  based on a set of noisy observations. This sparsity recovery problem is of broad interest, arising in various areas, including subset selection in regression, structure estimation in graphical models [16], sparse approximation and signal denoising [4], and compressive sensing [6], [2].

A great deal of work over the past few years, which we review briefly in Section I-A, has focused on the performance of computationally tractable methods, many based on  $\ell_1$  or other convex relaxations, both for recovering the exact sparsity pattern as well as related problems in sparse approximation. Of equal interest and complementary in nature, however, are the information-theoretic limits associated with the performance of *any* procedure for sparsity recovery. Such understanding of fundamental limitations is crucial in assessing the behavior of computationally tractable methods. In particular, there is little point in proposing novel methods for sparsity recovery, possibly with higher computational complexity, if currently extant and computationally tractable methods

achieve the information-theoretic limits. On the other hand, an information-theoretic analysis can reveal where there currently exists a gap between the performance of computationally tractable methods, and the fundamental limits. Indeed, the information-theoretic analysis of this paper makes contributions of both types.

With this motivation in mind, the focus of this paper is on the information-theoretic limitations of sparsity recovery. In particular, our analysis focuses on the noisy and high-dimensional setting, meaning that the observations are contaminated by noise, and all three problem parameters—the *number of observations*  $n$ , the *model dimension*  $p$ , and the *sparsity index*  $s$ , defined below—tend to infinity simultaneously. Our main results, stated more precisely in Section I-B, are necessary and sufficient conditions on the triplet  $(n, p, s)$  for asymptotically reliable sparsity recovery. The analysis given here complements our earlier paper [20] that established precise thresholds on the success/failure of  $\ell_1$ -constrained quadratic programming for sparsity recovery. Full details of the results described here can be found in the technical report [21].

### A. Problem formulation and past work

We begin with a more precise formulation of the problem, as well as a discussion of previous work, with emphasis on that most closely related to the results in this paper. Let  $\beta^* \in \mathbb{R}^p$  be a fixed but unknown vector; we refer to the ambient dimension  $p$  as the *model dimension*. Define the support set of  $\beta^*$  as

$$S := \{i \in \{1, \dots, p\} \mid \beta_i^* \neq 0\}. \quad (1)$$

We refer to its size  $s := |S|$  as the *sparsity index*. Finally, suppose that we are given a set of  $n$  observations, of the form

$$Y_i = x_i^T \beta^* + W_i, \quad i = 1, \dots, n \quad (2)$$

where each  $x_i \in \mathbb{R}^p$  is a measurement vector, and  $W_i \sim N(0, \sigma^2)$  is additive Gaussian noise. Of interest are conditions on the triplet  $(n, p, s)$  under which a given method either succeeds or fails in recovering the sparsity pattern  $S$ .

**Observation models:** The linear observation model (2) can be studied in either its noiseless variant ( $\sigma^2 = 0$ ), or the noisy setting ( $\sigma^2 > 0$ ); this paper focuses exclusively the noisy setting. In addition, previous work has addressed both deterministic families and random ensembles of measurement vectors  $\{x_i\}_{i=1}^n$ . The analysis in this paper is based on the

*standard Gaussian measurement ensemble*, in which each measurement vector  $x_i$  is drawn from the zero-mean isotropic Gaussian distribution  $N(0, I_{p \times p})$ .

**Error metrics:** There are various distinct criteria for assessing how close an estimate  $\hat{\beta} \in \mathbb{R}^p$  is to the truth  $\beta^*$ , including (a) various  $\ell_p$  norms  $\|\hat{\beta} - \beta^*\|_p$ , or (b) some measurement of predictive power (e.g.,  $\mathbb{E}[\|Y_i - \hat{Y}_i(\hat{\beta})\|_2^2]$ , where  $\hat{Y}_i$  is the estimate based on  $\hat{\beta}$ ). Given the abundance of recent results on sparse approximation (not all of which are mutually comparable), it is particularly important to specify the choice of error metric. In this paper, we focus exclusively on the sparsity recovery problem, for which the error metric is simply the 0–1 loss associated with the event of recovering the correct support  $S$ —viz.:

