Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting

Martin J. Wainwright, *Member, IEEE*

***Abstract*—The problem of sparsity pattern or support set recovery refers to estimating the set of nonzero coefficients of an unknown vector $\beta^* \in \mathbb{R}^p$ based on a set of n noisy observations. It arises in a variety of settings, including subset selection in regression, graphical model selection, signal denoising, compressive sensing, and constructive approximation. The sample complexity of a given method for subset recovery refers to the scaling of the required sample size n as a function of the signal dimension p, sparsity index k (number of non-zeroes in β^*), as well as the minimum value $\beta_{\min}$ of β^* over its support and other parameters of measurement matrix. This paper studies the information-theoretic limits of sparsity recovery: in particular, for a noisy linear observation model based on random measurement matrices drawn from general Gaussian measurement matrices, we derive both a set of sufficient conditions for exact support recovery using an exhaustive search decoder, as well as a set of necessary conditions that any decoder, regardless of its computational complexity, must satisfy for exact support recovery. This analysis of fundamental limits complements our previous work on sharp thresholds for support set recovery over the same set of random measurement ensembles using the polynomial-time Lasso method (ℓ_1-constrained quadratic programming).**

***Index Terms*—Compressed sensing, ℓ_1-relaxation, Fano's method, high-dimensional statistical inference, information-theoretic bounds, Lasso, model selection, signal denoising, sparsity pattern, sparsity recovery, subset selection, support recovery.**

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 β^* is sparse, meaning that its support set S—corresponding to those indices i for which ${\beta^*}_i$ is nonzero—is relatively small, say with size $|S| =: k \ll p$. 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 [24], structure estimation in graphical models [22], sparse approximation [10], [25], signal denoising [7], and compressive sensing [11], [5].

Manuscript received August 28, 2007; revised April 20, 2009. Current version published November 20, 2009. This work was supported in part by the National Science Foundation under Grants NSF DMS-0528488, CAREER-CCF-0545862, a Microsoft Research Grant, and a Sloan Foundation Fellowship. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Nice, France, June 2007, and was posted on arXiv in February 2007 (math/0702301).

The author is with the Department of Statistics, and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720 USA.

Communicated by A. Krzyżak, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2009.2032816

A great deal of work over the past few years has focused on the performance of computationally tractable methods, many based on ℓ_1-norm 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 I-A. 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 observationsn*, the *model dimension* p, and the *sparsity index* k, defined below—may tend to infinity. Our main results, stated more precisely in Section II, are necessary and sufficient conditions for subset recovery, stated in terms of the triplet (n, p, k) as well as signal-to-noise parameters such as the minimum value $\beta_{\min}$ of the signal $\beta^* \in \mathbb{R}^p$ and the noise variance σ^2. More specifically, our analysis applies to the class of random Σ-Gaussian measurement ensembles, in which each measurement is based on the inner product between β^* and a random p-vector $X_i \sim N(0, \Sigma)$. As a special but important case, this model includes the standard Gaussian ensemble in which $X_{ij} \sim N(0,1)$ are independent and identically distributed (i.i.d.), obtained by setting $\Sigma = I_{p \times p}$. In this paper, we derive a set of sufficient conditions for asymptotically perfect recovery using an exhaustive search 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 [31] 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 I-A, we provide a more precise formulation of the problem, and a brief discussion of past work, whereas Section II provides a precise statement of our main results, and a discussion of

their consequences. Section IV provides the proof of the sufficient conditions, based on analyzing the oracle decoder, whereas Section V provides the proof of the necessary conditions. More technical aspects of these proofs are provided in the Appendices. We conclude in Section VI with a discussion of open directions.

A. Problem Formulation

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

$$S := \{i \in \{1, \ldots, p\} \,|\, {\beta^*}_i \neq 0\}. \quad (1)$$

We refer to its size $k := |S|$ as the *sparsity index*. Consider the observation model

$$y = X\beta^* + w \quad (2)$$

where $y \in \mathbb{R}^n$ is a vector of observations, $X \in \mathbb{R}^{n \times p}$ is the measurement matrix, and $w \in \mathbb{R}^n$ is additive observation noise. We assume throughout the paper that $w \sim N(0, \sigma^2 I_{n \times n})$.

1) Error Metrics: Consider some method that generates the vector $\hat{\beta} \in \mathbb{R}^p$ 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_q norms $\mathbb{E}\|\hat{\beta} - \beta^*\|_q^q$, especially ℓ_2 and ℓ_1, or
- some measurement of predictive power (e.g., $\mathbb{E}\|y - \hat{y}\|_2^2$, where $\hat{y}$ 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 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.

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

Of interest are conditions on the triplet (n, p, k) as well as properties of the signal vector β^* and design matrix X under which exact support recovery is either possible, or impossible.

2) Past and Ongoing Work: A great deal of recent work has studied the behavior of ℓ_1-relaxations for sparse approximation, including linear programming techniques [7], [12], [5], [14] and ℓ_1-constrained quadratic programming [7], [13], [29], known as the Lasso in the statistics literature [22], [28], [36]. Some papers in this growing literature have provided conditions under which estimation of a noise-contaminated vector via the Lasso [29], [13] or other types of convex relaxation [6] is guaranteed to be stable in the ℓ_2 sense; however, it should be noted that such ℓ_2-stability does not guarantee exact recovery of the underlying support set. Most directly related to this paper are results, applicable to ℓ_1-constrained quadratic programming or the Lasso, that provide sufficient conditions [22], [36], [31] or necessary conditions [31] on the amount of data required for subset recovery (i.e., with the error metric (3)). These results isolate a mutual incoherence property [17], [29] of the design matrix that must be satisfied for the Lasso to succeed in recovering the support, and the paper [31] provides sharp scalings on (n, p, k) that demarcate the boundary between success and failure. As we discuss in the sequel, our results show that the exhaustive search decoder can recover the support set for design matrices for which the ℓ_1-based Lasso fails with high probability (see Section III-B), or for sample sizes in which the Lasso fails with high probability (see Section III-A.2).

Some past work on sparse linear regression [1], [27], [16] shares the information-theoretic motivation of this paper, but focuses on the rate–distortion perspective (i.e., under the ℓ_2-loss), as opposed to the subset recovery metric (3) of interest here. Since this paper was first posted [30], a number of papers have followed up on the information-theoretic limits of the subset recovery problem. Akcaya and Tarokh [2] analyzed the performance of a certain type of "joint typicality decoder," obtaining similar results for the support recovery problem studied here as well as various results for partial support recovery metrics (e.g., metrics in which it is sufficient to recover a large fraction of the support, as opposed to any element of the support). Their analysis is based on the same framework and type of partitioning scheme, but uses alternative large deviation bounds based on concentration of empirical entropies (joint typicality). Reeves and Gastpar [26] analyzed the partial support recovery problem in the regime of linear sparsity (i.e., $k = \alpha p$ for some $\alpha \in (0, 1)$), and showed that the signal-to-noise ratio (SNR) must tend to infinity in order for exhaustive search decoders to succeed. In subsequent work, Fletcher *et al.* [15] used direct methods to show that for any signal β^* with squared minimum value $\beta^2_{\min}$, any decoder applied to measurement matrices drawn from the standard Gaussian ensemble requires $n = \Omega\left(\frac{\log(p-k)}{\beta^2_{\min}}\right)$ measurements. Concurrent work by Wang *et al.* [32] used refinements of the Fano approach from the initial posting of this work [30], to establish the same scaling for general i.i.d. measurement matrices. Although these extensions did not appear in the original posting of this work [30], following the reviewers' suggestion, we have also included in this revised version some consequences of the refined Fano approach [32] for necessary conditions (Theorem 2) as applied to the non-i.i.d. Σ-Gaussian ensembles considered here.

Notation: 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 $a_n \leq Cb_n$ 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 $a_n/b_n \to 0$.

II. Statement of Main Results

The analysis of this paper applies to the high-dimensional setting, in that all three elements of the triplet (n, p, k) are permitted to tend to infinity. We provide both *positive results*—that is, scalings of the triplet (n, p, k) and associated signal/measurement parameters such that an exhaustive search decoder can recover the exact support with high probability—and also *converse results*, meaning scalings for which the probability of successful support recovery remains bounded away from zero for any method. Although we allow for completely general scaling of this triplet, our results can also be specialized to two particular cases of sparsity scaling: (a) the *linear sparsity regime* e.g.,

[5], [12], in which $k = \alpha p$ for some $\alpha \in (0,1)$; or (b) the *sublinear sparsity regime*, e.g., [22], [36], in which k/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 [31] that provided threshold results, also applicable to general scaling of (n,p,k), for the success/failure of the Lasso when used for sparsity recovery with general Gaussian measurement ensembles.

We focus on the linear observation model (2) in the noisy setting ($\sigma^2 > 0$), with the measurement matrix $X \in \mathbb{R}^{n \times p}$ drawn from the Σ-*Gaussian ensemble,* meaning that each row $X_i \in \mathbb{R}^p$ is drawn i.i.d. as $X_i \sim N(0,\Sigma)$ for $i = 1,2,\ldots,n$. Note that setting $\Sigma = I_{p\times p}$ yields as a special case the *standard Gaussian ensemble*, for which $X_{ij} \sim N(0,1)$ is i.i.d.

In addition to the three parameters (n,p,k), our results also highlight the importance of some other parameters associated with the signal ensemble. In particular, both the sufficient and necessary conditions require control of the *minimum value* of the unknown vector β^* on its support. Consequently, for a given minimum value $\beta_{\min}$, we consider the class of signals

$$\mathsf{U}(\beta_{\min}) = \{\beta^* \in \mathbb{R}^p \,|\, |\beta_i^*| \geq \beta_{\min} \text{ for all } i \in S(\beta^*)\} \quad (4)$$

where $S(\beta^*)$ is the support of β^*. Our results show that the SNR parameter $\beta_{\min}^2/\sigma^2$, as opposed to the more traditional measure $\|\beta^*\|_2^2/\sigma^2$ that would arise in assessing ℓ_2 error, is the key quantity that controls subset selection. Indeed, we show that the quantity $\|\beta^*\|_2^2/\sigma^2$ can be arbitrarily large without having any effect on the difficulty of subset recovery.

