Sharp thresholds for high-dimensional and noisy recovery of sparsity

Martin J. Wainwright
Department of Statistics, and
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
wainwrig@{eecs,stat}.berkeley.edu

Technical Report, UC Berkeley, Department of Statistics
May 2006

Abstract

The problem of consistently estimating the sparsity pattern of a vector $\beta^* \in \mathbb{R}^p$ based on observations contaminated by noise arises in various contexts, including subset selection in regression, structure estimation in graphical models, sparse approximation, and signal denoising. We analyze the behavior of ℓ_1-constrained quadratic programming (QP), also referred to as the Lasso, for recovering the sparsity pattern. Our main result is to establish a sharp relation between the problem dimension p, the number s of non-zero elements in β^*, and the number of observations n that are required for reliable recovery. For a broad class of Gaussian ensembles satisfying mutual incoherence conditions, we establish existence and compute explicit values of thresholds θ_ℓ and θ_u with the following properties: for any $\nu > 0$, if $n > 2\,(\theta_u + \nu) \log(p-s) + s + 1$, then the Lasso succeeds in recovering the sparsity pattern with probability converging to one for large problems, whereas for $n < 2\,(\theta_\ell - \nu) \log(p-s) + s + 1$, then the probability of successful recovery converges to zero. For the special case of the uniform Gaussian ensemble, we show that $\theta_\ell = \theta_u = 1$, so that the threshold is sharp and exactly determined.

Keywords: Quadratic programming; Lasso; subset selection; consistency; thresholds; sparse approximation; signal denoising; sparsity recovery; ℓ_0-regularization; model selection.

1 Introduction

The problem of recovering the sparsity pattern of an unknown vector β^*—that is, the positions of the non-zero entries of β^*— based on noisy observations arises in a broad variety of contexts, including subset selection in regression [29], structure estimation in graphical models [28], sparse approximation [8, 30], and signal denoising [6]. A natural optimization-theoretic formulation of this problem is via ℓ_0-minimization, where the ℓ_0 "norm" of a vector corresponds to the number of non-zero elements. Unfortunately, however, ℓ_0-minimization problems are known to be NP-hard in general [30], so that the existence of polynomial-time algorithms is highly unlikely. This challenge motivates the use of computationally tractable approximations or relaxations to ℓ_0 minimization. In particular, a great deal of research over the past decade has studied the use of the ℓ_1-norm as a computationally tractable surrogate to the ℓ_0-norm.

In more concrete terms, suppose that we wish to estimate an unknown but fixed vector $\beta^* \in \mathbb{R}^p$ on the basis of a set of n observations of the form

$$Y_k = x_k^T \beta^* + W_k, \qquad k = 1, \ldots n, \tag{1}$$

where $x_k \in \mathbb{R}^p$, and $W_k \sim N(0, \sigma^2)$ is additive Gaussian noise. In many settings, it is natural to assume that the vector β^* is *sparse*, in that its *support*

$$S := \{i \in \{1, \ldots p\} \mid \beta_i^* \neq 0\} \tag{2}$$

has relatively small cardinality $s = |S|$. Given the observation model (1) and sparsity assumption (2), a reasonable approach to estimating β^* is by solving the ℓ_1-constrained quadratic program (QP)

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{k=1}^{n} \|Y_k - x_k^T \beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}, \tag{3}$$

where $\lambda_n \geq 0$ is a regularization parameter. Of interest are conditions on the *ambient dimension* p, the *sparsity index* s, and the *number of observations* n for which it is possible (or impossible) to recover the support set S of β^*.

1.1 Overview of previous work

Given the substantial literature on the use of ℓ_1 constraints for sparsity recovery and subset selection, we provide only a very brief (and hence necessarily incomplete) overview here. In the *noiseless version* ($\sigma^2 = 0$) of the linear observation model (1), one can imagine estimating β^* by solving the problem

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \qquad \text{subject to} \quad x_k^T \beta = Y_k, \quad k = 1, \ldots, n. \tag{4}$$

This problem is in fact a linear program (in disguise), and corresponds to a method in signal processing known as basis pursuit, pioneered by Chen et al. [6]. For the noiseless setting, the interesting regime is the underdetermined setting (i.e., $n < p$). With contributions from a broad range of researchers [e.g., 3, 6, 12, 10, 14, 15, 26, 33], there is now a fairly complete understanding of conditions on deterministic vectors $\{x_k\}$ and sparsity index s for which the true solution β^* can be recovered exactly. Without going into technical details, the rough idea is that the *mutual incoherence* of the vectors $\{x_k\}$ must be large relative to the sparsity index s, and indeed we impose similar conditions to derive our results (e.g., conditions (14a) and (18) in the sequel). Most closely related to the current paper—as we discuss in more detail in the sequel—are recent results by Donoho [9], as well as Candes and Tao [4] that provide high probability results for random ensembles. More specifically, as independently established by both sets of authors using different methods, for uniform Gaussian ensembles (i.e., $x_k \sim N(0, I_p)$) with the ambient dimension p scaling linearly in terms of the number of observations (i.e., $p = \gamma n$, for some $\gamma > 1$), there exists a constant $\alpha > 0$ such that all sparsity patterns with $s \leq \alpha p$ can be recovered with high probability.

There is also a substantial body of work focusing on the noisy setting ($\sigma^2 > 0$), and the use of quadratic programming techniques for sparsity recovery [e.g., 6, 17, 18, 34, 11, 16, 28, 35]. The ℓ_1-constrained quadratic program (3), also known as the Lasso [32, 13], has been the focus of considerable research in recent years. Knight and Fu [23] analyze the asymptotic behavior of the optimal solution, not only for ℓ_1 regularization but for ℓ_p-regularization with $p \in (0, 2]$. Fuchs [17, 18] investigates optimality conditions for the constrained QP (3), and provides deterministic conditions, of the mutual incoherence form, under which a sparse solution, which is known to be within ϵ of the observed values, can be recovered exactly. Among a variety of other results, both Tropp [34] and Donoho et al. [11] also provide sufficient conditions for the support of the optimal solution to the constrained QP (3) to be contained within the true support of β^*. Most directly related to the current paper is recent work by both Meinshausen and Buhlmann [28], focusing on Gaussian noise, and extensions by Zhao and Yu [35] to more general noise distributions, on the use of the Lasso for model selection. For the case of Gaussian noise, both papers established that under mutual incoherence conditions and appropriate choices of the regularization parameter λ_n, the Lasso can recover the sparsity pattern with probability converging to one for particular regimes of n, p and s, when x_k drawn randomly from random Gaussian ensembles. We discuss connections to our results at more length in the the sequel.

1.2 Our contributions

Recall the linear observation model (1). 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:

$$X \;:=\; \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \;=\; \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}. \tag{5}$$

Consider the (random) set $\mathcal{S}(X, \beta^*, W, \lambda_n)$ of optimal solutions to this constrained quadratic program (3). By convexity and boundedness of the cost function, the solution set is always non-empty. For any vector $\beta \in \mathbb{R}^p$, we define the sign function

$$\operatorname{sgn}(\beta_i) \;:=\; \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0. \end{cases} \tag{6}$$

Of interest is the event that the Lasso (3) succeeds in recovering the sparsity pattern of the unknown β^*:

Property $\mathcal{R}(X, \beta^*, W, \lambda_n)$: There exists an optimal solution $\widehat{\beta} \in \mathcal{S}(X, \beta^*, W, \lambda_n)$ with the property $\operatorname{sgn}(\widehat{\beta}) = \operatorname{sgn}(\beta^*)$.

Our main result is that for a broad class of random Gaussian ensembles based on covariance matrices satisfying mutual incoherence conditions, there exist fixed constants $0 < \theta_\ell \leq 1$ and $1 \leq \theta_u < +\infty$ such that for all $\nu > 0$, property $\mathcal{R}(X, \beta^*, W, \lambda_n)$ holds with high probability (over the choice of noise vector W and random matrix X) whenever

$$n \;>\; 2(\theta_u + \nu)\, s \, \log(p - s) + s + 1, \tag{7}$$

and *conversely*, fails to hold with high probability whenever

$$n \;<\; 2(\theta_\ell - \nu)\, s \, \log(p - s) + s + 1. \tag{8}$$

Moreover, for the special case of the uniform Gaussian ensemble (i.e., $x_k \sim N(0, I_p)$), we show that $\theta_\ell = \theta_u = 1$, so that the threshold is sharp. This threshold result has a number of connections to previous work in the area that focuses on special forms of scaling. More specifically, as we discuss in more detail in Section 3.2, in the special case of linear scaling (i.e., $n = \gamma p$ for some $\gamma > 0$), this theorem provides a noisy analog of results previously established for basis pursuit in the noiseless case [9, 4]. Moreover, our result can also be adapted to an entirely different scaling regime for n, p and s, as considered by a separate body of recent work [28, 35] on the high-dimensional Lasso.

The remainder of this paper is organized as follows. We begin in Section 2 with some necessary and sufficient conditions, based on standard optimality conditions for convex programs, for property $\mathcal{R}(X, \beta^*, W, \lambda_n)$ to hold. We then prove a consistency result for the case of deterministic design matrices X. Section 3 is devoted to the statement and proof of our main result on the asymptotic behavior of the lasso for random Gaussian ensembles. We illustrate this result via simulation in Section 4, and conclude with a discussion in Section 5.

2 Some preliminary analysis

In this section, we provide necessary and sufficient conditions for property $\mathcal{R}(X, \beta^*, W, \lambda_n)$ to hold. Based on these conditions, we then define collections of random variables that play a central role in our analysis. In particular, the study of $\mathcal{R}(X, \beta^*, W, \lambda_n)$ is reduced to the study of the extreme order statistics of these random variables. We then state and prove a result about the behavior of the Lasso for the case of a deterministic design matrix X.

2.1 Necessary and sufficient conditions

We begin with a simple set of necessary and sufficient conditions for property $\mathcal{R}(X, \beta^*, W, \lambda_n)$ to hold. We note that this result is not essentially new (e.g., see [17, 18, 28, 34, 35] for variants), and follows in a straightforward manner from optimality conditions for convex programs [21]; see Appendix A for further details. We define $S := \{i \in \{1, \ldots, p\} \mid \beta_i^* \neq 0\}$ to be the support of β^*, and let S^c be its complement. For any subset $T \subseteq \{1, 2, \ldots, p\}$, let X_T be the $n \times |T|$ matrix with the vectors $\{X_i, i \in T\}$ as columns.

Lemma 1. *Assume that the matrix $X_S^T X_S$ is invertible. Then, for any given $\lambda > 0$ and noise vector $w \in \mathbb{R}^n$, property $\mathcal{R}(X, \beta^*, w, \lambda_n)$ holds if and only if*

$$\left| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \left[\frac{1}{n} X_S^T w - \lambda \operatorname{sgn}(\beta_S^*)\right] - \frac{1}{n} X_{S^c}^T w \right| \leq \lambda, \quad \textit{and} \tag{9a}$$

$$\left| \beta_S^* + \left(\frac{1}{n} X_S^T X_S\right)^{-1} \left[\frac{1}{n} X_S^T w - \lambda \operatorname{sgn}(\beta_S^*)\right] \right| > 0, \tag{9b}$$

where both of these vector inequalities should be taken elementwise.

For shorthand, define $\vec{b} := \operatorname{sgn}(\beta_S^*)$, and denote by $e_i \in \mathbb{R}^s$ the vector with 1 in the i^{th} position, and zeroes elsewhere. Motivated by Lemma 1, much of our analysis is based on the collections of random variables, defined each index $i \in S$ and $j \in S^c$ as follows:

$$U_i := e_i^T \left(\frac{1}{n} X_S^T X_S\right)^{-1} \left[\frac{1}{n} X_S^T W - \lambda_n \vec{b}\right] \tag{10a}$$

$$V_j := X_j^T \left\{ X_S \left(X_S^T X_S\right)^{-1} \lambda_n \vec{b} - \left[X_S \left(X_S^T X_S\right)^{-1} X_S^T - I_{n \times n}\right] \frac{W}{n} \right\}. \tag{10b}$$

Recall that $s = |S|$ and $N = |S^c| = p - s$. From Lemma 1, the behavior of $\mathcal{R}(X, \beta^*, W, \lambda_n)$ is determined by the behavior of $\max_{j \in S^c} |V_j|$ and $\max_{i \in S} |U_i|$. In particular, condition (9a) holds if and only if the event

$$\mathcal{M}(V) := \left\{ \max_{j \in S^c} |V_j| \leq \lambda_n \right\} \tag{11}$$

holds. On the other hand, if we define $\rho_n := \min_{i \in S} |\beta_i^*|$, then the event

$$\mathcal{M}(U) := \left\{ \max_{i \in S} |U_i| \leq \rho_n \right\} \tag{12}$$

is sufficient to guarantee that condition (9b) holds. Consequently, our proofs are based on analyzing the asymptotic probability of these two events.

2.2 Recovery of sparsity: deterministic design

We now show how Lemma 1 can be used to analyze the behavior of the Lasso for the special case of a deterministic (non-random) design matrix X. To gain intuition for the conditions in the theorem statement, it is helpful to consider the *zero-noise condition* $w = 0$, in which each observation $Y_k = x_k^T \beta^*$ is uncorrupted. In this case, the conditions of Lemma 1 reduce to

$$\left| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \mathrm{sgn}(\beta_S^*) \right| \leq 1 \tag{13a}$$

$$\left| \beta_S^* - \lambda \left(\frac{1}{n} X_S^T X_S\right)^{-1} \mathrm{sgn}(\beta_S^*) \right| > 0. \tag{13b}$$

Of course, if the conditions of Lemma 1 fail to hold in the zero-noise setting, then there is little hope of succeeding in the presence of noise.

The zero-noise conditions motivate imposing the following set of conditions on the design matrix:

$$\left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \right\|_\infty \leq (1 - \epsilon) \quad \text{for some } \epsilon \in (0, 1], \text{ and} \tag{14a}$$

$$\Lambda_{min}\left(\frac{1}{n} X_S^T X_S\right) \geq C_{min} > 0, \tag{14b}$$

where Λ_{min} denotes the minimal eigenvalue. Under these conditions, we have the following:

Proposition 1. *Suppose that we observe $Y = X\beta^* + W$, where each column X_j of X is normalized to ℓ_2-norm n, and $W \sim N(0, \sigma^2 I)$. Assume β^* and X satisfy conditions* (14)*, and define $\rho_n := \min_{i \in S} |\beta_i^*|$. If $\lambda_n \to 0$ is chosen such that*

$$(a) \quad \frac{n\lambda_n^2}{\log(p-s)} \to +\infty, \qquad and \quad (b) \quad \frac{1}{\rho_n}\left\{\sqrt{\frac{\log s}{n}} + \lambda_n \, \|(\frac{1}{n}X_S^T X_S)^{-1}\|_\infty\right\} \to 0, \tag{15}$$

then $\mathbb{P}(\mathcal{R}(X, \beta^*, W, \lambda_n) \to 1$ as $n \to +\infty$.

Before proving the proposition, we pause to make a number of comments. First, conditions of the form (14a) have been considered in previous work on the lasso [17, 18, 28, 34, 35]. In particular, various authors [34, 28, 35] provide examples and results on matrix families that satisfy this type of condition. Moreover, previous work [28, 35] provides asymptotic results for particular scalings of p, s and n for random design matrices, as we discuss in more detail in Section 3. To the best of our knowledge, Proposition 1 is the first result to provide sufficient conditions for exact recovery in deterministic designs with general scaling of p, s and n.

Second, it is worthwhile to consider Proposition 1 in the classical setting (i.e., in which the number of samples $n \to +\infty$ with p and s remaining fixed). In this setting, the quantity $\rho_n = \min_{i \in S} |\beta_i^*|$ does not depend on n. Hence, in addition to the condition (14), the requirements reduce to $\lambda_n \to 0$ and $n\lambda_n^2 \to +\infty$. Note that $\lambda_n = \frac{\log n}{\sqrt{n}}$ is one suitable choice. This classical case is also covered by previous work [23, 28, 35].

Last, consider the more general setting where all three parameters (n, p, s) grow to infinity, and suppose for simplicity that ρ_n stays bounded away from 0. The conditions $\lambda_n^2 \to 0$ and $\lambda_n^2 \frac{n}{\log(p-s)} \to +\infty$ imply that the number of observations n must grow at a rate faster than $\log(p-s)$. In the following section, in which we consider the more general case of random Gaussian ensembles, we will see that for ensembles satisfying mutual incoherence conditions, we in fact require that $\frac{n}{\log(p-s)} = \Theta(s) \to +\infty$.

2.3 Proof of Proposition 1

Recall the events $\mathcal{M}(V)$ and $\mathcal{M}(U)$ defined in equations (11) and (12) respectively. To establish the claim, we must show that that $\mathbb{P}[\mathcal{M}(V)^c$ or $\mathcal{M}(U)^c] \to 0$, where $\mathcal{M}(V)^c$ and $\mathcal{M}(U)^c$ denote the complements of these events. By union bound, it suffices to show both $\mathbb{P}[\mathcal{M}(V)^c]$ and $\mathbb{P}[\mathcal{M}(U)^c]$ converge to zero, or equivalently that $\mathbb{P}[\mathcal{M}(V)]$ and $\mathbb{P}[\mathcal{M}(U)]$ both converge to one.

Analysis of $\mathcal{M}(V)$: We begin by establishing that $\mathbb{P}[\mathcal{M}(V)] \to 1$. Throughout the proof, we use the shorthand $\vec{b} := \mathrm{sgn}(\beta^*)$ and $N := p - s = |S^c|$.

Recalling the definition (10b) of the random variables V_j, note that $\mathcal{M}(V)$ holds holds if and only $\frac{\min_{j \in S^c} V_j}{\lambda_n} \geq -1$ and $\frac{\max_{j \in S^c} V_j}{\lambda_n} \leq 1$. Moreover, we note that each V_j is Gaussian with mean

$$\mu_j \;=\; \mathbb{E}[V_j] \;\;=\;\; \lambda_n X_j^T X_S \left(X_S^T X_S\right)^{-1} \vec{b}\,.$$

Using condition (14a), we have $|\mu_j| \leq (1-\epsilon)\,\lambda_n$ for all indices $j = 1, \ldots, N$, from which we obtain that

$$\frac{\max_{j \in S^c} V_j}{\lambda_n} \;\leq\; (1-\epsilon) + \frac{1}{\lambda_n}\max_j \widetilde{V}_j, \qquad \text{and} \qquad \frac{\min_{j \in S^c} V_j}{\lambda_n} \;\geq |; -(1-\epsilon) + \frac{1}{\lambda_n}\min_j \widetilde{V}_j,$$

where $\widetilde{V}_j := X_j^T \left[I_{n\times n} - X_S \left(X_S^T X_S\right)^{-1} X_S^T\right] W$ are zero-mean (correlated) Gaussian variables. Hence, in order to establish condition (9a) of Lemma 1, we need to show that

$$\mathbb{P}\left[\frac{1}{\lambda_n}\min_{j\in S^c}\widetilde{V}_j < -\epsilon, \quad \text{or} \quad \frac{1}{\lambda_n}\max_{j\in S^c}\widetilde{V}_j > \epsilon\right] \to 0. \tag{16}$$

In fact, using Lemma 11 (see Appendix C), it is sufficient to show that $\mathbb{P}[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_n} > \epsilon] \to 0$. By applying Markov's inequality and Gaussian comparison results [25] (see Lemma 9 in Appendix B), we obtain

$$\mathbb{P}\left[\frac{\max_{j\in S^c}|\widetilde{V}_j|}{\lambda_n} > \epsilon\right] \;\leq\; \frac{\mathbb{E}[\max_{j\in S^c}|\widetilde{V}_j|]}{\lambda_n} \;\leq\; \frac{3\sqrt{\log N}}{\lambda_n}\max_j\sqrt{\mathbb{E}[\widetilde{V}_j^2]}.$$

Straightforward computation yields that

$$\mathbb{E}[\widetilde{V}_j^2] = \frac{\sigma^2}{n^2}\, X_j^T\left[I_{n\times n} - X_S\left(X_S^T X_S\right)^{-1} X_S^T\right] X_j \;\leq \frac{\sigma^2}{n^2}\|X_j\|^2 \;=\; \frac{\sigma^2}{n},$$

since the matrix $I_{n\times n} - X_S\left(X_S^T X_S\right)^{-1} X_S^T$ has maximum eigenvalue equal to one, and $\|X_j\|_2^2 = n$ by construction. Consequently, condition (a) in the theorem statement—namely, that $\frac{\log N}{n\lambda_n^2} \to 0$ is sufficient to ensure that $\mathbb{E}[\widetilde{V}_{(N)}]/\lambda_n \to 0$. Thus, we have established $\mathbb{P}(\mathcal{M}(V)) \to 1$ (i.e., that condition (9a) holds w.p. one as $n \to +\infty$).

Analysis of $\mathcal{M}(U)$: We now show that $\mathbb{P}(\mathcal{M}(U)) \to 1$. Beginning with the triangle inequality, we upper bound $\max_i |U_i| := \|(\frac{1}{n}X_S^T X_S)^{-1}[\frac{1}{n}X_S^T W - \lambda_n \operatorname{sgn}(\beta_S^*)]\|_\infty$ as

$$\max_i |U_i| \;\leq\; \left\|(\frac{1}{n}X_S^T X_S)^{-1}\frac{1}{n}X_S^T W\right\|_\infty + \left\|(\frac{1}{n}X_S^T X_S)^{-1}\right\|_\infty \lambda_n$$

Let e_i denote the unit vector with one in position i and zeroes elsewhere. Now define, for each index $i \in S$, the Gaussian random variable $Z_i := e_i^T(\frac{1}{n}X_S^T X_S)^{-1}\frac{1}{n}X_S^T W$. Each such Z_i is a zero-mean Gaussian with variance given by

$$\operatorname{var}(Z_i) \;=\; \frac{\sigma^2}{n}e_i^T(\frac{1}{n}X_S^T X_S)^{-1}e_i \;\leq\; \frac{\sigma^2}{C_{min}n}$$

Hence, by a standard Gaussian comparison theorem [25] (in particular, see Lemma 9 in Appendix B), we have

$$\begin{aligned}\mathbb{E}[\max_{1\leq i\leq s}|Z_i|] \;&=\; \mathbb{E}\left[\left\|(\frac{1}{n}X_S^T X_S)^{-1}\frac{1}{n}X_S^T W\right\|_\infty\right] \\ &\leq\; 3\sqrt{\frac{\sigma^2\log s}{nC_{min}}}.\end{aligned}$$

Thus, recalling the defining $\rho_n := \min_{i \in S} |\beta_i^*|$, we apply Markov's inequality to conclude that

$$
\begin{aligned}
1 - \mathbb{P}\left[\left|\beta_S^* + \left(\frac{1}{n} X_S^T X_S\right)^{-1}\left[\frac{1}{n} X_S^T w - \lambda \operatorname{sgn}(\beta_S^*)\right]\right| > 0\right] &\leq \mathbb{P}\left[\frac{1}{\rho_n} \max_{1 \leq i \leq s} |U_i| > 1\right] \\
&\leq \mathbb{P}\left[\frac{1}{\rho_n}\left\{\max_{1 \leq i \leq s} |Z_i| + \lambda_n \|(\frac{1}{n} X_S^T X_S)^{-1}\|_\infty\right\} > 1\right] \\
&\leq \frac{1}{\rho_n}\left\{\mathbb{E}\left[\max_{1 \leq i \leq s} |Z_i|\right] + \lambda_n \|(\frac{1}{n} X_S^T X_S)^{-1}\|_\infty\right\} \\
&\leq \frac{1}{\rho_n}\left\{3\sqrt{\frac{\sigma^2 \log s}{n C_{min}}} + \lambda_n \|(\frac{1}{n} X_S^T X_S)^{-1}\|_\infty\right\},
\end{aligned}
$$

which converges to zero as $n \to +\infty$, using condition (b) in the theorem statement. □

3 Recovery of sparsity: random Gaussian ensembles

We now turn to the analysis of random design matrices X, in which each row x_k is chosen as an i.i.d. