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

Martin J. Wainwright
Department of Statistics, and
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
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 β^*), 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, regardless of its computational complexity, must satisfy for perfect recovery. This analysis of optimal decoding limits complements our previous work [24] on sharp thresholds for sparsity recovery using the Lasso (ℓ_1-constrained quadratic programming) with Gaussian measurement ensembles.

Keywords: High-dimensional statistical inference; subset selection; signal denoising; compressive sensing; model selection; sparsity recovery; information-theoretic bounds; Fano's method.

1 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 β^* is sparse, meaning that its support set S—corresponding to those indices i for which β_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 [20], structure estimation in graphical models [19], sparse approximation [7, 21], signal denoising [5], and compressive sensing [8, 3].

A great deal of work over the past few years has focused on the performance of computationally tractable methods, many based on ℓ_1 or other convex relaxations, both for recovering the exact sparsity pattern as well as related problems in sparse approximation. We provide a brief overview of those parts of this extensive literature most relevant to our work in Section 1.1 below. 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—may tend to infinity. Our main results, stated more precisely in Section 1.2, are necessary and sufficient conditions on the triplet (n, p, s) for exact recovery. In particular, given noisy linear observations 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. The analysis given here complements our earlier paper [24] that established precise thresholds on the success/failure of the Lasso (i.e., ℓ_1-constrained quadratic programming) for sparsity recovery.

The remainder of this paper is organized as follows. In Section 1.1, we provide a more precise formulation of the problem, and a brief discussion of past work, whereas Section 1.2 provides a precise statement of our main results, and a discussion of their consequences. Section 2 and the appendices are devoted to the proofs of our main results, and we conclude in Section 3 with a discussion of open directions.

1.1   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 β^* as

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

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

$$Y_i \quad = \quad x_i^T \beta^* + W_i, \qquad i = 1, \ldots, n \tag{2}$$

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.

2

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: Consider some method that generates the vector $\widehat{\beta} \in \mathbb{R}^p$ as an estimate of the truth β^*. There are various distinct criteria for assessing how close the estimate is to the truth, including

- various ℓ_p norms $\mathbb{E}\|\widehat{\beta} - \beta^*\|_p$, especially ℓ_2 and ℓ_1, or

- some measurement of predictive power (e.g., $\mathbb{E}[\|Y_i - \widehat{Y}_i\|_2^2]$, where $\widehat{Y}_i$ is the estimate based on $\widehat{\beta}$).

Given the abundance of recent results on sparse approximation (not all of which are mutually comparable), it is particularly important to specify up front the choice of error metric. In this paper, we focus exclusively on the sparsity recovery problem, for which the appropriate error metric is simply the $0 - 1$ loss associated with the event of recovering the correct support S—viz.:

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

Past work: Closely related in its information-theoretic spirit is the earlier paper of Fletcher et al. [14] that analyzed the standard Gaussian ensemble from a rate-distortion perspective, studying the average ℓ_2-error of the optimal decoder. The results given here also address the information-theoretic limitations, albeit of the sparsity recovery problem, using the error metric (3) as opposed to ℓ_2-norm. In a related but distinct line of work, the use of ℓ_1-relaxation for sparse approximation has a lengthy history; relatively early papers from the 1990s include the work of Chen, Donoho and Saunders [5], as well as Tibshirani [22] on ℓ_1-constrained quadratic programming (known as the Lasso in the statistics literature). A great deal of subsequent work has analyzed the performance of ℓ_1-relaxations, both in the noiseless [12, 13, 18] and noisy setting [23] for deterministic ensembles, as well as the noiseless [10, 3, 11] and noisy setting [4, 2, 9, 19, 27, 24] for random ensembles. Other work has provided conditions under which estimation of a noise-contaminated vector via the Lasso [2, 9] or other types of convex relaxation [4] is stable in the ℓ_2 sense; however, such ℓ_2-stability does not guarantee exact recovery of the underlying sparsity pattern.

A notable feature of the results given here is that they apply to completely general scaling of the triplet (n, p, s). In contrast, most previous work has addressed one of two possible special cases of sparsity scaling: (a) either the *linear sparsity regime* [e.g 3, 10, 9], in which $s = \alpha p$ for some $\alpha \in (0, 1)$; or (b) the *sublinear sparsity regime* [e.g., 19, 27], 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 [24] that provided threshold results, also 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.