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

**Related work:** This paper focuses on information-theoretic limitations of sparsity recovery (i.e., using the error metric (3)) as applied to the standard Gaussian ensemble. Independent work analyzed information-theoretic aspects of sparse estimation problems, including rate-distortion analysis of the  $\ell_2$ -error [12], [17], and modified independent-subspace ensembles [11]. On the other hand, in terms of practical methods, the use of  $\ell_1$ -relaxation for sparse approximation has a lengthy history [4], [18]. There are now various results on the performance of  $\ell_1$ -relaxations, both in the noiseless [10], [15] and noisy setting [19] for deterministic ensembles, as well as the noiseless [8], [2], [9] and noisy setting [3], [1], [7], [16], [22], [20] for random ensembles. Other work has provided conditions under which estimation of a noise-contaminated vector via the Lasso or  $\ell_1$ -constrained quadratic programming [1], [7] or other types of convex relaxation [3] is stable in the  $\ell_2$  sense; however, such  $\ell_2$ -stability does not guarantee exact sparsity recovery.

It should be noted that the results given here apply to completely general scaling of the triplet  $(n, p, s)$ . In contrast, much previous work has addressed one of two possible special cases of sparsity scaling: (a) either the *linear sparsity regime* [e.g. 2], [8], [7], in which  $s = \alpha p$  for some  $\alpha \in (0, 1)$ ; or (b) the *sublinear sparsity regime* [e.g., 16], [22], in which  $s/p$  tends to zero. Depending on the underlying motivation for sparse approximation, both of these sparsity regimes are of independent interest. In covering the full range of scaling, the results given here are complementary to those of our previous paper [20] that provided threshold results, applicable to general scaling of  $(n, p, s)$ , for the success/failure of the Lasso when used for sparsity recovery with random Gaussian measurement ensembles. We discuss connections to previous work in more technical detail following the statement of our main results below.

## B. Our contributions

A decoder is a mapping from the  $n$ -vector of observations  $Y$  to an estimated subset—say of the form  $\hat{S} = \phi(Y)$ . We

think of the underlying true vector  $\beta^* \in \mathbb{R}^p$  with its support  $S$  randomly chosen, uniformly over all  $\binom{p}{s}$  subspaces of size  $s$ . Accordingly, the average error probability  $p_{\text{err}}$  of any decoder is given by

$$p_{\text{err}}(\phi) = \frac{1}{\binom{p}{s}} \sum_{S, |S|=s} \mathbb{P}[\phi(Y) \neq S \mid S].$$

Here the term  $\mathbb{P}[\phi(Y) \neq S \mid S]$  corresponds to the probability, conditioned on the true underlying support being  $S$  and averaging over the measurement noise  $W$ , the choice of Gaussian random matrix  $X$ , and the choice of the entries  $\beta_S^*$  on the fixed support  $S$ , that the decoder makes an error. We say that (a) the sparsity recovery is *asymptotically reliable* (error-free) if  $p_{\text{err}}(\phi) \rightarrow 0$  as  $n \rightarrow +\infty$ , and (b) the sparsity recovery is *asymptotically unreliable* if for some constant  $c > 0$ , the error probability stays bounded  $p_{\text{err}}(\phi) \geq c$  as  $n \rightarrow +\infty$ .

In addition to the three parameters  $(n, p, s)$ , our results also involve the minimum value of the unknown vector  $\beta^*$  on its support, given by

$$\mathcal{M}(\beta^*) := \min_{i \in S} |\beta_i^*|. \quad (4)$$

We begin by stating a set of conditions on the triplet  $(n, p, s)$  which are sufficient to ensure asymptotically perfect recovery of the sparsity pattern:

**Theorem 1** (Sufficient conditions). *If  $(n-s)\mathcal{M}^2(\beta^*) \rightarrow +\infty$ , then the following condition suffices to ensure asymptotically reliable recovery: for some fixed constant  $C > 0$ ,*

$$n > C \max \left\{ s \log(p/s), \frac{1}{\mathcal{M}^2(\beta^*)} \log(p-s) \right\}. \quad (5)$$

The proof of this claim, given in Section II-B, is constructive in nature, based on direct analysis of the error probability associated with the optimal decoder.