A. Decoders and Error Probabilities

A decoder ψ is a mapping from the pair $(y,X) \in \mathbb{R}^n \times \mathbb{R}^{n\times p}$ to the family

$$\mathfrak{S}_k := \{T \,|\, T \subseteq \{1,2,\ldots,p\},\ |T| = k\}$$

of all k-sized subsets of $\{1,2,\ldots,p\}$. The output $\widehat{S} = \psi(y,X)$ corresponds to the decoder's best estimate of the unknown underlying subset. The underlying true vector $\beta^* \in \mathbb{R}^p$ is fixed but unknown. We focus on three different types of error, depending on whether (a) the error probability is taken conditionally on a fixed support set S, or (b) the error probability is averaged over a support set S chosen uniformly at random (u.a.r.) from all $\binom{p}{k}$ possible k-sized subsets, or (c) the error probability is worst case over a support set S chosen in an adversarial manner. In particular, in the case that β^* has fixed but unknown support S, we define the S-based error

$$q_S(\psi) := \mathbb{P}[\psi(y,X) \neq S \,|\, S]. \quad (5)$$

Here the probability $\mathbb{P}[\psi(y,X) \neq S \,|\, S]$ is taken over the random measurement y, or equivalently, over the observation noise w and random design matrix X, with the underlying support being S, with the probability taken over the measurement noise w and the choice of random design matrix X. When S is viewed as a random variable, chosen u.a.r., we define the *average error probability*

$$q_{\text{ave}}(\psi) := \mathbb{E}_S[q_S(\psi)] = \frac{1}{\binom{p}{k}} \sum_{\{S \in \mathfrak{S}_k\}} q_S(\psi). \quad (6)$$

Finally, when S is chosen in an adversarial manner, we define the *worst case error probability*

$$q_{\max}(\psi) := \max_{\{S \in \mathfrak{S}_k\}} q_S(\psi). \quad (7)$$

B. Design Covariance Parameters

Our second set of parameters involve the covariance matrix Σ that defines the Σ-Gaussian ensemble of design matrices, in which each row of the design matrix $X \in \mathbb{R}^{n\times p}$ is drawn i.i.d. from the p-dimensional $N(0,\Sigma)$ normal distribution. We begin with the key quantities that arise in our sufficient conditions, as stated in Theorem 1. Given a pair $(S,T) \in \mathfrak{S}_k \times \mathfrak{S}_k$ with $S \neq T$, we define the $|S\backslash T| \times |S\backslash T|$ matrix

$$\Gamma(T,S) := \Sigma_{(S\backslash T)S\backslash T} - \Sigma_{(S\backslash T)T}(\Sigma_{TT})^{-1}\Sigma_{T(S\backslash T)}. \quad (8)$$

Note that $\Gamma(T,S)$ corresponds to the Schur complement [19] of the $|S \cup T| \times |S \cup T|$ matrix $\Sigma_{S\cup T, S\cup T}$ with respect to the submatrix Σ_{TT}. With the support set S viewed as fixed, we define the quantity

$$\rho_S(\Sigma) := \min_{\{T \in \mathfrak{S}_k \backslash S\}} \lambda_{\min}(\Gamma(T,S)). \quad (9)$$

As will be clarified in our analysis, this quantity controls the relative distinguishability of subsets S and T under the exponential search decoder. For the case of average and worst case error probabilities, we define the uniform bound

$$\begin{aligned}\rho_{\text{uni}}(\Sigma) :&= \min_{\{S\in\mathfrak{S}_k\}} \rho_S(\Sigma) \\ &= \min_{\{S\in\mathfrak{S}_k\}} \min_{T\in\mathfrak{S}_k\backslash S} \lambda_{\min}(\Gamma(T,S)). \end{aligned} \quad (10)$$

Note that all of these quantities are extremely simple in the case of the standard Gaussian ensemble ($\Sigma = I_{p\times p}$); in particular, we have $\Gamma(T,S) = I_{|S\backslash T|\times|S\backslash T|}$ for all pairs of distinct subsets $(S,T) \in \mathfrak{S}_k \times \mathfrak{S}_k$, and hence $\rho_S(\Sigma) = 1$ for all $S \in \mathfrak{S}_k$, and moreover $\rho_{\text{uni}}(\Sigma) = 1$.

A closely related set of quantities arise in the statement of our necessary conditions on any algorithm, as stated in Theorem 2. In particular, letting S denote a subset chosen uniformly at random from $\mathfrak{S}_k$, we define

$$\omega_{\text{ave}}(\Sigma) := \mathbb{E}_S \left[\min_{t\in S} \min_{\{z_u, u \in \{t\}\cup S^c, |z_u| \geq \frac{1}{\sqrt{2}}\}} \sum_{u,v\in\{t\}\cup S^c} \left(\Sigma_{uu} z_u^2 + \Sigma_{vv} z_v^2 - 2\Sigma_{uv} z_u z_v\right) \right]. \quad (11)$$

As our analysis will demonstrate, this quantity measures the difficulty of distinguishing a subset S from the collection of subsets T that differ from it in only one position. The second quantity that we define plays a role in specifying the bulk effect of all subsets in $\mathfrak{S}_k$

$$\omega_{\mathrm{bu}}(\Sigma) := \mathbb{E}_S \left[\min_{\{z_S \in \mathbb{R}^k \,|\, |z_j| \geq 1 \ \forall j\}} z_S^T \Sigma_{SS} z_S \right]. \quad (12)$$

As with the quantities $(\rho_S, \rho_{\mathrm{uni}})$ involved that arise in the statement of Theorem 1, these quantities are especially simple for the case of the standard Gaussian ensemble ($\Sigma = I_{p\times p}$); in particular, we have $\omega_{\mathrm{ave}}(I_{p\times p}) = 1$ and $\omega_{\mathrm{bu}}(I_{p\times p}) = k$. More generally, the quantity $\omega_{\mathrm{ave}}(\Sigma)$ is closely related to the quantity $\rho_{\mathrm{uni}}(S)$; in particular, letting $\Sigma_{u,v}$ denote the 2×2 submatrix of Σ indexed by (u, v), we have the inequality

$$\omega_{\mathrm{ave}}(\Sigma) \overset{(a)}{\geq} \min_{u \neq v} \lambda_{\min}(\Sigma_{u,v}) \overset{(b)}{\geq} \rho_{\mathrm{uni}}(\Sigma). \quad (13)$$

Inequality (a) follows from the definition (11), whereas inequality (b) follows because by choosing subsets S and T such that $S \backslash T = \{u, v\}$, we have

$$\begin{aligned} \lambda_{\min}(\Sigma_{u,v}) &\geq \lambda_{\min}(\Sigma_{u,v} - \Sigma_{(u,v),T}(\Sigma_{TT})^{-1}\Sigma_{T,(u,v)}) \\ &\geq \rho_{\mathrm{uni}}(\Sigma). \end{aligned}$$

As mentioned above, these inequalities are met with equality for the standard Gaussian ensemble ($\Sigma = I_{p\times p}$); in Section III-B, we provide a more general family of matrices for which $\omega_{\mathrm{ave}}(\Sigma) = \rho_{\mathrm{uni}}(\Sigma)$ (in particular, see Example 2).

C. Statement of Sufficient Conditions

We now have the necessary ingredients to state conditions on the triplet (n, p, k), minimum value $\beta_{\min}$, and design condition parameters $\rho_S(\Sigma)$ or $\rho_{\mathrm{uni}}(\Sigma)$ that are sufficient to ensure exact support recovery using an optimal decoder ψ^* (to be specified later). So as to simplify the statements of our results, we define the function

$$\begin{aligned} &g(c_1, p, k, \beta_{\min}, \sigma, \rho(\Sigma)) \\ &\quad = (c_1 + 2048) \max \left\{ \log \binom{p-k}{k}, \frac{\log(p-k)}{\rho(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}} \right\}. \quad (14) \end{aligned}$$

Here, the quantity $\rho(\Sigma)$ will be set to either ρ_S or ρ_{uni}, depending on the error probability under discussion.

Theorem 1 (Sufficient Conditions): Given a problem instance (β^*, y, X) from the Σ-Gaussian linear observation model (2), there exists a decoder ψ^* with the following characteristics.

(a) For any fixed vector $\beta^* \in \mathbb{U}(\beta_{\min})$ with fixed support $S \in \mathfrak{S}_k$, if the sample size satisfies

$$n > \bar{n}_S := g(c_1, p, k, \beta_{\min}, \sigma, \rho_S(\Sigma)) \quad (15)$$

for some $c_1 > 0$, then $q_S(\psi^*) \leq \exp(-c_1(n-k))$.

(b) For the support set S chosen uniformly at random from $\mathfrak{S}_k$, if the sample size satisfies

$$n > \bar{n}_{\mathrm{ave}} := g(c_1, p, k, \beta_{\min}, \sigma, \rho_{\mathrm{uni}}(\Sigma)) \quad (16)$$

for some $c_1 > 0$, then $q_{\mathrm{ave}}(\psi^*) \leq \exp(-c_1(n-k))$.

(c) For the support set S chosen adversarially from $\mathfrak{S}_k$, if the sample size satisfies

$$n > \bar{n}_{\mathrm{uni}} := \log \binom{p}{k} + g(c_1, p, k, \beta_{\min}, \sigma, \rho_{\mathrm{uni}}(\Sigma)) \quad (17)$$

for some $c_1 > 0$, then $q_{\max}(\psi^*) \leq \exp(-c_1(n-k))$.

Remarks: Note that there are only minor differences on the conditions required for the three different types of error probability. The mildest conditions are required for q_S corresponding to the error probability associated with a fixed subset. It requires only bounds on $\rho_S(\Sigma)$ from (9)—that is, only a uniform lower bound on the eigenvalues of the matrices $\{\Gamma(T, S), T \in \mathfrak{S}_k \backslash S\}$, as defined previously (8). The error probabilities q_{ave} and $q_{\max}$ involve any possible subset, and so require lower bounds on the quantity $\rho_{\mathrm{uni}}(\Sigma)$, which measures eigenvalues uniformly over $\Gamma(T, S)$ for all distinct pairs $(S, T) \in \mathfrak{S}_k \times \mathfrak{S}_k$. In addition, the worst case error probability $q_{\max}$ requires an additional term $\log \binom{p}{k}$ in the definition of $\bar{n}_{\mathrm{uni}}$, corresponding to the (logarithm of the) number of possible subsets of cardinality k.

D. Statement of Necessary Conditions

Thus far, we have provided sufficient conditions for an exhaustive search decoder to succeed with high probability in recovering the support set. Of equal interest and complementary in nature are necessary conditions that must be satisfied by any method for reliable recovery to be possible. We state a result of this nature in this subsection.

Before proceeding, note that for any fixed subset S, it is not possible to provide any type of lower bound on the probability $\inf_\psi \mathbb{P}[\psi(y, X) \neq S \,|\, S]$, since the trivial decoder $\psi(y, X) = S$ for all (y, X) always achieves perfect recovery in this setting. Accordingly, it is necessary to lower-bound either the average probability of error (with S drawn uniformly at random from $\mathfrak{S}_k$) or the worst case probability of error. The following result provides lower bounds on the sample size for the average error probability. Since the adversarial setting is not any easier, the following theorem also provides lower bounds for the worst case error.

Theorem 2 (Necessary Conditions): Consider the family of problem instances (β^*, y, X) defined by random Σ-Gaussian designs and the linear observation model (2), with S chosen uniformly at random from $\mathfrak{S}_k$. If the sample size is upper-bounded as

$$n < \underline{n} := \max \left\{ \frac{\log \binom{p}{k}}{8\omega_{\mathrm{bu}}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}}, \frac{\log(p-k)}{4\omega_{\mathrm{ave}}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}} \right\} \quad (18)$$

then for any decoding algorithm $\psi : (y, X) \to \mathfrak{S}_k$, there exists a vector $\beta^* \in \mathbb{U}(\beta_{\min})$ such that

$$q_{\max}(\psi) \geq q_{\mathrm{ave}}(\psi) = \mathbb{P}[\psi(y, X) \neq S] \geq \frac{1}{2}.$$

The proof of this claim, given in Section V, is somewhat more indirect in nature, based on the Fano method [8], [18], [20], [35], [34] in order to lower-bound the probability of error for

a restricted ensemble, which can be viewed as a certain type of hypothesis testing or channel decoding problem.