1.2 Our contributions

The analysis of this paper procedure is asymptotic in nature, focusing on scaling conditions on the triplet (n, p, s) under which asymptotically exact recovery is either possible or impossible. As mentioned previously, we focus on the linear observation model (2) in the noisy setting ($\sigma^2 > 0$), and with the measurement vectors x_i drawn in an i.i.d. manner from the standard Gaussian $N(0, I_{p \times p})$ ensemble. A decoder is a mapping from the n-vector of observations Y to an estimated subset—say of the form $\widehat{S} = \phi(Y)$. 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_{err} of any decoder is given by

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

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 β^*_S on the fixed support S, that the decoder makes an error. We say that

- the sparsity recovery is *asymptotically reliable* (error-free) if $p_{\mathrm{err}}(\phi) \to 0$ as $n \to +\infty$, and

- the sparsity recovery is *asymptotically unreliable* if for some constant $c > 0$, the error probability stays bounded $p_{\mathrm{err}}(\phi) \geq c$ as $n \to +\infty$.

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

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

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

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

The proof of this claim, given in Section 2.2, 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 $C' > 0$:*

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

The proof of this claim, given in Section 2.3, is somewhat more indirect in nature, based on exploiting a corollary of Fano's inequality [6, 15, 16, 26], in order to lower bound the probability of 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 $s = o(p)$. 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 $\log \frac{p}{s} = \Omega(\log(p - s)) - o(1)$, 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 [24] on computationally tractable methods, we established that ℓ_1-constrained quadratic programming (Lasso) has a threshold[1] for success/failure of order $n = \Theta\left(s \log(p - s)\right)$, so that the Lasso essentially 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 $\mathcal{M}^2(\beta^*)s \to +\infty$, then $n = \Theta(p)$ observations are sufficient to ensure asymptotically reliable recovery. This information-theoretic condition should be compared with our earlier analysis [24] of ℓ_1-constrained quadratic programming (the Lasso); one consequence of this work is 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 that $2s \log(p - s) \gg \Theta(p)$ for linear sparsity $s = \alpha p$, we see that there is a substantial gap between the performance of the Lasso and the optimal decoder in the linear sparsity regime. Thus, Theorem 1 raises the interesting question as to the existence of computationally efficient techniques for asymptotically reliable recovery in the regime of linear sparsity.

2 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.

2.1 Notation and set-up

For compactness in notation, let us use X to denote the $n \times p$ matrix formed with the vectors $x_k = (x_{k1}, x_{k2}, \ldots, x_{kp}) \in \mathbb{R}^p$ as rows, and the vectors $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T \in \mathbb{R}^n$ as columns, as follows:

$$X \quad := \quad \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \quad = \quad \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}. \tag{7}$$

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:

$$Y \quad = \quad X\beta^* + W. \tag{8}$$

---

[1] Those results [24] allowed the minimum value to scale as $\mathcal{M}^2(\beta^*) = f(s)/s$, where f is any function such that $\lim_{s \to +\infty} f(s) = +\infty$.

Given any subset $V \subseteq \{1, \ldots, p\}$, we use the notation β_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 σ can be accounted for in the scaling of β^*, via the parameter $\mathcal{M}(\beta^*)$.)

In addition, we use the following standard notation for asymptotics of real sequences $\{a_n\}$ and $\{b_n\}$: (i) $a_n = \mathcal{O}(b_n)$ means that $a_n \leq C b_n$ for some constant $C \in (0, \infty)$; (ii) $a_n = \Omega(b_n)$ means that $a_n \geq C' b_n$ for some constant $C' \in (0, \infty)$; (iii) $a_n = \Theta(b_n)$ is shorthand for $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$, and (iv) $a_n = o(b_n)$ means that $a_n / b_n \to 0$.

2.2 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 $U \subset \{1, \ldots, p\}$, as

$$f(U; Y, X, \beta^*) \;=\; \arg\min_{\beta_U} \left\{ \|Y - X_U \beta_U\|_2^2 \right\}. \tag{9}$$

We frequently write $f(U)$ as a shorthand; note that this value corresponds to the error associated with the best estimator of Y that lies in $\text{Ra}(X_U)$. The optimal decoder chooses the best subset $\widehat{S}$ based on the minimal value of this error, ranging over all subsets U of size s:

$$\widehat{S} \;=\; \phi_{\text{opt}}(Y) \;:=\; \arg\min_{|U|=s} f(U; Y, X, \beta^*). \tag{10}$$

Note that by symmetry, the error probability $\mathbb{P}[\widehat{S} \neq S \mid S]$ 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

$$p_{\text{err}}(\phi) \;=\; \mathbb{P}[\phi(Y) \neq S], \tag{11}$$

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

$$\Pi_U \;:=\; X_U \left[X_U^T X_U \right]^{-1} X_U^T, \qquad \text{and} \tag{12a}$$

$$\Pi_U^\perp \;:=\; I_{n \times n} - X_U \left[X_U^T X_U \right]^{-1} X_U^T. \tag{12b}$$

Note that Π_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, we state the following result (see Appendix A for a proof):

Lemma 1. *For a given vector β^* with support S, the optimal decoder declares U over S if and only if the random variable*

$$\Delta(U) \;=\; \left\| \Pi_U^\perp \left(X_{S \setminus U} \beta_{S \setminus U}^* + W \right) \right\|^2 - \left\| \Pi_S^\perp W \right\|^2. \tag{13}$$

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

$$\mathbb{P}[\widehat{S} \neq S] = \mathbb{P}\Big[\bigcup_{U \neq S, \, |U| = s} \{\Delta(U) < 0\} \Big]. \tag{14}$$

In order to analyze this error probability, we begin by considering the range of possible integers $k := |S \backslash U|$, corresponding to the complement of the overlap. The following lemma characterizes the exponential decay rates of the random variable $\Delta(U)$:

Lemma 2. *For fixed k (with $1 \leq k \leq s$), we have for any U with $|S \backslash U| = k$,*

$$\mathbb{P}[\Delta(U) < 0] \leq \exp\left\{ \frac{-(n-s)\|\beta^*_{S \backslash U}\|^2}{12 \left(\|\beta^*_{S \backslash U}\|^2 + 4 \right)} \right\} + 2\exp\left\{ -\frac{k}{4}\left[-1 + \frac{1}{4}(n-s)\frac{\|\beta^*_{S \backslash U}\|^2}{k} \right]^2 \right\}. \tag{15}$$

Proof. We begin by conditioning on the Gaussian noise vector W. Since each element of $X_{S \backslash U}$ is standard normal, each entry of the random vector $X_{S \backslash U}\beta^*_{S \backslash U}$ is zero-mean Gaussian with variance $\|\beta^*_{S \backslash U}\|^2$. Consequently, if we rescale by the standard deviation, then the random vector

$$\|\beta^*_{S \backslash U}\|^{-1} \left(X_{S \backslash U}\beta^*_{S \backslash U} + W \right)$$

is an n-dimensional Gaussian random vector with independent entries, each with with unit variance, and mean vector W. Applying the orthogonal transform $\Pi_U^\perp$ reduces the number of degrees of freedom to $(n - s)$, so that we conclude that

$$\|\beta^*_{S \backslash U}\|^{-2} \left\| \Pi_U^\perp \left(X_{S \backslash U}\beta^*_{S \backslash U} + W \right) \right\|^2$$

is a non-central χ^2 variate with $d = n - s$ degrees of freedom, and non-centrality parameter $\nu = \|\beta^*_{S \backslash U}\|^{-2}\|\Pi_U^\perp W\|^2$. With these choices of (d, ν), we have

$$\mathbb{P}[\Delta(U) < 0 \mid W] = \mathbb{P}\left[\chi^2(d, \nu) < t \right]$$

where we have set $t := \frac{\|\Pi_S^\perp W\|^2}{\|\beta^*_{S \backslash U}\|^2}$ for shorthand. Thus, conditioned on W, our problem reduces to bounding the tail of a non-central χ^2 variate. In Appendix D, we state some known tail bounds [1] on such variates, which we use here. In order to apply these bounds, we condition on the following "good event", defined in terms of W

$$\mathcal{A} = \left\{ \left| \frac{\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2}{\|\beta^*_{S \backslash U}\|^2} \right| \leq \frac{n-s}{2} \right\} \bigcap \left\{ \|\Pi_U^\perp W\|^2 \leq 2(n-s) \right\}.$$