**Theorem 2** (Necessary conditions). *Asymptotically reliable recovery is impossible under the following condition: for some fixed constant  $C' > 0$ :*

$$n < \left[ \frac{C'}{s \mathcal{M}^2(\beta^*)} \right] s \log \frac{p}{s}. \quad (6)$$

The proof of this claim, given in Section II-C exploits a corollary of Fano's inequality [13], [5], in order to lower bound the error for a restricted hypothesis testing problem. To interpret these results, we consider two distinct regimes of sparsity:

**Regime of sublinear sparsity:** First suppose that the sparsity is sublinear, meaning that  $s = o(p)$  (including, for instance, the scaling  $s = \mathcal{O}(\sqrt{p})$ ). Based on the two theorems, we identify the critical scaling as  $\mathcal{M}^2(\beta^*) = \Theta(1/s)$ . With this scaling, the sufficient condition in Theorem 1 reduces to  $n > C s \max\{\log(p-s), \log \frac{p}{s}\}$ , whereas the necessary condition in Theorem 2 reduces to  $n < C' s \log \frac{p}{s}$ . For many choices of sublinear sparsity (e.g.,  $s = \mathcal{O}(\sqrt{p})$ ), we have  $\log(p-s) = \Theta(\log \frac{p}{s}) + o(1)$ , so that we can

summarize the two conditions as a threshold of the order  $n = \Theta(s \log(p-s))$ . To compare with our previous work [20] on computationally tractable methods, we established that  $\ell_1$ -constrained quadratic programming (Lasso) has a threshold<sup>1</sup> for success/failure of order  $n = \Theta(s \log(p-s))$ , so that the Lasso very nearly achieves the information-theoretic bounds.

**Regime of linear sparsity:** Next consider the regime of linear sparsity, in which  $s = \alpha p$  for some  $\alpha \in (0, 1)$ . Considering first the sufficient conditions of Theorem 1, we see that as long as  $\mathcal{M}^2(\beta^*)s \rightarrow +\infty$ , then  $n = \Theta(p)$  observations are sufficient to ensure asymptotically reliable recovery. Comparing with our earlier analysis [20] on  $\ell_1$ -constrained quadratic programming, this work showed that if  $n < 2s \log(p-s)$ , then the Lasso fails with probability converging to one, even if  $\mathcal{M}^2(\beta^*)$  stays bounded away from zero. Given this dichotomy, Theorem 1 raises the interesting question: does there exist a computationally efficient technique for reliably recovering a linear sparsity pattern ( $s = \alpha p$ ) based on only a linear fraction of observations ( $n = \Theta(p)$ )?

## II. ANALYSIS

This section is devoted to the proofs of Theorems 1 and 2. We begin by setting up some useful notation to be used throughout the remainder of the paper.

### A. Notation and set-up

For compactness in notation, let us use  $X$  to denote the  $n \times p$  matrix formed with the vectors  $x_k = (x_{k1}, x_{k2}, \dots, x_{kp}) \in \mathbb{R}^p$  as rows, and the vectors  $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T \in \mathbb{R}^n$  as columns. Using  $Y$  and  $W$  to denote the  $n$ -dimensional observation and noise vectors respectively, we can re-write our linear observation model (2) in matrix-vector form as follows:

$$Y = X\beta^* + W. \quad (7)$$

Given any subset  $V \subseteq \{1, \dots, p\}$ , we use the notation  $\beta_V^*$  to denote the  $|V|$ -dimensional subvector  $\{\beta_i^*, i \in V\}$ , and similarly for other vectors (e.g.,  $Y$ , etc.). In an analogous manner, we use  $X_V$  to denote the  $n \times |V|$  matrix with columns  $\{X_i, i \in V\}$ . From herein, we assume without loss of generality that  $\sigma^2 = 1$ , so that  $W \sim N(0, I_{n \times n})$  is simply a standard Gaussian vector. (Note that any scaling of  $\sigma$  can be accounted for in the scaling of  $\beta^*$ , via the parameter  $\mathcal{M}(\beta^*)$ ).