III. Some Consequences of Our Results

In this section, we explore some consequences of our results. We begin by discussing two regimes in which Theorems 1 and 2 provide a sharp characterization of the sample complexity of the subset recovery problem. By comparison to known threshold results on the Lasso (ℓ_1-constrained quadratic programming), these results reveal that the Lasso is information-theoretically optimal in some regimes, while dramatically suboptimal in others. We then discuss conditions on the design covariance matrix Σ, and show with an explicit construction that the exhaustive search decoder can succeed for designs Σ for which the Lasso fails with high probability.

A. Consequences for Different SNR and Sparsity Regimes

We begin by discussing some regimes of SNR and sparsity in which the results of Theorems 1 and 2 provide a sharp characterization of the sample complexity of the subset recovery problem. In order to make explicit comparisons to the Lasso (ℓ_1-constrained quadratic programming), we begin by stating its sample complexity. For random design matrices X drawn from any Σ-Gaussian ensemble satisfying a certain mutual incoherence condition (see (25) to follow), Wainwright [31] establishes a sharp phase transition for the success/failure of the Lasso. If we assume that $\lambda_{\min}(\Sigma_{SS})$ and the incoherence parameter remain bounded away from (p, k), the Lasso threshold [31] is of the form

$$n_{\mathrm{LAS}} := \left[c_1 k + \frac{{c_1}'}{\beta_{\min}^2/\sigma^2}\right]\log(p-k) \tag{19}$$

for constants $c_1, {c_1}' > 0$.

1) Regime of Bounded Norm Vectors: We begin by considering the regime of bounded norm vectors (i.e., $\|\beta^*\|_2 = \mathcal{O}(1)$), which implies (due to k-sparsity of β^*) that $\beta_{\min}^2 \leq \frac{c_2}{k}$ for some constant $c_2 > 0$. In this regime, we have the following corollary of Theorems 1 and 2.

Corollary 1: Consider a signal $\beta^* \in \mathbb{R}^p$ with $\beta_{\min}^2 = \mathcal{O}(\frac{1}{k})$. Then the information-theoretic sample complexity of subset selection is given by

$$n \asymp \left(k + \frac{1}{\beta_{\min}^2/\sigma^2}\right)\log(p-k)$$

More precisely, there exist constants $(c_\ell, {c_\ell}', c_u, {c_u}')$ as follows.

(a) For sequences (n, p, k) such that

$$n > \left(c_u k + \frac{{c_u}'}{\beta_{\min}^2/\sigma^2}\right)\log(p-k) \tag{20}$$

the exhaustive search decoder has error probability $q_{\max}[\psi^*] \leq 4\exp(-c(n-k))$.

(b) For sequences (n, p, k) such that

$$n < \left(c_\ell k + \frac{{c_\ell}'}{\beta_{\min}^2/\sigma^2}\right)\log(p-k) \tag{21}$$

any algorithm fails often—that is, $q_{\mathrm{ave}}[\psi] \geq 1/2$.

Remarks: By comparison to the Lasso threshold (19), Corollary 1 implies that for signals with $\beta_{\min}^2 = \mathcal{O}(1/k)$, the sample complexity of the Lasso is equal, up to constants independent of p, k, and $\beta_{\min}^2$, to the information-theoretic capacity.

Although Theorems 1 and 2 provide matching scalings for $\beta_{\min}^2 = \mathcal{O}(1/k)$, it should be noted that the conditions do not match for all scalings of the squared minimum value $\beta_{\min}^2$. For instance, if the squared minimum value is constant (i.e., $\beta_{\min}^2 = c_2$ for some constant c_2), then Theorem 2 implies that $n = c\log(p-k)$ samples are needed, whereas Theorem 1 dictates that $n = c'\log\binom{p}{k}$ samples are sufficient. It remains an open question to determine the sharp order of scaling for such regimes.

2) Consequences for Linear Sparsity: We have seen that the Lasso is information-theoretically optimal for certain regimes of the SNR parameter $\beta_{\min}^2$. In contrast, for other regimes of SNR and sparsity, Theorem 1 reveals a dramatic difference between the ℓ_1-based Lasso, and the performance of the optimal decoder. This difference appears in the regime of linear sparsity, in which $k = \alpha p$ for some $\alpha \in (0,1)$. This linear sparsity regime is particularly relevant for compressed sensing [5], [11], where the parameter α corresponds to the fraction of nonzero entries in a signal to be reconstructed. First, note that if $k = \alpha p$, then according to the previously stated Lasso threshold (19), there is a constant c_4 such that (for any scaling of the squared minimum value $\beta_{\min}^2$) the Lasso fails unless the sample size satisfies $n > c_4 p\log p$. Hence, the number of samples required by the Lasso grows faster than linearly (i.e., $p\log p \gg p$). In sharp contrast, as long as the squared minimum value $\beta_{\min}^2$ does not decrease too quickly (as made precise below), Theorems 1 and 2 imply that the information-theoretic threshold is $n = \Theta(p)$ observations.

The following corollary makes these observations precise. To simplify the statement, for $\alpha \in (0, 1/2)$, define the function $t_u(\alpha, c_1, \frac{c_2}{\sigma^2}, c_3)$

$$\alpha + (2048 + c_1)\max\left\{(1-\alpha)h\left(\frac{\alpha}{1-\alpha}\right), \frac{2\alpha}{c_3 c_2/\sigma^2}\right\} \tag{22}$$

as well as the function

$$t_\ell\left(\alpha, \frac{c_2}{\sigma^2}, c_3\right) := \frac{\alpha}{4c_3 c_2/\sigma^2}. \tag{23}$$

Here $h : [0,1] \to [0,1]$ is the binary entropy function $h(s) = -s\log s - (1-s)\log(1-s)$. With this notation, we havethe following.

Corollary 2: Consider a signal $\beta^* \in \mathbb{R}^p$ with linear sparsity (i.e., $k = \alpha p$ for some $\alpha \in (0, 1/2)$). Suppose that the minimum value $\beta_{\min}^2 = c_2\frac{\log k}{k}$ for some $c_2 > 0$. Then the information-theoretic threshold for subset recovery is $n \asymp p$. More precisely:

(a) Given size $n \geq t_u(\alpha, c_1, \frac{c_2}{\sigma^2}, \rho_S(\Sigma))p$, the exhaustive search decoder has error probability $q_S(\psi^*) \leq 4\exp(-c_1 n)$.

(b) Conversely, if $n \leq t_\ell(\alpha, \frac{c_2}{\sigma^2}, \omega_{\mathrm{ave}}(\Sigma))p$, then the probability of error of any algorithm is at least $1/2$.

Remark: Note that we have

$$\lim_{\alpha \to 0^+} t_u\left(\alpha, c_1, \frac{c_2}{\sigma^2}, \rho_S(\Sigma)\right) = 0.$$

Consequently, for any fixed SNR constant c_2 and design parameter $\rho_S(\Sigma)$, for sufficiently small fractions $\alpha \in (0,1)$, the optimal decoder can recover with $n/p < 1$. We note that the constant 2048 in the definition (22) of the threshold function t_u is far from optimal,[1] but it certainly could be improved by more careful control of constants in the large deviations analysis.

Proof: Recall the required sample size from (15) of Theorem 1. Under the stated conditions of the corollary, the ratio $\bar{n}_S/p$ is given by

$$\alpha + (c_1 + 2048)\max\left\{\frac{\log\binom{(1-\alpha)p}{\alpha p}}{p}, \frac{\log((1-\alpha)p)}{\log \alpha p}\frac{\alpha}{c_2 c_3}\right\}.$$

For $\alpha \in [0, 1/2]$ and $p \geq 2$, we have $\frac{\log(1-\alpha)p}{\log \alpha p} \leq 2$. Moreover, from standard bounds on binomial coefficients (see bound (54) in Appendix C), for $\alpha \in [0, 1/2]$, we have

$$\log\binom{(1-\alpha)p}{\alpha p} \leq p(1-\alpha)h\left(\frac{\alpha}{1-\alpha}\right).$$

Combining the pieces yields the stated claim in part (a).

Turning to the claim in part (b), from Theorem 2, we know that at least $\frac{\log(p-k)}{4\omega_{\text{ave}}(\Sigma)\beta^2_{\min}/\sigma^2}$ samples are required. Substituting in $\beta^2_{\min} = \frac{c_2 \log k}{k}$ and $k = \alpha p$, we obtain

$$\begin{aligned}\frac{\log(p-k)}{4\omega_{\text{ave}}(\Sigma)\beta^2_{\min}/\sigma^2} &= \frac{\log((1-\alpha)p)}{4\omega_{\text{ave}}(\Sigma)\frac{c_2\log(\alpha p)}{\sigma^2 \alpha p}} \\ &= \frac{\alpha p}{4\omega_{\text{ave}}(\Sigma)\frac{c_2}{\sigma^2}}\left[\frac{\log((1-\alpha)p)}{\log \alpha p}\right] \\ &\geq \frac{\alpha p}{4\omega_{\text{ave}}(\Sigma)\frac{c_2}{\sigma^2}}\end{aligned}$$

where the final inequality uses the fact that $\alpha \in (0, 1/2)$. In summary, for a squared minimum value scaling as $\beta^2_{\min} \geq \frac{c_2 \log k}{k}$ for some constant c_2, Corollary 2 demonstrates that the Lasso is highly suboptimal in the linear sparsity regime $k = \alpha p$. Regardless of the linear fraction α and the squared minimum $\beta^2_{\min}$, success of the Lasso for support recovery [31] requires the number n of samples to scale so quickly such that $n/p \to +\infty$.

As pointed out by one of the reviewers, results by Candes and Tao [6] on the Dantzig selector (an ℓ_1-based relaxation) apply to the linear–linear regime of Corollary 2. In particular, for measurement matrices X drawn from the standard Gaussian ensemble, they establish bounds on the mean-square error (MSE) prediction as well as on the ℓ_2 error $\|\hat{\beta} - \beta\|_2^2$ of the Dantzig selector. These results show that for the case of design matrices X drawn from the standard Gaussian ensemble (with i.i.d. $N(0,1)$) entries, a sample size of

$$n \asymp k \log(p/k) \tag{24}$$

is sufficient to achieve a squared ℓ_2 reconstruction error that is bounded. Related results by Meinshausen and Yu [23] and Bickel *et al.* [3] have similar consequences for the Lasso.

[1]As pointed out by a reviewer, it requires that $\alpha \approx 10^{-4}$ for a meaningful result.

In the context of this paper (which focuses exclusively on support recovery), we note that the criteria of support recovery is related to but distinct from the criteria of prediction error $\mathbb{E}\|y - X\hat{\beta}\|_2^2$, or on the ℓ_2 reconstruction error $\|\hat{\beta} - \beta^*\|_2^2$. On one hand, given a procedure that correctly recovers the support S of the unknown vector β^*, then of course we can simply restrict our problem to the subset S, and use standard methods (e.g., ordinary linear regression) to obtain estimates with good MSE prediction or ℓ_2 error. In the opposite direction, however, an estimate $\hat{\beta}$ can be close to β^* but still have a different support than the true vector β^*. Indeed, as discussed above, for standard Gaussian matrices, the sample size (24) guarantees that the Dantzig selector and Lasso achieve ℓ_2 errors that are bounded. As pointed out by one of the reviewers, if the minimum value $\beta_{\min}$ were also strictly bounded away from zero and if $\hat{\beta}$ also had entries bounded above by $\beta_{\min}$, then an estimate $\hat{\beta}$ such that $\|\hat{\beta} - \beta^*\|_2^2 = \mathcal{O}(\frac{k \log \frac{p}{k}}{n}) = \mathcal{O}(1)$ would be sufficient to guarantee support recovery. However, in the regimes of practical interest, the minimum value $\beta_{\min}$ decreases to zero at some rate (e.g., $\beta_{\min} = \mathcal{O}(1/\sqrt{k})$ when β^* has constant ℓ_2 norm), so that ℓ_2 recovery with constant error bounds is not sufficient. Indeed, a consequence of the results of Wainwright [31] is that the Lasso requires $n \asymp k\log(p-k)$ samples to perform support recovery, which scales much more rapidly than $k\log(p/k)$ in the case of linear sparsity. Theorem 2 demonstrates that when $\beta_{\min} = \mathcal{O}(1/\sqrt{k})$, this $k\log(p-k)$ scaling—as opposed to $k\log(p/k)$—is unavoidable for subset selection.

Moving onto consideration of arbitrary methods, a consequence of Corollary 2 is that no method can recover the support exactly with $n = \Theta(p)$ observations unless the squared minimum value is lower-bounded as $\beta^2_{\min} \geq \frac{c_2 \log k}{k}$. Nonetheless, it is an interesting question to consider the subset selection performance of other computationally efficient methods.

B. Conditions on the Design Covariance

It is worthwhile comparing the conditions on the design matrix Σ imposed by Theorems 1 and 2 to those conditions imposed in past work on ℓ_1-based methods. One set of conditions, sufficient for guarantees in terms of ℓ_2 or prediction error, are based on restricted isometry properties (RIP) [5], [12], requiring that the condition numbers of various submatrices of the matrix $X^T X/n$ are uniformly very close to one. (For instance, among other conditions, RIP requires the bound $\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} \leq 1+\delta$, for a suitably small δ.) By known concentration results in random matrix theory [9], such RIP conditions hold with high probability for design matrices X whose columns are close to orthogonal (e.g., for X drawn from the standard Gaussian ensemble with $\Sigma = I_{p\times p}$ and $n \asymp k\log(p/k)$). It should be noted that these RIP conditions, while sufficient, are far from necessary to obtain bounds on ℓ_2 or prediction error; we refer the reader to Bickel *et al.* [3] for a much weaker set of conditions for ℓ_2 and prediction error consistency.

By contrast, the focus of this paper is on the problem of exact support recovery, for which a related but distinct set of conditions on the design covariance Σ are known to be necessary and sufficient. First, successful Lasso-based support recovery requires that the minimal eigenvalue $\lambda_{\min}(\Sigma_{SS})$ stay bounded

away from zero, and secondly (and more significantly), that a certain mutual incoherence parameter stays bounded strictly away from zero—namely

$$\gamma(\Sigma) := 1 - \max_{j \notin S} \|\Sigma_{jS}(\Sigma_{SS})^{-1}\|_1 > 0. \tag{25}$$

Whereas the lower bound on $\lambda_{\min}(\Sigma_{SS})$ is a mild condition, the *mutual incoherence condition* (25) is more restrictive. It was first defined in the context of sample design matrices independently by Fuchs [17] and Tropp [29], and also imposed in other high-dimensional analysis of the Lasso [22], [36], [31]. Note that the eigenvalue lower bound and mutual incoherence condition (25) are trivially satisfied for random design matrices drawn from the standard Gaussian ensemble ($\Sigma = I_{p\times p}$).

It is known [36], [31] that if the Lasso is applied to any ensemble of Σ-Gaussian measurement matrices for which the incoherence assumption (25) is violated and the noise w has a symmetric distribution, it will fail with probability at least $1/2$, regardless of the sample size (see Wainwright [31] for a precise statement). Exploiting this fact, the following examples show that there exist covariance matrix families for which the optimal decoder can succeed while the Lasso fails.

Example 1: Consider the Σ-Gaussian family with covariance matrices of the form

$$\Sigma := \begin{bmatrix} 1 & \mu & \mu & \dots & \dots & \mu \\ \mu & 1 & 0 & \dots & \dots & 0 \\ \mu & 0 & 1 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \tag{26}$$

for some $\mu \in \mathbb{R}$. In particular, for $\mu = (2\sqrt{p})^{-1}$, it can be verified that we have $\lambda_{\min}(\Sigma) \geq 1/2$ uniformly for all $p = 2, 3 \dots$.

Consider some k-sized subset $S \in \mathfrak{S}_k$ that does *not* include the index $\{1\}$, and let T be another k-sized subset. Using the notation $\ell = |S \backslash T|$, we have

$$\Gamma(T,S) = I_{\ell\times\ell} - \Sigma_{(S\backslash T),T}(\Sigma_{TT})^{-1}\Sigma_{T(S\backslash T)}.$$

If T also does not include the index $\{1\}$, then $\Sigma_{T(S\backslash T)} = 0$, so that $\Gamma(T,S) = I_{\ell\times\ell}$. In the more interesting case that T includes the index $\{1\}$, a little calculation shows that

$$\Gamma(T,S) = I_{\ell\times\ell} - \frac{\mu^2}{1-\mu^2(k-1)}\mathbf{1}_\ell\mathbf{1}_\ell^T$$

where $\mathbf{1}_\ell \in \mathbb{R}^\ell$ is a vector of all ones. Consequently, for this family

$$\begin{aligned} \rho_S(\Sigma) &= \min_{\ell=1,\dots,k} \left\{ 1 - \frac{\mu^2\ell}{1-\mu^2(k-1)} \right\} \\ &\geq 1 - \frac{\frac{1}{4}\frac{k}{p}}{1 - \frac{1}{4}\frac{k-1}{p}} \\ &\geq 1 - \frac{\frac{1}{4}}{\frac{3}{4}} = 2/3 \end{aligned}$$

where the first inequality uses the definition $\mu^2 = \frac{1}{4p}$, and the second inequality uses the fact that $k \leq p$. Consequently, the optimal decoder succeeds in recovering the support set S with $n > \bar{n}_S \asymp \max\{\log\binom{p}{4k}, \frac{\log(p-k)}{\beta^2_{\min}}\}$ observations.

On the other hand, suppose that the given subset S has cardinality $k > 2\sqrt{p}$. By definition of the covariance matrix (26) and the mutual incoherence parameter $\gamma(\Sigma)$ in (25), we have

$$\gamma(\Sigma) \leq 1 - \Sigma_{1S}(\Sigma_{SS})^{-1} = 1 - k\mu < 0$$

showing that with $\mu = (2\sqrt{p})^{-1}$, the mutual incoherence condition (25) is violated. Consequently, for this ensemble and for any subset with $k > 2\sqrt{p}$ elements that excludes the index $\{1\}$, the probability of *incorrect* support recovery using ℓ_1-constrained quadratic programming is at least one half [31], regardless of the sample size, whereas the optimal decoder will succeed with high probability for sample sizes $n > \bar{n}_S$. ♢

It is also interesting to consider the quantities $\rho_{\text{uni}}(\Sigma)$ and $\omega_{\text{ave}}(\Sigma)$ that arise in the necessary and sufficient condition of Theorems 1 and 2. As previously shown (13), the quantity $\rho_{\text{uni}}(\Sigma)$ always lower-bounds the quantity $\omega_{\text{ave}}(\Sigma)$. The following example provides a family of matrices for which this lower bound is met with equality, so that the dependence on the design covariance Σ identified by Theorems 1 and 2 is tight.

Example 2: Letting $\vec{1} \in \mathbb{R}^p$ denote the all-ones vector, consider the family of covariance matrices

$$\Sigma := (1-\mu)I_{p\times p} + \mu\vec{1}_p\vec{1}_p^T. \tag{27}$$

In this example, we show that for a squared minimum value of the order $\beta^2_{\min} = \frac{c_2}{k}$ and any $\mu \in [0, +1)$, Theorems 1 and 2 predict that the critical sample size scales as $n \asymp \frac{k\log(p-k)}{1-\mu}$. To establish this fact, we begin by calculating the matrices $\Gamma(T,S)$ that define the quantity $\rho_{\text{uni}}(\Sigma)$. For any p, any $\mu \in [0, +1)$, and any pair (S,T) of k-sized subsets with $\ell = |S\backslash T| \geq 1$, a little calculation (using the matrix inversion formula [19]) shows that

$$\Gamma(T,S) = (1-\mu)I_{\ell\times\ell} + \frac{\mu(1-\mu)}{1+\mu(k-1)}\mathbf{1}_\ell\mathbf{1}_\ell^T \tag{28}$$

so that for any subset $\rho_S(\Sigma) = 1-\mu$, and hence $\rho_{\text{uni}}(\Sigma) = 1-\mu$. Consequently, the exhaustive search decoder succeeds with high probability (w.h.p.) as long as the sample size satisfies

$$n > \frac{ck\log(p-k)}{(1-\mu)}$$

for some constant $c > 0$.

Let us compare this sufficient condition to the necessary conditions from Theorem 2. For this particular covariance matrix Σ, a little calculation shows that

$$\begin{aligned} \omega_{\text{ave}}(\Sigma) &= \min_{(z_1,z_2),|z_i|\geq\frac{1}{\sqrt{2}}} \begin{bmatrix} z_1 & -z_2 \end{bmatrix} \begin{bmatrix} 1 & \mu \\ \mu & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ -z_2 \end{bmatrix} \\ &= 1-\mu \end{aligned}$$

and moreover that $\omega_{\text{bu}}(\Sigma) = k(1-\mu)$. Therefore, for this ensemble with $\beta^2_{\min} = \frac{c_2}{k}$, Theorem 2 implies that if

$$n < \frac{c'k\log(p-k)}{(1-\mu)}$$

for some constant $c' \leq c$, then the error probability of any algorithm is at least $1/2$. Consequently, for this ensemble of matrices Σ parameterized by $\mu \in [0,1)$, Theorems 1 and 2 provide a set of necessary and sufficient conditions that are matching up to constant factors independent of (p,k,μ).

On the other hand, for any k-sized subset S, the condition number of the submatrix Σ_{SS} is given by

$$\frac{\lambda_{\max}(\Sigma_{SS})}{\lambda_{\min}(\Sigma_{SS})} = \frac{1+(k-1)\mu}{1-\mu}$$

which tends to infinity for any fixed $\mu \neq 0$. $\diamondsuit$

Finally, we provide an example to illustrate that an upper bound on the maximum eigenvalue $\lambda_{\max}(\Sigma_{SS})$ is not required for ℓ_1-based support recovery.

Example 3: For a given k-sized subset S, consider the covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ given by

$$\Sigma_{ij} = \begin{cases} 1, & \text{if } i = j \\ \mu, & \text{if } i \neq j,\ i,j \in S \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

For this matrix, a simple calculation shows that the maximum eigenvalue $\lambda_{\max}(\Sigma_{SS}) = 1+(k-1)\mu$, which tends to infinity for any fixed $\mu > 0$ as $k \to +\infty$. On the other hand, since $\Sigma_{ij} = 0$ for all $i \neq j$ outside of the subset S, the mutual incoherence condition (25) is satisfied with $\gamma = 1$. Moreover, a little calculation shows that $\lambda_{\min}(\Sigma_{SS}) = 1-\mu > 0$. Therefore, the Lasso will succeed using $n \asymp \frac{k\log(p-k)}{1-\mu}$ samples. $\diamondsuit$

IV. Proof of Theorem 1

This section is devoted to the proof of Theorem 1. We begin by setting up some useful notation to be used throughout the remainder of the analysis. 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., β, y, etc.). In an analogous manner, we use X_V to denote the $n \times |V|$ matrix with columns $\{X_i, i \in V\}$. We use X' to denote the transpose of a matrix X.

A. Exhaustive Search Decoder

Our route to establishing the sufficient conditions in Theorem 1 is by direct analysis of the decoder ψ^* that searches exhaustively over all $\binom{p}{k}$ possible subsets of size k. More specifically, the search decoder obtains its estimate $\hat{S}$ by the following two-step procedure.

(a) For each of the $\binom{p}{k}$ subsets subset $T \subset \{1,\ldots,p\}$ of size k, solve the quadratic program

$$Z(T) := \min_{\beta_T \in \mathbb{R}^k} \|y - X_T\beta_T\|_2^2. \tag{30}$$

(b) Return the subset $\hat{S} = \arg\min_{|T|=k} Z(T)$.

Of interest to us are various error probabilities associated with this procedure. We begin by bounding the S-based error

$$q_S(\psi) := \mathbb{P}[\psi(y,X) \neq S \,|\, S]$$

with the probability taken over X and the noise vector w, when the underlying support is fixed to S. Using this result, we then bound the average error probabilities $q_{\text{ave}}(\psi)$ and the worst case error probabilities $q_{\max}(\psi)$, as defined in (6) and (7), respectively.

With $n > k$ and random matrices X drawn from a Gaussian ensemble with nondegenerate covariance Σ, each of the $n \times k$ submatrices X_T has rank k with probability one.[2] Accordingly, we may define the $n \times n$ matrices

$$\Pi_T := X_T[X_T'X_T]^{-1}X_T' \quad \text{and} \tag{31a}$$

$$\Pi_T^{\perp} := I_{n\times n} - X_T[X_T'X_T]^{-1}X_T'. \tag{31b}$$

Note that Π_T and $\Pi_T^{\perp}$ are both orthogonal projection matrices, associated with the k-dimensional range space $\text{Ra}(X_T)$ and $(n-k)$-dimensional nullspace $\text{Ker}(X_T)$, respectively. For any pair of subsets T and S, each with k elements, define the random variable

$$\Delta(T;S) := ||\|\Pi_T^{\perp}y\|_2^2 - \|\Pi_S^{\perp}y\|_2^2. \tag{32}$$

With these definitions, we state the following result.

Lemma 1: For any given vector β^* with support S, the exhaustive search decoder prefers T to the true underlying S if and only if $\Delta(T;S) < 0$.

Proof: We begin by showing that for any subset T for which X_T is full rank, the quantity $Z(T)$ defined in (30) is equal to $Z(T) = \|\Pi_T^{\perp}y\|^2$. Under the given rank condition, the linear least squares estimator of ${\beta^*}_T$ is given by $\hat{\beta}_T = [X_T'X_T]^{-1}X_T^Ty$. Noting that by the definition (31a) of Π_T, we have $X_T\hat{\beta}_T = \Pi_T y$ we substitute into the quadratic norm and expand, thereby obtaining

$$Z(T) = \|y - X_T\widehat{\beta_T}\|_2^2 = \|(I-\Pi_T)y\|_2^2 = \|\Pi_T^{\perp}y\|_2^2.$$

Failure occurs if and only if $\Delta(T;S) = Z(T) - Z(S) < 0$, as claimed.

Overall, the search decoder fails if and only if at least one T (with cardinality $|T| = k$) is preferable to S; consequently, the overall probability of error can be written as

$$\mathbb{P}[\hat{S} \neq S \,|\, S] = \mathbb{P}\left[\bigcup_{T \in \mathfrak{S}_k \backslash \{S\}} \{\Delta(T;S) < 0\}\right]. \tag{33}$$

Consequently, assuming that β^* has support S, the technical result central to analyzing the error probability (33) is tight control on the probabilities of the events $\{\Delta(T;S) < 0\}$, for all k-sized subsets $T \in \mathfrak{S}_k \backslash S$.

B. Large Deviations Bound

Before stating a large deviations bound, we require some notation. Recalling the definition (8) of $\Gamma(T,S)$, we use $\Gamma^{1/2}(T,S)$ to denote its symmetric matrix square root. For each $T \in \mathfrak{S}_k \backslash \{S\}$, define the quantity

$$f({\beta^*}_{S\backslash T}) := \|\Gamma^{1/2}(T,S){\beta^*}_{S\backslash T}\|_2^2/\sigma^2 \tag{34}$$

[2]That is, the probability that k random Gaussian random vectors in $\mathbb{R}^n$ are linearly independent is equal to zero, which follows since the Gaussian has density with respect to Lebesgue measure.

representing a type of SNR, reflecting how distinguishable subset T is from S. With this notation, we have the following.

Lemma 2: As long as $n - k \geq \frac{64}{\rho_S(\Sigma)\beta_{\min}^2/\sigma^2}$, for any pair of distinct subsets S and T, we have

$$\mathbb{P}[\Delta(T;S) < 0] \leq 4\exp\left(-(n-k)\frac{f(\beta^*{}_{S\backslash T})}{64(f(\beta^*{}_{S\backslash T})+8)}\right). \tag{35}$$

The proof of Lemma 2 is somewhat technical in nature. However, the high-level strategy is straightforward: given some $\delta > 0$, we define the events

$$\mathcal{E}_1(\delta) := \left\{\left|\frac{||\Pi_S^{\perp}(y)||_2^2 - ||\Pi_T^{\perp}(w)||_2^2}{\sigma^2(n-k)}\right| \geq \delta\right\} \text{ and} \tag{36a}$$

$$\mathcal{E}_2(\delta) := \left\{\frac{||\Pi_T^{\perp}(y)||_2^2 - ||\Pi_T^{\perp}(w)||_2^2}{\sigma^2(n-k)} \leq 2\delta\right\}. \tag{36b}$$

We now observe that for any $\delta > 0$, the event $\{\Delta(T;S) < 0\}$ implies that at least one of the events $\mathcal{E}_2(\delta)$ or $\mathcal{E}_1(\delta)$ holds. Indeed, supposing that neither $\mathcal{E}_2(\delta)$ nor $\mathcal{E}_1(\delta)$ is true, then the quantity

$$\frac{\Delta(T;S)}{\sigma^2(n-k)} = \left\{\frac{||\Pi_T^{\perp}(y)||_2^2 - ||\Pi_T^{\perp}(w)||_2^2}{\sigma^2(n-k)}\right\} + \left\{\frac{||\Pi_T^{\perp}(w)||_2^2 - ||\Pi_S^{\perp}(y)||_2^2}{\sigma^2(n-k)}\right\}$$

is lower-bounded by $2\delta - \delta > 0$.

Consequently, by union bound, it suffices to control the two probabilities $\mathbb{P}[\mathcal{E}_1(\delta)]$ and $\mathbb{P}[\mathcal{E}_2(\delta)]$. This argument applies for any choice of $\delta > 0$; a convenient choice turns out to be $\delta^* = \frac{f(\beta^*{}_{S\backslash T})}{4}$.

With this setup, the proof of Lemma 2 is a consequence of the following two results, proved in Appendices A and B, respectively.

Lemma 3: For all $\delta \geq \frac{16\ell}{n-k}$

$$\mathbb{P}[\mathcal{E}_1(\delta)] \leq 2\exp\left(-\frac{\delta(n-k)}{16}\right). \tag{37}$$

Moreover, the choice $\delta^* = \frac{f(\beta^*{}_{S\backslash T})}{4}$ is valid as long as $n - k \geq \frac{64}{\rho_S(\Sigma)\beta_{\min}^2/\sigma^2}$.

Lemma 4: With $\delta^* = \frac{f(\beta^*{}_{S\backslash T})}{4}$, we have

$$\mathbb{P}[\mathcal{E}_2(\delta^*)] \leq 2\exp\left(-(n-k)\frac{f(\beta^*{}_{S\backslash T})}{16(f(\beta^*{}_{S\backslash T})+8)}\right) \tag{38}$$

for any pair of distinct subsets S and T.

Combining these two results yields the claim of Lemma 2.

C. Analysis of Error Probability

Using Lemma 2, we are now equipped to complete the proof of Theorem 1. Denote by $N(\ell)$ the number of subsets T with cardinality k, such that $|S\backslash T| = \ell$. (Moreover, note that since both S and T have cardinality k, we have $|T\backslash S| = \ell$ as well.) A counting argument yields that, for each ℓ with $1 \leq \ell \leq k$, there are

$$N(\ell) = \binom{k}{\ell}\binom{p-k}{\ell} \tag{39}$$

such subsets.

In order to simplify the statement of the result, we begin by deriving a weaker form of the large deviations bounds from Lemma 2, albeit one that leads to simpler expressions. Observe that the function $f(t) = t/(t+8)$ is increasing on the interval $[0,\infty)$. Consider a pair of subsets S and T with overlap of size $\ell = |S\backslash T|$. Using the definition (9) of ρ_S, we have

$$f(\beta^*{}_{S\backslash T}) \geq \ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2.$$

Consequently, for any pair (S,T) with $|S\backslash T| = \ell$, the bound (35) implies that

$$\mathbb{P}[\Delta(T;S) < 0] \leq 4\exp\left(-(n-k)\frac{\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2}{64\,(\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2+8)}\right). \tag{40}$$

Combining this upper bound with the union bound applied to the expression (33) yields that $\mathbb{P}[\psi^*(y) \neq S \,|\, S]$ is upper-bounded by

$$4\sum_{\ell=1}^{k} N(\ell)\exp\left(-(n-k)\frac{\ell\rho_S(\Sigma)\beta_{\min}^2}{64\,(\ell\rho_S(\Sigma)\beta_{\min}^2+8)}\right),$$

which is further upper bounded by

$$4k\max_{\ell=1,\ldots,k}\left\{\binom{k}{\ell}\binom{p-k}{\ell} \times \exp\left(-(n-k)\frac{\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2}{64(\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2+8)}\right)\right\}.$$

Consequently, in order for the error probability to vanish asymptotically, it suffices to take $(n-k) \to +\infty$ such that $(n-k)$ is greater than

$$\max_{\ell=1,\ldots,k}\left\{64 + \frac{512}{\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2} \times \left[\log 4k + \log\binom{k}{\ell} + \log\binom{p-k}{\ell}\right]\right\}.$$

Let us upper-bound this quantity (denote it T_1). By our assumption that $k \leq p/2$, we have $\binom{k}{\ell} \leq \binom{p-k}{\ell}$. Moreover, we have

$$\log 4k \leq \max_{\ell}\log 4\binom{k}{\ell} \leq 2\max_{\ell}\log\binom{k}{\ell}.$$

Overall, we conclude that

$$T_1 \leq \max_{\ell=1,2,\ldots,k}\left\{\left[256 + \frac{2048}{\ell\rho_S(\Sigma)\beta_{\min}^2/\sigma^2}\right]\log\binom{p-k}{\ell}\right\}. \tag{41}$$

Given that the conditions of Theorem 1 certainly imply that $n > k + 256 \log \binom{p-k}{k}$, it suffices to restrict attention to the term involving $(\rho_S(\Sigma), \beta^2_{\min})$—namely, to upper-bound the quantity

$$T_1' := \frac{2048}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2} \max_{\ell=1,2,\ldots,k} \frac{\log \binom{p-k}{\ell}}{\ell} \tag{42}$$

Using standard bounds on binomial coefficients (see Appendix C), we have

$$\begin{aligned} T_1' &\leq \frac{2048}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2} \max_{\ell=1,2,\ldots,k} \log \frac{e(p-k)}{\ell} \\ &= \frac{4096 \log(p-k)}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2}. \end{aligned}$$

Returning to the upper bound (Section IV-C), we conclude that for a sample size n satisfying

$$n > k + (c_1 + 2048) \max \left\{ \log \binom{p-k}{k}, \frac{\log(p-k)}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2} \right\} \tag{43}$$

for some $c_1 > 0$, the error probability associated with detecting support set S decays exponentially as

$$q_S(\psi^*) = \mathbb{P}[\psi^*(y) \neq S \,|\, S] \leq \exp(-c_1(n-k))$$

thereby establishing the claim of Theorem 1(a).

Turning to Theorem 1(b), if we replace the quantity $\rho_S(\Sigma)$ in the lower bound (43) by $\rho_{\text{uni}}(\Sigma) := \max_{|S|=k} \rho_S(\Sigma)$, then we can conclude that the average probability of error

$$q_{\text{ave}}(\psi^*) = \frac{1}{\binom{p}{k}} \sum_{|S|=k} \mathbb{P}[\psi^*(y) \neq S \,|\, S]$$

also vanishes exponentially fast.

Finally, turning to Theorem 1(c), let us consider the case of the worst case error probability taken over all subsets S. In this case, in addition to using the worst case measure $\rho_{\text{uni}}(\Sigma)$, we also need the probability of error for any given subset S to converge to zero sufficiently fast. In particular, by union bound, we have

$$\begin{aligned} q_{\max}(\psi^*) &= \mathbb{P}[\cup_{|S|=k} \{\psi^*(y) \neq S \,|\, S\}] \\ &\leq \binom{p}{k} \mathbb{P}[\psi^*(y) \neq S \,|\, S]. \end{aligned}$$

Consequently, if for some $c_1 > 0$, we choose a sample size n such that

$$n > k + (c_1 + 2048) \max \left\{ \log \binom{p-k}{k}, \frac{\log(p-k)}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2} \right\} + \log \binom{p}{k}$$

then we have $q_{\max}(\psi^*) \leq \exp(-c_1(n-k))$.

V. Proof of Theorem 2

In this section, we prove the necessary conditions stated in Theorem 2. Our method involves two *restricted versions* of the subset recovery problem, for which the analysis can be reduced to a type of channel decoding problem. We then apply a variant of Fano's bound [8] to analyze the error probability over these restricted ensembles. We note the Fano method is a standard approach for obtaining minimax lower bounds in nonparametric statistical problems [18], [20], [35], [34].

A. Basic Setup

We begin by describing the basic setup for the proof of Theorem 2. Let $\mathcal{E} \subseteq \mathfrak{S}_k$ denote a particular subset of the set $\mathfrak{S}_k$ of all k-sized subsets, and let Ξ be a set-valued random variable, uniformly distributed over $\mathcal{E}$—that is, $\mathbb{P}[\Xi = S] = \frac{1}{|\mathcal{E}|}$ for all $S \in \mathcal{E}$. Suppose that the decoder is told that the selected subset S is a member of $\mathcal{E}$, and moreover it is provided with the form of the vectors ${\beta^*}_S$ for all $S \in \mathcal{E}$. Its goal is to use the pair (y, X) to recover the unknown subset S. Note that the two forms of side information—namely, that $S \in \mathcal{E}$, and the form of ${\beta^*}_S$ for any fixed $S \in \mathcal{E}$—cannot harm the decoder's performance, since the decoder can always choose to ignore this information. Consequently, the error probability of the decoder for the original problem is lower-bounded by the error probability $\mathbb{P}[\psi(y, X) \neq \Xi]$, where Ξ is uniform over $\mathcal{E}$. This is a multi-way hypothesis testing problem, and we may lower-bound the probability of error of any decoder using Fano's inequality [8].

We lower-bound this error probability as follows: first, for any fixed $X \in \mathbb{R}^{n \times p}$ and for any decoder ψ

$$\mathbb{P}[\psi(y, X) \neq \Xi \,|\, X] \geq 1 - \frac{I(\Xi; y \,|\, X) + \log 2}{\log |\mathcal{E}|} \tag{44}$$

where $I(\Xi; y \,|\, X)$ is the mutual information between Ξ and y conditioned on X; explicitly, it is given by $I(\Xi; y \,|\, X) = \mathbb{E}_{\Xi, y}\left[\log \frac{\mathbb{P}(y, \Xi \,|\, X)}{\mathbb{P}(y \,|\, X)\mathbb{P}(\Xi)}\right]$. Taking expectations of both sides (44), we conclude that

$$\mathbb{P}[\psi(y, X) \neq \Xi] \geq 1 - \frac{\mathbb{E}_X[I(\Xi; y \,|\, X)] + \log 2}{\log |\mathcal{E}|}. \tag{45}$$

Consequently, in order to make effective use of the Fano lower bound (44) or its averaged form (45), we need to construct ensembles $\mathcal{E}$ for which $\log |\mathcal{E}|$ is relatively large while the mutual information $I(\Xi; y \,|\, X)$ is relatively small. In our analysis, we make use of the upper bound on the mutual information

$$I(\Xi; y \,|\, X) \leq \frac{1}{|\mathcal{E}|^2} \sum_{(S,T) \in \mathcal{E} \times \mathcal{E}} D(\mathbb{P}_{y \,||\, S,X} \,|\, \mathbb{P}_{y \,|\, T,X}) \tag{46}$$

which follows from the convexity of mutual information [8]. Here, the quantity

$$D(\mathbb{P}_{y \,||\, S,X} \,|\, \mathbb{P}_{y \,|\, T,X}) := \mathbb{E}_y \left[\log \frac{\mathbb{P}_{y \,|\, S,X}(y \,|\, S, X)}{\mathbb{P}_{y \,|\, S,X}(y \,|\, T, X)} \right] \tag{47}$$

is the Kullback–Leibler divergence between the distributions $\mathbb{P}_{y \,|\, S,X}$ and $\mathbb{P}_{y \,|\, T,X}$.

B. Bound for Bulk Ensemble

The bulk ensemble is defined by the choice $\mathcal{E} = \mathfrak{S}_k$, and then setting

$$\beta^*{}_S = \beta^2_{\min} \left[\arg \min_{\{u \in \mathbb{R}^k \,|\, |u_j| \geq 1 \quad \forall j\}} u^T \Sigma_{SS} u \right] \tag{48}$$

for each $S \in \mathcal{E}$. A straightforward computation yields that

$$\begin{aligned} D(\mathbb{P}_{y\,\|\,S,X}\,|\,\mathbb{P}_{y\,|\,T,X}) &= \frac{n\beta_{\min}^2}{2\sigma^2}\left(\frac{1}{n}\|X_S\beta^*{}_S - X_T\beta^*{}_T\|_2^2\right) \\ &\leq \frac{n\beta_{\min}^2}{2\sigma^2}\left(\frac{2}{n}\|X_S\beta^*{}_S\|_2^2 + \frac{2}{n}\|X_T\beta^*{}_T\|_2^2\right) \end{aligned}$$

so that we have

$$I(\Xi; y\,|\,X) \leq \frac{n\beta_{\min}^2}{\sigma^2}\left[\frac{1}{\binom{p}{k}^2}\sum_{(S,T)\in\mathfrak{S}_k\times\mathfrak{S}_k} Z(S,T)\right] \quad (49)$$

where

$$Z(S,T) := \frac{1}{n}\|X_S\beta^*{}_S\|_2^2 + \frac{2}{n}\|X_T\beta^*{}_T\|_2^2,$$

and

$$\bar{Z} := \binom{p}{k}^{-2}\sum_{(S,T)\in\mathfrak{S}_k\times\mathfrak{S}_k} Z(S,T).$$

Note that each $Z(S,T)$ is a random variable (as a function of the random design matrix X); a little calculation shows that

$$\mathbb{E}[Z(S,T)] = (\beta^*{}_S)^T\Sigma_{SS}\beta^*{}_S + (\beta^*{}_T)^T\Sigma_{TT}\beta^*{}_T$$

and, moreover, that

$$\mathbb{E}[\bar{Z}] \leq \frac{2}{\binom{p}{k}}\sum_{S\in\mathfrak{S}_k}(\beta^*{}_S)^T\Sigma_{SS}\beta^*{}_S = 2\omega_{\mathrm{bu}}(\Sigma)$$

where the final inequality follows by our choice (48) of $\beta^*{}_S$, and the definition (12) of ω_{bu}. Consequently, using the bound (49), we have $\mathbb{E}_X[I(\Xi; y\,|\,X)] \leq 2\omega_{\mathrm{bu}}(\Sigma)n\frac{\beta_{\min}^2}{\sigma^2}$, and hence, using the bound (45)

$$\begin{aligned} \mathbb{P}[\psi(y,X)\neq\Xi] &\geq 1 - \frac{2\omega_{\mathrm{bu}}(\Sigma)n\frac{\beta_{\min}^2}{\sigma^2} + \log 2}{\log\binom{p}{k}} \\ &= 1 - \frac{2\omega_{\mathrm{bu}}(\Sigma)n\frac{\beta_{\min}^2}{\sigma^2}}{\log\binom{p}{k}} - o(1). \end{aligned}$$

Consequently, if the sample size is upper bounded as

$$n < \frac{\log\binom{p}{k}}{8\omega_{\mathrm{bu}}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}}$$

then $\mathbb{P}[\psi(y,X)\neq\Xi] \geq 1/2$ as claimed.

C. Bound for Nearby Subsets

We now describe bounds based on a second ensemble, introduced by Wang *et al.* [32]. For any subset S, let $t(S)$ be an index that achieves the minimum in the definition (11) of the function ω_{ave}. We then let the ensemble $\mathcal{E}$ consist of all $(p-k+1)$ subsets that contain all $(k-1)$ indices in the set $S\backslash\{t(S)\}$, and then one more index chosen from the set $S^c \cup \{t(S)\}$. Observe that the resulting family of subsets has cardinality $|\mathcal{E}| = p-k+1$, with the property that for each distinct pair $(U,V) \in \mathcal{E}\times\mathcal{E}$ of subsets, the Hamming distance is equal to two.

For each subset $U \in \mathcal{E}$, we define its signal vector $\beta^*{}_U \in \mathbb{R}^k$ as follows:

$$\beta^*{}_{U,u} = \begin{cases} \beta_{\min}, & \text{if } u \in S\backslash\{t(S)\} \\ \beta_{\min}(\sqrt{2}z_u), & \text{if } u \in S^c \cup \{t(S)\} \end{cases}$$

where z_u achieves the minimum in (11). Note that we have $\min_{u\in U}|\beta^*{}_{U,u}| = \beta_{\min}$ by construction.

Now consider a pair of distinct subsets $U \neq V$, say with $U\backslash V = \{u\}$ and $V\backslash U = \{v\}$. With the choices given above, a little calculation shows that

$$D\big(\mathbb{P}_{y\,\|\,S,X}\,|\,\mathbb{P}_{y\,|\,T,X}\big) = \frac{n\beta_{\min}^2}{\sigma^2}(X_uz_u - X_vz_v)^2$$

so that we have

$$I(\Xi; y\,|\,X) \leq \frac{n\beta_{\min}^2}{\sigma^2(p-k+1)^2}\sum_{u,v\in S^c\cup\{t(S)}(X_uz_u - X_vz_v)^2.$$

Taking expectations and using the definition (11) of the function ω_{ave}, we obtain that $\mathbb{E}_X[I(\Xi; y\,|\,X)]$ is equal to

$$\frac{n\beta_{\min}^2}{\sigma^2(p-k+1)^2}\sum_{u,v\in S^c\cup\{t(S)}\mathbb{E}[(X_uz_u - X_vz_v)^2].$$

Expanding the expectation yields that $\mathbb{E}_X[I(\Xi; y\,|\,X)]$ is equal to

$$\frac{n\beta_{\min}^2}{\sigma^2(p-k+1)^2}\sum_{u,v\in S^c\cup\{t(S)}\left(\Sigma_{uu}z_u^2 - 2\Sigma_{uv}z_uz_v + \Sigma_{vv}z_v^2\right)$$

which is upper-bounded by $\frac{n\beta_{\min}^2\omega_{\mathrm{ave}}(\Sigma)}{\sigma^2}$, using the definition (11). Consequently, by the Fano bound, we obtain

$$\begin{aligned} \mathbb{P}[\psi(y,X)\neq\Xi] &\geq 1 - \frac{n\frac{\beta_{\min}^2}{\sigma^2}\omega_{\mathrm{ave}}(\Sigma) + \log 2}{\log(p-k)} \\ &= 1 - \frac{n\omega_{\mathrm{ave}}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}}{\log(p-k)} - o(1). \end{aligned}$$

Consequently, if $n < \frac{\log(p-k)}{4\omega_{\mathrm{ave}}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}}$, then the probability of error remains bounded above by $1/2$, as claimed.

VI. 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 $X_i \sim N(0,\Sigma)$ drawn from Σ-Gaussian ensembles, including the standard Gaussian one ($\Sigma = I_{p\times p}$) as a special case. We have established both lower and upper bounds on the number of observations n as a function of the model dimension p, signal sparsity k, squared minimum value $\beta_{\min}^2$, and noise variance σ^2 as well as other parameters of the design covariance Σ that are required for asymptotically reliable recovery. In conjunction with previous

work [31] on the limits of the Lasso (ℓ_1-constrained quadratic programming), this analysis has some consequences.

(a) For signals β^* of bounded norm, the Lasso achieves the information-theoretically optimal order of scaling as a function of (n, p, k) and $\beta^2_{\min}$ (see Corollary 1), whereas
(b) for signals β^* with linear sparsity ($k = \alpha p$) and squared minimum value $\beta^2_{\min} = \Theta(\frac{\log k}{k})$, the Lasso is suboptimal (see Corollary 2).

There are a variety of open directions suggested by our analysis. First, while our upper and lower bounds are essentially matching for certain regimes of scaling, it is likely that the analysis can be tightened in other regimes. In particular, the necessary conditions stated in Theorem 2 certainly involve some slack, since they are obtained by analyzing restricted ensembles in which the value of β^* on the subset is known *a priori* to the decoder. It would be interesting to see if sharper results could be obtained via analysis of a less restrictive ensembles. Second, our work has revealed the suboptimality of current practical methods in the linear sparsity regime ($k = \alpha p$ for some $\alpha \in (0,1)$) for sufficiently high SNR (in particular, $\beta^2_{\min} = \Omega(\frac{\log k}{k})$). It is possible that multistage methods (e.g., [33], [23]) could be helpful in closing these gaps. Third, our results highlight various differences between the conditions on the design covariance matrix Σ (from which the random measurement matrices are generated) required by ℓ_1-based methods such as the Lasso, as contrasted with exponential-complexity methods. It would be interesting to see to what extent the mutual incoherence conditions that affect standard ℓ_1 methods can be relaxed; see Meinshausen and Yu [23] for some progress in this direction.

APPENDIX

A. Proof of Lemma 3

Using the linear observation model (2), we note that $\Pi^{\perp}_S(y) = \Pi^{\perp}_S(w)$, so that for any $\delta > 0$, we can write

$$\mathcal{E}_1(\delta) = \left\{ \frac{\left| ||\Pi^{\perp}_S(w)||_2^2 - ||\Pi^{\perp}_T(w)||_2^2 \right|}{\sigma^2} \geq \frac{d\delta}{2} \right\} \qquad (50)$$

where we have adopted the shorthand notation $d = n - k$. The following lemma characterizes the distribution of the random variable to be bounded.

Lemma 5: For any two k-sized subsets T and S with overlap $\ell = |S \backslash T| = |T \backslash S|$, we have

$$\frac{||\Pi^{\perp}_T(w)||_2^2 - ||\Pi^{\perp}_S(w)||_2^2}{\sigma^2} = \tilde{Z}_\ell - Z_\ell$$

where Z and $\tilde{Z}$ are chi-squared variables with ℓ degrees of freedom.

Proof: Note that by the Pythagorean Theorem for projections, we have

$$||\Pi^{\perp}_T(w)||_2^2 - ||\Pi^{\perp}_S(w)||_2^2 = ||\Pi_S(w)||_2^2 - ||\Pi_T(w)||_2^2.$$

Again using the Pythagorean Theorem, we can write

$$\begin{aligned} ||\Pi_S(w)||_2^2 &= ||\Pi_{S\cap T}\Pi_S w||^2 + ||(I - \Pi_{S\cap T})\Pi_S w||^2 \\ &= ||\Pi_{S\cap T} w||^2 + ||(\Pi_S - \Pi_{S\cap T})w||^2, \end{aligned}$$

where we have used the facts that $\Pi_T \Pi_{S\cap T} = \Pi_{S\cap T} = \Pi_{S\cap T}\Pi_T$. Since there is an analogous decomposition for $||\Pi_T(w)||_2^2$, we can write

$$\begin{aligned} ||\Pi_S(w)||_2^2 - ||\Pi_T(w)||_2^2 = ||(\Pi_S - \Pi_{S\cap T})w||^2 \\ - ||\big(\Pi_T - \Pi_{S\cap T}\big)w||^2. \end{aligned}$$

Now the matrix $\Pi_S - \Pi_{S\cap T}$ is a projection matrix with rank equal to $\{|S| - |S \cap T|\} = |S \backslash T| = \ell$, and similarly for the matrix $\Pi_S - \Pi_{S\cap T}$. Consequently, $||(\Pi_S - \Pi_{S\cap T})w||^2$ is distributed as χ^2_ℓ, and similarly for the second term.

Using this lemma and the decomposition (50), we may write

$$\mathbb{P}[\mathcal{E}_1(\delta)] = \mathbb{P}\left[|Z - \tilde{Z}| \geq \frac{d\delta}{2} \right].$$

By triangle inequality, we have $|Z - \tilde{Z}| \leq |Z - \ell| + |\tilde{Z} - \ell|$, so that by union bound

$$\begin{aligned} \mathbb{P}\left[|Z - \tilde{Z}| \geq \frac{d\delta}{2} \right] &\leq \mathbb{P}\left[|Z - \ell| \geq \frac{d\delta}{4} \right] + \mathbb{P}\left[|\tilde{Z} - \ell| \geq \frac{d\delta}{4} \right] \\ &= 2\mathbb{P}\left[\frac{|Z - \ell|}{\ell} \geq \frac{d\delta}{4\ell} \right] \end{aligned}$$

where $Z \sim \chi^2_\ell$. As long as $t = \frac{d\delta}{16\ell} \geq 1$, we may apply the chi-squared tail bound (56) to conclude that

$$\mathbb{P}[\mathcal{E}_1(\delta)] \leq 2\exp\left(-\ell \frac{d\delta}{16\ell} \right) = 2\exp\left(-\frac{d\delta}{16} \right)$$

as claimed.

To establish the validity of the choice $\delta^* = \frac{f(\beta^*{}_{S\backslash T})}{4}$, we note that

$$\begin{aligned} f(\beta^*{}_{S\backslash T}) &= ||\Gamma^{1/2}(T, S)\beta^*{}_{S\backslash T}||_2^2/\sigma^2 \\ &\geq \frac{\rho_S(\Sigma)\ell\beta^2_{\min}}{\sigma^2}. \end{aligned}$$

Consequently, we have

$$\begin{aligned} t^* &= \frac{(n-k)\frac{f(\beta^*{}_{S\backslash T})}{4}}{16\ell} \\ &\geq \frac{(n-k)\rho_S(\Sigma)\beta^2_{\min}}{64\sigma^2} \end{aligned}$$

so that it suffices to have $n - k \geq \frac{64}{\rho_S(\Sigma)\beta^2_{\min}/\sigma^2}$, as claimed.

B. Proof of Lemma 4

We begin by conditioning on w and X_T, and showing that the random variable $||\Pi^{\perp}_T(y)||_2^2$ follows a noncentral chi-squared distribution. By conditioning on X_U, we can decompose $X_{S\backslash T}$ into a linear prediction based on X_U and a zero-mean error term. In particular, we have

$$X_{S\backslash T} = X_T(\Sigma_{TT})^{-1}\Sigma_{T(S\backslash T)} + E_{S\backslash T}$$

where $E_{S\backslash T} \in \mathbb{R}^{n \times |S\backslash T|}$ is a Gaussian random matrix independent of X_T, with i.i.d. rows drawn from the zero-mean Gaussian

distribution with covariance matrix $\Gamma(T,S)$ from (8). Using this decomposition, we have

$$\Pi_T^{\perp}(y) = \Pi_T^{\perp}(X_S {\beta^*}_S + w) = \Pi_T^{\perp}(E_{S\setminus T}{\beta^*}_{S\setminus T} + w)$$

since the orthogonal projection $\Pi_T^{\perp}$ annihilates any terms in the column space X_T.