Note that the first event defining $\mathcal{A}$ ensures that

$$d + \nu - t = (n-s) + \frac{1}{\|\beta^*_{S \backslash U}\|^2}\left(\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2 \right) \geq \frac{n-s}{2} \geq 0. \tag{16}$$

7

Consequently, conditioned on $\mathcal{A}$, we may set $x := \frac{(d+\nu-t)^2}{4(d+2\nu)}$ in equation (35b) to obtain the upper bound

$$
\begin{aligned}
\log \mathbb{P}[\Delta(U) < 0 \mid \mathcal{A}] \;\leq\;& -\frac{(d+\nu-t)^2}{4(d+2\nu)} \\[2ex]
=\;& -\frac{\left([n-s] + \frac{\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2}{\|\beta_{S\setminus U}^*\|^2}\right)^2}{4\left([n-s] + 2\frac{\|\Pi_U^\perp W\|^2}{\|\beta_{S\setminus U}^*\|^2}\right)} \\[2ex]
=\;& -(n-s)\,\frac{\left(1 + \frac{\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2}{(n-s)\,\|\beta_{S\setminus U}^*\|^2}\right)^2}{4\left(1 + 2\frac{\|\Pi_U^\perp W\|^2}{(n-s)\,\|\beta_{S\setminus U}^*\|^2}\right)} \\[2ex]
\overset{(b)}{\leq}\;& -(n-s)\,\frac{1/2}{4\left(1 + 4/\|\beta_{S\setminus U}^*\|^2\right)} \\[2ex]
=\;& -(n-s)\,\frac{\|\beta_{S\setminus U}^*\|^2}{8\left(\|\beta_{S\setminus U}^*\|^2 + 4\right)},
\end{aligned}
\tag{17}
$$

where inequality (b) makes use of the second event defining $\mathcal{A}$.

We complete the proof by observing that

$$
\mathbb{P}[\Delta(U) < 0] \;\leq\; \mathbb{P}[\Delta(U) < 0 \mid \mathcal{A}] + \mathbb{P}[\mathcal{A}^c], \tag{18}
$$

so that it suffices to upper bound $\mathbb{P}[\mathcal{A}^c]$. By union bound, we have

$$
\mathbb{P}[\mathcal{A}^c] \;\leq\; \mathbb{P}\left[\left|\frac{\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2}{\|\beta_{S\setminus U}^*\|^2}\right| \geq \frac{n-s}{2}\right] + \mathbb{P}\left[\|\Pi_U^\perp W\|^2 \geq 2(n-s)\right]. \tag{19}
$$

Since $\|\Pi_U^\perp W\|^2$ is a central χ^2 with $(n-s)$ degrees of freedom, we may apply the tail bounds from Appendix D to conclude that

$$
\mathbb{P}\left[\|\Pi_U^\perp W\|^2 \geq 2(n-s)\right] \;\leq\; \exp(-(n-s)/12). \tag{20}
$$

Turning to the first term on the RHS on equation (19), we observe that

$$
\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2 \;=\; \|\Pi_U W\|^2 - \|\Pi_S W\|^2 \overset{d}{=} \sum_{i \in U\setminus S} Z_i^2 - \sum_{j \in S\setminus U} Z_j^2,
$$

where $\{Z_i, Z_j\}$ are i.i.d. standard normal variates. Now if the difference $\sum_{i\in U\setminus S} Z_i^2 - \sum_{j\in S\setminus U} Z_j^2$ is to exceed $\frac{1}{2}(n-s)\|\beta_{S\setminus U}^*\|^2$, then at least one of the terms must exceed $\frac{1}{4}(n-s)\|\beta_{S\setminus U}^*\|^2$. Moreover,

we observe that $\sum_{j \in S \backslash U} Z_j^2$ is χ_k^2, where $k = |S \backslash U|$. Hence, we have

$$
\begin{aligned}
\log \mathbb{P} \left[\left| \frac{\|\Pi_U^\perp W\|^2 - \|\Pi_S^\perp W\|^2}{\|\beta_{S \backslash U}^*\|^2} \right| \geq \frac{n-s}{2} \right] & \leq \log 2\, \mathbb{P} \left[\frac{\chi_k^2}{k} \geq \frac{1}{4}(n-s) \frac{\|\beta_{S \backslash U}^*\|^2}{k} \right] \\
& = \log 2 \mathbb{P} \left[\chi_k^2 - k \geq k \left\{ -1 + \frac{1}{4}(n-s) \frac{\|\beta_{S \backslash U}^*\|^2}{k} \right\} \right] \\
& \leq -\frac{k}{4} \left[-1 + \frac{1}{4}(n-s) \frac{\|\beta_{S \backslash U}^*\|^2}{k} \right]^2 + \log 2,
\end{aligned}
$$