### B. Proof of Theorem 1

**Optimal decoding:** We begin by describing the “best” decoder, that is optimal in terms of minimizing the probability of error  $p_{\text{err}}(\phi)$  over all decoding rules. It is based on the following real-valued function, defined on the subsets  $U \subseteq \{1, \dots, p\}$ , as

$$f(U; Y, X, \beta^*) = \arg \min_{\beta_U} \{\|Y - X_U \beta_U\|_2^2\}. \quad (8)$$

We frequently write  $f(U)$  as a shorthand; note that this value corresponds to the error associated with the best estimator of

<sup>1</sup>Those results [20] allowed the minimum value to scale as  $\mathcal{M}^2(\beta^*) = f(s)/s$ , where  $f$  is any function such that  $\lim_{s \rightarrow +\infty} f(s) = +\infty$ .

$Y$  that lies in  $\text{Ra}(X_U)$ . The optimal decoder chooses the best subset  $\hat{S}$  based on the minimal value of this error, ranging over all subsets  $U$  of size  $s$ :

$$\hat{S} = \phi_{\text{opt}}(Y) := \arg \min_{|U|=s} f(U; Y, X, \beta^*). \quad (9)$$

Note that by symmetry, the error probability  $\mathbb{P}[\hat{S} \neq S \mid S]$  is in fact the same regardless of which underlying set  $S$  acts as the true one. Consequently, we can view the choice of  $S$  as fixed (and hence non-random), and write

$$p_{\text{err}}(\phi) = \mathbb{P}[\phi(Y) \neq S], \quad (10)$$

which should now be understood as an unconditional probability (with  $S$  fixed).

**Analysis of error probability:** Consider the difference  $\Delta(U) := f(U) - f(S)$  between the reconstruction error  $f(S)$  using the true subset  $S$ , versus the error  $f(U)$  candidate subset  $U$ . For any subset  $U$  such that  $X_U$  is full rank, define the  $n \times n$  matrices

$$\Pi_U := X_U [X_U^T X_U]^{-1} X_U^T, \quad \text{and} \quad (11a)$$

$$\Pi_U^\perp := I_{n \times n} - X_U [X_U^T X_U]^{-1} X_U^T. \quad (11b)$$

Note that  $\Pi_U$  and  $\Pi_U^\perp$  are both orthogonal projection matrices, associated with the  $s$ -dimensional range space  $\text{Ra}(X_U)$  and  $(n-s)$ -dimensional nullspace  $\text{Ker}(X_U)$  respectively. With these definitions, some algebraic manipulation yields that for a given vector  $\beta^*$  with support  $S$ , the optimal decoder declares  $U$  over  $S$  if and only if the random variable

$$\Delta(U) = \left\| \Pi_U^\perp \left( X_{S \setminus U} \beta_{S \setminus U}^* + W \right) \right\|^2 - \left\| \Pi_S^\perp W \right\|^2 \quad (12)$$

is negative. Overall, the optimal decoder fails if and only if at least one  $U$  (with cardinality  $|U| = s$ ) is preferable to  $S$ ; consequently, the probability of error can be written as

$$\mathbb{P}[\hat{S} \neq S] = \mathbb{P} \left[ \bigcup_{U \neq S, |U|=s} \{\Delta(U) < 0\} \right]. \quad (13)$$

In order to analyze this error probability, we begin by considering the range of possible integers  $k := |S \setminus U|$ , corresponding to the complement of the overlap. The following lemma makes use of known large-deviation bounds for  $\chi^2$  variates [14] to characterize the exponential decay rates of the random variable  $\Delta(U)$ :

**Lemma 1.** *For fixed  $k$  (with  $1 \leq k \leq s$ ), we have for any  $U$  with  $|S \setminus U| = k$ ,*

$$\begin{aligned} \mathbb{P}[\Delta(U) < 0] &\leq \exp \left\{ \frac{-(n-s) \|\beta_{S \setminus U}^*\|^2}{12 \left( \|\beta_{S \setminus U}^*\|^2 + 4 \right)} \right\} \\ &+ \exp \left\{ -\frac{k}{4} \left[ -1 + \frac{1}{4} (n-s) \frac{\|\beta_{S \setminus U}^*\|^2}{k} \right]^2 \right\}. \quad (14) \end{aligned}$$

**Weakened but simpler bound:** In order to make further progress, we simplify the bound (14) (though possibly weakening it) by noting that for all  $k \geq 1$ , we have  $\|\beta_{S \setminus U}^*\|^2 \geq k \mathcal{M}^2(\beta^*)$ , so that  $\mathbb{P}[\Delta(U) \leq 0]$  is upper bounded by

$$\exp \left\{ \frac{-(n-s)k \mathcal{M}^2(\beta^*)}{12(k \mathcal{M}^2(\beta^*) + 4)} \right\} + \exp \left\{ -\frac{k}{4} \left[ \frac{n-s}{4} \mathcal{M}^2(\beta^*) - 1 \right]^2 \right\}. \quad (15)$$

The advantage of this weakened bound is that it is independent of the subset  $U$ , and depends only on the parameter  $k = |S \setminus U|$ .