Let us diagonalize the orthogonal projection matrix $\Pi_T^{\perp}$, writing it as $\Pi_T^{\perp} = M^T D M$ where D is diagonal with $(n-k)$ ones, and k zeros, and M is a unitary matrix. With this transformation, we have

$$\begin{aligned}||\Pi_T^{\perp}(y) &= ||M^T D M(E_{S\setminus T}{\beta^*}_{S\setminus T} + w)||_2^2 \\ &= ||D(ME_{S\setminus T}{\beta^*}_{S\setminus T} + Mw)||_2^2\end{aligned}$$

since M is unitary. The random vector $E_{S\setminus T}{\beta^*}_{S\setminus T}$ has i.i.d. Gaussian entries with variance $\gamma^2 = ||\Gamma^{1/2}(T,S){\beta^*}_{S\setminus T}||_2^2$, so that multiplication by M leaves its distribution unchanged. We conclude that conditioned on X_T and w, the rescaled variable

$$Z(T) := \frac{||\Pi_T(y)||_2^2}{\gamma^2} \stackrel{d}{=} \left|\left| D\left(\frac{E_{S\setminus T}{\beta^*}_{S\setminus T}}{\gamma} + \frac{Mw}{\gamma}\right)\right|\right|_2^2.$$

Since the rescaled vector $\frac{E_{S\setminus T}{\beta^*}_{S\setminus T}}{\gamma}$ has i.i.d. $N(0,1)$ entries, Z has a noncentral chi-squared distribution with $n-k$ degrees of freedom, and noncentrality parameter

$$\nu := \frac{||DMw||_2^2}{\gamma^2} = \frac{||\Pi_T^{\perp} w||_2^2}{\sigma^2 f({\beta^*}_{S\setminus T})}.$$

Now the event $\mathcal{E}_2(\delta)$ can be expressed in terms of Z and ν as $\mathcal{E}_2(\delta) = \{\frac{Z-\nu}{d} \leq \frac{2\delta}{f({\beta^*}_{S\setminus T})}\}$, where we have introduced the convenient shorthand $d = n-k$. For the choice $\delta^* = \frac{f({\beta^*}_{S\setminus T})}{4}$, we have $\mathcal{E}_2(\delta^*) = \{\frac{Z-\nu}{d} \leq 1/2\}$. Consequently, setting $x = \frac{d^2}{16(d+2\nu)}$ in (58b), we obtain $\mathbb{P}[\mathcal{E}_2(\delta)] \leq \exp(-x)$, with

$$\begin{aligned}x &:= \frac{1}{16}\frac{d}{1+2\nu/d} \\ &= \frac{1}{16}\frac{(n-k)f({\beta^*}_{S\setminus T})}{f({\beta^*}_{S\setminus T}) + 2\frac{||\Pi_T^{\perp}(w)||^2}{\sigma^2(n-k)}}.\end{aligned} \tag{51}$$

Finally, let us define the event $\mathcal{A}(w) = \{\frac{||\Pi_T^{\perp}(w)||^2}{\sigma^2(n-k)})|| \leq 2\}$. For each fixed X, the variable $||\Pi_T^{\perp}(w)||^2/\sigma^2$ is (central) chi-squared variate with $(n-k)$ degrees of freedom, so that by the tail bound (57), we have $\mathbb{P}[\mathcal{A}^c(w)] \leq \exp(-(n-k)/16)$. Putting together the pieces, we have

$$\mathbb{P}[\mathcal{E}_2(\delta)] \leq \mathbb{P}[\mathcal{E}_2(\delta) \,|\, \mathcal{A}(w)] + \mathbb{P}[\mathcal{A}^c(w)].$$

Since the event $\mathcal{A}(w)$ is a function only of X_T and w, our earlier tail bound (51) may be applied. Moreover, conditioned on $\mathcal{A}(w)$, we have $x \geq \frac{1}{16}\frac{(n-k)f({\beta^*}_{S\setminus T})}{f({\beta^*}_{S\setminus T})+4}$, so that we obtain

$$\begin{aligned}\mathbb{P}[\mathcal{E}_2(t^*)] \leq \exp\left(-(n-k)\frac{f({\beta^*}_{S\setminus T})}{16(f({\beta^*}_{S\setminus T})+4)}\right) \\ + \exp\{-(n-k)/16\}\end{aligned}$$

so that we conclude that

$$\mathbb{P}[\mathcal{E}_2(t^*)] \leq 2\exp\left(-(n-k)\frac{f({\beta^*}_{S\setminus T})}{16(f({\beta^*}_{S\setminus T})+4)}\right) \tag{52}$$

as claimed.

C. Bounds on Binomial Coefficients

We make use of the following crude bounds on the binomial coefficients:

$$\left(\frac{p}{k}\right)^k \leq \binom{p}{k} \leq \left(\frac{pe}{k}\right)^k. \tag{53}$$

In addition, the following bound is also standard [8]:

$$\log\binom{p}{k} \leq ph\left(\frac{k}{p}\right) \tag{54}$$

where $h(t) := -t\log t - (1-t)\log(1-t)$ is the binary entropy function.

D. Tail Bounds for χ^2-Variates

The following large deviations bounds for centralized χ^2 are taken from Laurent and Massart [21]. Given a centralized χ^2-variate Z with m degrees of freedom, then for all $x \geq 0$

$$\mathbb{P}[Z - m \geq 2\sqrt{mx} + 2x] \leq \exp(-x) \quad \text{and} \tag{55a}$$

$$\mathbb{P}[Z - m \leq -2\sqrt{mx}] \leq \exp(-x). \tag{55b}$$

The following consequences of these bounds are useful in our analysis. First, for $t \geq 1$, we have

$$\mathbb{P}\left[\frac{|Z-m|}{m} \geq 4t\right] \leq \exp(-mt). \tag{56}$$

Starting with the bound (55a), setting $x = tm$ yields $\mathbb{P}[\frac{Z-m}{m} \geq 2\sqrt{t} + 2t] \leq \exp(-tm)$, Since $4t \geq 2\sqrt{t} + 2t$ for $t \geq 1$, we have $\mathbb{P}[\frac{Z-m}{m} \geq 4t] \leq \exp(-tm)$ for all $t \geq 1$. On the other hand, for all $t \geq 1$, we have $\mathbb{P}[\frac{Z-m}{m} \leq -4t] = 0$, so that the claim (56) follows. Secondly, we have

$$\mathbb{P}[Z/m \geq 2] \leq \exp(-m/16) \tag{57}$$

which follows by setting $x = m/16$ in (55a).

More generally, the analogous tail bounds for *noncentral* χ^2, taken from Birgé [4], can be established via the Chernoff technique, and careful bounding of the moment generating function. Let X be a noncentral χ^2 variable with d degrees of freedom and noncentrality parameter $\nu \geq 0$. Then for all $x > 0$

$$\mathbb{P}[X \geq (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x] \leq \exp(-x) \tag{58a}$$

$$\mathbb{P}[X \leq (d+\nu) - 2\sqrt{(d+2\nu)x}] \leq \exp(-x). \tag{58b}$$

ACKNOWLEDGMENT

The author wishes to thank Peter Bickel for helpful discussions and pointers, and the anonymous reviewers for careful

reading and helpful comments and suggestion that improved the presentation.

REFERENCES

[1] S. Aeron, M. Zhao, and S. Venkatesh, "Information-theoretic bounds to sensing capacity of sensor networks under fixed snr," in *Proc. IEEE Information Theory Workshop*, San Diego, CA, Sep. 2007.

[2] M. Akcakaya and V. Tarokh, "Shannon Theoretic Limits on Noisy Compressive Sampling," Harvard Univ., Cambridge, MA, Tech. Rep., Nov. 2007 [Online]. Available: arXiv:cs.IT:0711.0366

[3] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, to be published.

[4] L. Birgé, "An alternative point of view on Lepski's method," in *State of the Art in Probability and Statistics*, ser. IMS Lecture Notes, no. 37. Beachwood, OH: Inst. Math. Statist., 2001, pp. 113–133.

[5] E. Candés and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[6] E. Candés and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[7] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

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

[9] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices, and Banach spaces," in *Handbook of Banach Spaces*. Amsterdan, The Netherlands: Elsevier, 2001, vol. 1, pp. 317–336.

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

[11] D. Donoho, "Compressed sensing," *IEEE Trans. Inf.. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] D. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure and Appl. Math.*, vol. 59, no. 6, pp. 797–829, Jun. 2006.

[13] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Commun. Pure and Appl. Math.*, vol. 59, no. 7, pp. 907–934, Jul. 2006.

[14] D. L. Donoho and J. M. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, pp. 1–53, Jul. 2009.

[15] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and Sufficient Conditions on Sparsity Pattern Recovery," Univ. Calif., Berkeley, Tech. Rep., Apr. 2008 [Online]. Available: arXiv:cs.IT/0804.1839

[16] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *J. Appl. Signal Process.*, vol. 10, pp. 1–19, 2006.

[17] J. J. Fuchs, "Recovery of exact sparse representations in the presence of noise," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, QC, Canada, 2004, vol. 2, pp. 533–536.

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

[19] R. A. Horn and C. R. Johnson, *Matrix Analysis.*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

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

[21] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1303–1338, 1998.

[22] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, pp. 1436–1462, 2006.

[23] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Statist.*, to be published.

[24] A. J. Miller, *Subset Selection in Regression.*. New York: Chapman&Hall, 1990.

[25] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[26] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. Information Theory*, Toronto, ON, Canada, Jul. 2008, pp. 2187–2191.

[27] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements versus bits: Compressed sensing meets information theory," in *Proc. Allerton Conf. Communication , Control, and Computing*, Monticello, IL, Sep. 2006.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc.*, ser. B, vol. 58, no. 1, pp. 267–288, 1996.

[29] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[30] M. J. Wainwright, "Information-Theoretic Bounds for Sparsity Recovery in the High-Dimensional and Noisy Setting," Dep. Statist., Univ. Calif., Berkeley, Tech. Rep. 725, Jan. 2007 [Online]. Available: arxiv:math.ST/0702301, presented at the IEEE Int. Symp. Information Theory, Nice, France, June 2007

[31] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[32] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-Theoretic Limits on Sparse Signal Recovery: Dense Versus Sparse Measurement Matrices," Univ. Calif., Berkeley, Tech. Rep., June 2008 [Online]. Available: arXiv:0806.0604, presented at the IEEE Int. Symp. Information Theory, Toronto, ON, Canada, Jul. 2008

[33] L. Wasserman and K. Roeder, "Multi-stage variable selection: Screen and clean," *Ann. Statist.*, to be published.

[34] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

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

[36] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2567, 2006.

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

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

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