where we have used the upper bound (34a) from Appendix D with $x := \frac{k}{4} \left(-1 + \frac{1}{4}(n-s) \frac{\|\beta_{S \backslash U}^*\|^2}{k} \right)^2$ in the final inequality.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

Weakened but simpler bound: In order to make further progress, we simplify the bound (15) from Lemma 2, at the expense of weakening it, by noting that for all $k \geq 1$, we have $\|\beta_{S \backslash U}^*\|^2 \geq k \mathcal{M}^2(\beta^*)$, so that

$$
\mathbb{P}[\Delta(U) \leq 0] \leq \exp \left\{ \frac{-(n-s)k \, \mathcal{M}^2(\beta^*)}{12 \left(k \, \mathcal{M}^2(\beta^*) + 4 \right)} \right\} + 2 \exp \left\{ -\frac{k}{4} \left[\frac{n-s}{4} \mathcal{M}^2(\beta^*) - 1 \right]^2 \right\}. \quad (21)
$$

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

From this weakened bound (21), we see the necessity (at least for this analysis) of the requirement $(n-s) \, \mathcal{M}^2(\beta^*) \to +\infty$, 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

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

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

$$
\mathbb{P}[\Delta(U) \leq 0] \leq 3 \exp \left\{ \frac{-(n-s)k \, \mathcal{M}^2(\beta^*)}{12 \left(k \, \mathcal{M}^2(\beta^*) + 8 \right)} \right\}. \quad (22)
$$

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

$$
N(k) = \binom{s}{k} \binom{p-s}{k} \quad (23)
$$

9

such subsets. Using this simple bound (22) and union bound applied to the representation (14), we can upper bound the error probability as

$$\mathbb{P}[\widehat{S} \neq S] \;\leq\; 3 \sum_{k=1}^{s} \binom{s}{k} \binom{p-s}{k} \, \exp \left\{ \frac{-(n-s)k \, \mathcal{M}^2(\beta^*)}{12 \, (k \, \mathcal{M}^2(\beta^*) + 8)} \right\}. \tag{24}$$

Analysis of the upper bound: We now analyze the upper bound (24); in particular, our goal is to derive sufficient conditions for each of the terms in the summation to vanish asymptotically. In order to deal with the binomial coefficients, we make use of the bounds (see Appendix C)

$$\log \binom{s}{k} \leq k \log \frac{se}{k}, \qquad \text{and} \qquad \log \binom{p-s}{k} \leq k \log \frac{(p-s)e}{k}. \tag{25}$$

Applying these two bounds, we conclude that the (logarithm of the) k^{th} term 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

$$\begin{aligned}
(n-s) \;\geq\; & \frac{12 \left(k\mathcal{M}^2(\beta^*) + 8 \right)}{k \, \mathcal{M}^2(\beta^*)} \, k \left[2 + \log \frac{s}{k} + \log \frac{p-s}{k} \right] \\
= \; & 12 \left(k + \frac{8}{\mathcal{M}^2(\beta^*)} \right) \left\{ 2 + \log \frac{s}{k} + \log \frac{p-s}{k} \right\}.
\end{aligned} \tag{26}$$

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 (26) 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 the condition

$$n \;>\; C \, \max \left\{ s \log(p/s), \; \frac{1}{\mathcal{M}^2(\beta^*)} \log(p-s) \right\}, \tag{27}$$

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

2.3 Proof of Theorem 2

We now turn to the proof of the necessary conditions given in Theorem 2.

Fano method: Our analysis is based on a well-known lower bound on the probability of error in a multiway hypothesis testing problem in terms of Kullback-Leibler divergences. In the non-parametric statistics literature [15, 16, 26], this approach is referred to as the Fano method, since the bound is a corollary of Fano's inequality from information theory [6]. Here we state and make use of the following variant [25]:

Lemma 3. *Consider a family of N distributions $\{\mathbb{P}_1, \ldots, \mathbb{P}_N\}$. Then the average probability of error in performing in a hypothesis test over this family is lower bounded as*

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

where $D(\mathbb{P}_i \,\|\, \mathbb{P}_j)$ denotes the Kullback-Leibler divergence between distributions $\mathbb{P}_i$ and $\mathbb{P}_j$.