From this weakened bound (15), we see the necessity (at least for this analysis) of the requirement  $(n-s) \mathcal{M}^2(\beta^*) \rightarrow +\infty$ , so that the second error term decays asymptotically. Under this requirement, we have (for sufficiently large  $n$ ) that the second error exponent can be bounded as

$$\begin{aligned} -\frac{k}{4} \left[ \frac{n-s}{4} \mathcal{M}^2(\beta^*) - 1 \right]^2 &\leq -\frac{k}{12} \left[ \frac{n-s}{4} \mathcal{M}^2(\beta^*) - 1 \right] \\ &\leq \frac{-(n-s)k \mathcal{M}^2(\beta^*)}{12(k \mathcal{M}^2(\beta^*) + 8)}. \end{aligned}$$

The first error exponent is also upper bounded by this same quantity, so that we can simplify the upper bound to

$$\mathbb{P}[\Delta(U) \leq 0] \leq \exp \left\{ \frac{-(n-s)k \mathcal{M}^2(\beta^*)}{12(k \mathcal{M}^2(\beta^*) + 8)} \right\}. \quad (16)$$

Denote by  $N(k)$  the number of subsets  $U$  of size  $s$ , with overlap exactly equal to  $k$ . A standard counting argument yields that, for each  $k$  with  $1 \leq k \leq s$ , there are  $N(k) := \binom{s}{k} \binom{p-s}{k}$  such subsets. Using this simple bound (16) and union bound applied to the representation (13), we can upper bound the error probability as

$$\mathbb{P}[\hat{S} \neq S] \leq \sum_{k=1}^s \binom{s}{k} \binom{p-s}{k} \exp \left\{ \frac{-(n-s)k \mathcal{M}^2(\beta^*)}{12(k \mathcal{M}^2(\beta^*) + 8)} \right\}. \quad (17)$$

To complete the proof, we use the bound (17) to derive sufficient conditions for each of the terms in the summation (17) to vanish asymptotically. In order to deal with the binomial coefficients, we make use of the standard bounds

$$\log \binom{s}{k} \leq k \log \frac{se}{k}, \quad \text{and} \quad \log \binom{p-s}{k} \leq k \log \frac{(p-s)e}{k}.$$

Applying these two bounds, we conclude that the (logarithm of the)  $k^{\text{th}}$  term in the summation (17) is upper bounded by

$$k \left[ 2 + \log \frac{s}{k} + \log \frac{p-s}{k} \right] - \frac{(n-s)k \mathcal{M}^2(\beta^*)}{12(k \mathcal{M}^2(\beta^*) + 8)}.$$

Requiring this term to be negative asymptotically is equivalent to having

$$(n-s) \geq 12 \left( k + \frac{8}{\mathcal{M}^2(\beta^*)} \right) \left\{ 2 + \log \frac{s}{k} + \log \frac{p-s}{k} \right\}.$$

In order to understand the behavior of this lower bound, we consider  $k$  in two distinct regimes. On one hand, if  $k = \gamma s$  for some  $\gamma \in (0, 1)$ , then the second term on the RHS of the bound (18) is dominated by the term  $\log \frac{p-s}{\gamma s} =$

$\Omega(\log \frac{p}{s})$ , so that the overall lower bound is dominated by  $\max\{s, \mathcal{M}^{-2}(\beta^*)\} \log(p/s)$ . On the other hand, if  $k = o(s)$ , the lower bound is dominated by the maximum of linear growth  $s$ , and the quantity  $\mathcal{M}^{-2}(\beta^*) \log(p-s)$ . Overall, we conclude that

$$n > C \max \left\{ s \log(p/s), \frac{1}{\mathcal{M}^2(\beta^*)} \log(p-s) \right\}, \quad (18)$$

for some constant  $C > 0$  is sufficient for asymptotically reliable recovery, as claimed in Theorem 1.

### C. Proof of Theorem 2

We now turn to the proof of the necessary conditions given in Theorem 2. Our analysis is based on the following well-known lower bound [13] on the probability of error in a multiway hypothesis testing problem in terms of Kullback-Leibler divergences:

**Lemma 2.** *The average probability of error in performing in a hypothesis test over a family of distributions  $\{\mathbb{P}_1, \dots, \mathbb{P}_N\}$  is lower bounded as*

$$p_{\text{err}} \geq 1 - \frac{\frac{1}{N^2} \sum_{i,j=1}^N D(\mathbb{P}_i \parallel \mathbb{P}_j) + \log 2}{\log(N-1)},$$

where  $D(\mathbb{P}_i \parallel \mathbb{P}_j)$  denotes the Kullback-Leibler divergence.

Note that this bound is actually a weakened form of Fano's inequality [5], obtained by upper bounding the mutual information.

**Restricted problem:** Consider the collection of all  $N = \binom{p}{s}$  subsets of size  $s$  chosen from  $\{1, \dots, p\}$ . In order to produce lower bounds, we analyze the behavior of the optimal decoder for a restricted problem, in which we assume that for any fixed support  $S$ , it is known *a priori* that  $\beta_i^* = \mathcal{M}(\beta^*)$  for all indices  $i \in S$ . (Recall that  $\mathcal{M}(\beta^*)$  is the minimum absolute value of entries in the support of  $\beta^*$ .) Consequently, the optimal decoder for this modified problem is based on searching over all  $N$  subsets in our collection, seeking to minimize the quantity

$$g(U) := \|Y - X_U \vec{v}\|_2^2 = \|(X_S - X_U) \vec{v} + W\|_2^2,$$

where  $\vec{v} = \mathcal{M}(\beta^*) \vec{1}_s$  is a rescaled  $s$ -vector of ones.

Let us index the collection of all  $s$ -sized subsets with  $i = 1, 2, \dots, N(\delta)$ , and use  $U[i]$  to denote the corresponding support. For each index  $i$ , let  $\mathbb{P}_i$  denote the multivariate Gaussian distribution with mean  $X_{U[i]} \vec{v}$  and covariance matrix  $I_n$ . Note that the Kullback-Leibler divergence between any such pair is given by  $D(\mathbb{P}_i \parallel \mathbb{P}_j) = \frac{1}{2} \|X_{U[i]} \vec{v} - X_{U[j]} \vec{v}\|_2^2$ , so that the corresponding Fano bound takes the form

$$p_{\text{err}} \geq 1 - \frac{\frac{1}{N^2(\delta)} \sum_{i,j=1}^N \|X_{U[i]} \vec{v} - X_{U[j]} \vec{v}\|_2^2 + 2 \log 2}{\log[N-1]}.$$

**Upper bounds via concentration:** Thus, in order to ensure that  $p_e$  stays bounded away from zero, we need

to (upper) bound the quantity  $\frac{1}{2} \frac{1}{N^2} \sum_{i,j=1}^N \|X_{U[i]}\vec{v} - X_{U[j]}\vec{v}\|_2^2 / \log[N-1]$  away from one. For a given pair of subsets  $(U, V)$  in our collection, consider the random variable  $Z_{U,V} := \|X_U\vec{v} - X_V\vec{v}\|_2^2$ . A little calculation shows that  $Z_{U,V} \sim \gamma(U, V)\chi_n^2$ , where

$$\gamma(U, V) = 2\mathcal{M}^2(\beta^*) (s - |U \cap V|). \quad (19)$$

The following result bounds the upper tail behavior of the random variable  $Z = \frac{1}{N^2(\delta)} \sum_{U \neq V} Z_{U,V}$ , and follows from an application of Markov's inequality:

**Lemma 3.** *We have the tail bound  $\mathbb{P}[Z \geq 4\mathcal{M}^2(\beta^*)sn] \leq \frac{1}{2}$ .*

Using this lemma, we are guaranteed that at least 1/2 of Gaussian ensembles satisfy the upper bound

$$\frac{\frac{1}{2N^2} \sum_{i,j=1}^N D(\mathbb{P}_i \| \mathbb{P}_j)}{\log[N-1]} = \frac{\frac{1}{2N^2} \sum_{U \neq V} Z_{U,V}}{\log[N-1]} \leq \frac{4\mathcal{M}^2(\beta^*)sn}{\log[N-1]}.$$