Restricted problem: Consider the collection of all $N = \binom{p}{s}$ subsets of size s chosen from $\{1, \ldots, p\}$. 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 β^*.) This problem is simply an N-way hypothesis testing problem, in which the observation under the hypothesis associated with subset U takes the form

$$Y = X_U \vec{v} + W, \tag{28}$$

where $\vec{v} = \mathcal{M}(\beta^*) \vec{1}_s$ is a rescaled s-vector of ones, and $W \sim N(0, I_{n \times n})$.

Let us index the collection of all s-sized subsets with $i = 1, 2, \ldots, N$, 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 $I_{n \times n}$; note that $\mathbb{P}_i$ is simply the class-conditional distribution of Y under the hypothesis $U[i]$. Moreover, the Kullback-Leibler divergence between any such pair is given by $D(\mathbb{P}_i \,\|\, \mathbb{P}_j) = \frac{1}{2} \|X_{U[i]} \vec{v} - X_{U[j]} \vec{v}\|_2^2$, so that the corresponding Fano bound takes the form

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

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

$$\gamma(U, V) = 2 \mathcal{M}^2(\beta^*) \, (s - |U \cap V|). \tag{29}$$

The following result bounds the upper tail behavior of the random variable $Z = \frac{1}{N^2} \sum_{U \neq V} Z_{U,V}$.

Lemma 4. *The tail of Z obeys the bound*

$$\mathbb{P}\left[Z \geq 4\mathcal{M}^2(\beta^*) sn\right] \leq \frac{1}{2}.$$

11

Using this lemma (see Appendix B for a proof of this claim), we are guaranteed that at least $1/2$ of the Gaussian ensembles satisfy the upper bound

$$\frac{1}{2} \frac{\frac{1}{N^2} \sum_{i,j=1}^{N} D(\mathbb{P}_i \,\|\, \mathbb{P}_j)}{\log[N-1]} \;=\; \frac{1}{2} \frac{\frac{1}{N^2} \sum_{U \neq V} Z_{U,V}}{\log[N-1]} \;\leq\; \frac{4\mathcal{M}^2(\beta^*)sn}{\log[N-1]}. \tag{30}$$

Hence, as long as the quantity (30) 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. Consequently, we obtain the necessary condition that

$$n \;>\; \frac{\log[N-1]}{4\mathcal{M}^2(\beta^*)s}$$

for reliable recovery with probability one asymptotically. To obtain a more transparent bound, we first lower bound N via $\log[N-1] \geq \frac{1}{2}\log N$, and then further via

$$\frac{1}{2}\log N \;=\; \frac{1}{2}\log\binom{p}{s} \;\geq\; \frac{1}{2}s\,\log\frac{p}{s},$$

as stated in Appendix C. Consequently, we obtain the necessary condition

$$n > \Omega\left(\frac{1}{s\,\mathcal{M}^2(\beta^*)}s\log\frac{p}{s}\right), \tag{31}$$

as stated in Theorem 2.

3 Conclusion

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 (see proof of Theorem 2) involves some slack since it is based on analyzing a very restricted ensemble. Second, our results (in particular, a corollary of Theorem 1) reveal 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 $\beta > 0$). Since our previous work [24] established that the Lasso (ℓ_1-constrained quadratic programming) cannot achieve reliable recovery in this particular (n, p, s) regime, it remains to determine a computationally tractable method that approaches such performance in the regime of linear sparsity. Third, whereas the current analysis has focused on a very special class of Gaussian ensemble, the analysis given here could be extended to a broader class of measurement ensembles.

A Proof of Lemma 1

We begin by showing that for any subset U for which X_U is full rank, the function f has the equivalent form $f(U) = \|\Pi_U^\perp Y\|_2^2$. Under the given rank condition, the linear least squares estimator of β_U^* is given by $\widehat{\beta}_U = \left[X_U^T X_U\right]^{-1} X_U^T Y$. Noting that $X_U \widehat{\beta}_U = \Pi_U Y$, we substitute into the quadratic norm and expand, thereby obtaining

$$f(U) \;=\; \|Y - X_U \widehat{\beta}_U\|_2^2 \;=\; \|(I - \Pi_U)Y\|_2^2 \;=\; \|\Pi_U^\perp Y\|_2^2$$

as claimed. Lastly, to establish equation (13), we note that

$$f(U) \;=\; \|\Pi_U^\perp \left(X_S \beta_S^* + W\right)\|_2^2 \;=\; \|\Pi_U^\perp \left(X_{S\setminus U}\beta_S^* {\setminus U} + W\right)\|_2^2,$$

since $\Pi_U^\perp v = 0$ for any vector v belonging to the range of X_U.

B Proof of Lemma 4

Note that $Z = \frac{1}{N^2} \sum_{U,V} Z_{U,V}$ is a rescaled sum of a total number N^2 variables (neither independent nor identically distributed). However, since Z is a non-negative random variable, we may apply Markov's inequality for any $t > 0$ to conclude that

$$\mathbb{P}\left[Z \geq t\right] \;\leq\; \frac{\mathbb{E}[Z]}{t}. \tag{32}$$

Since each $Z_{U,V}$ has distribution $\gamma(U,V)\chi_n^2$, we have $\mathbb{E}[Z_{U,V}] = \gamma(U,V)n$. From equation (29), we note that $\gamma(U,V) \leq 2\mathcal{M}^2(\beta^*)s$, and hence

$$\mathbb{E}[Z] \;\leq\; \max_{U \neq V}\left(\gamma(U,V)\right)n \;=\; 2\mathcal{M}^2(\beta^*)sn,$$

Hence setting $t = 4\mathcal{M}^2(\beta^*)sn$ in the bound (32) yields the claim.

C Bounds on binomial coefficients

Although more refined results are certainly possible, we make frequent use of the following crude bounds on the binomial coefficients

$$\left(\frac{n}{k}\right)^k \;\leq\; \binom{n}{k} \;\leq\; \left(\frac{n\,e}{k}\right)^k. \tag{33}$$

D   Tail bounds for chi-square variables

The following large-deviations bounds for centralized χ^2 are taken from Laurent and Massart [17]. Given a centralized χ^2-variate X with d degrees of freedom, then for all $x \geq 0$,

$$\mathbb{P}\left[X - d \geq 2\sqrt{dx} + 2x\right] \leq \mathbb{P}\left[X - d \geq 2\sqrt{dx}\right] \leq \exp(-x), \qquad \text{and} \qquad (34\text{a})$$

$$\mathbb{P}\left[X - d \leq -2\sqrt{dx}\right] \leq \exp(-x). \qquad (34\text{b})$$

More generally, the analogous tail bounds for *non-central* χ^2, taken from Birgé [1], can be established via the Chernoff bound. Let X be a non-central χ^2 variable with d degrees of freedom and non-centrality parameter $\nu \geq 0$. Then for all $x > 0$,

$$\mathbb{P}\left[X \geq (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x\right] \leq \exp(-x), \qquad \text{and} \qquad (35\text{a})$$

$$\mathbb{P}\left[X \leq (d+\nu) - 2\sqrt{(d+2\nu)x}\right] \leq \exp(-x). \qquad (35\text{b})$$

References

[1] 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.

[2] 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.

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

[4] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

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

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

[7] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer-Verlag, New York, NY, 1993.

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

[9] D. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, July 2006.

[10] D. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, June 2006.

[11] D. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via ℓ_1 minimization. In *Proc. Conf. on Inofrmation Sciences and Systems*, March 2006.

[12] 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.

[13] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Trans. Info Theory*, 49(6):1579–1581, 2003.

[14] 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.

[15] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[16] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

[17] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, 1998.

[18] 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.

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

[20] A. J. Miller. *Subset selection in regression*. Chapman-Hall, New York, NY, 1990.

[21] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 24(2):227–234, 1995.

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

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

[24] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs. In *Proc. Allerton Conference on Communication, Control and Computing*, October 2006. Long version appeared as UC Berkeley Technical Report 709.

[25] Y. G. Yatracos. A lower bound on the error in non-parametric regression type problems. *Annals of Statistics*, 16(3):1180–1187, 1988.

[26] B. Yu. Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, Berlin, 1997.

[27] 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.