Hence, as long as the RHS remains bounded from above away from one, the Fano bound implies that the probability of error averaged over the whole ensemble will remain bounded away from zero, which yields that  $n > \frac{\log[N-1]}{4\mathcal{M}^2(\beta^*)s}$  is necessary for asymptotically reliable recovery. To obtain a more transparent bound, we lower bound  $\log[N-1] = \log\binom{p}{s} - 1$  via

$$\log[N-1] \geq \frac{1}{2} \log N \geq \frac{1}{2} s \log \frac{p}{s},$$

where we have used standard lower bounds on binomial coefficients. Consequently, we obtain the necessary condition

$$n > \Omega\left(\frac{1}{s\mathcal{M}^2(\beta^*)} s \log \frac{p}{s}\right), \quad (20)$$

which concludes the proof of Theorem 2.

#### D. Concluding remarks

In this paper, we have analyzed the information-theoretic limits of the sparsity recovery problem for the linear observation model (2) with measurement vectors drawn from the standard Gaussian ensemble. We have established both lower and upper bounds on the number of observations  $n$  as a function of the model dimension  $p$  and sparsity index  $s$  that are required for asymptotically reliable recovery. There are a variety of open questions raised by our analysis. First, while our upper and lower bounds are essentially matching for certain regimes of scaling (e.g., sublinear sparsity with the minimum  $\mathcal{M}^2(\beta^*) = \Theta(1/s)$ ), it is likely that the analysis can be tightened in other regimes. In particular, the analysis of the necessary conditions (Theorem 2) is loose at several points, since it is based on analyzing a restricted ensemble, and using a weakened form of Fano's inequality. Second, one corollary of Theorem 1 is that with the sparsity index scaling linearly ( $s = \alpha p$  for some  $\alpha \in (0, 1)$ ), as long as the minimum value  $\mathcal{M}^2(\beta^*)$  decays sufficiently slowly, then asymptotically reliable recovery is possible with only a linear number of observations (i.e.,  $n = \beta p$  for some  $\beta > 0$ ). It remains to determine if there exists a computationally tractable method that approaches such performance in the linear sparsity regime.

#### Acknowledgements

This work was partially supported by NSF CAREER Award CCF-0545862, NSF Grant DMS-0605165, and an Alfred P. Sloan Foundation Fellowship. We thank Peter Bickel for helpful discussions and pointers.

#### REFERENCES

- [1] 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.
- [2] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory*, 51(12):4203–4215, December 2005.
- [3] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *Annals of Statistics*, 2006.
- [4] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.
- [5] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- [6] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, April 2006.
- [7] D. Donoho. For most large underdetermined systems of linear equations, the minimal  $\ell_1$ -norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, July 2006.
- [8] D. Donoho. For most large underdetermined systems of linear equations, the minimal  $\ell_1$ -norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.
- [9] D. L. Donoho and J. M. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. Technical report, Stanford University, 2006. Submitted to Journal of the AMS.
- [10] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 48(9):2558–2567, September 2002.
- [11] A. K. Fletcher, S. Rangan, and V. K. Goyal. Error bounds on sparse approximation. In *ICASSP*, April 2007.
- [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] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.
- [14] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, 1998.
- [15] D. M. Malioutov, M. Cetin, and A. S. Willsky. Optimal sparse representations in general overcomplete bases. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages II–793–796, May 2004.
- [16] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 2006. To appear.
- [17] 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.
- [18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [19] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030–1051, March 2006.
- [20] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programs. In *Proc. Allerton Conference on Communication, Control and Computing*, October 2006. Long version appeared as UC Berkeley Technical Report 709.
- [21] 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. Posted as arxiv:math.ST/0702301; To be presented at International Symposium on Information Theory, June 2007.
- [22] P. Zhao and B. Yu. Model selection with the lasso. Technical report, UC Berkeley, Department of Statistics, March 2006. Accepted to Journal of Machine Learning Research.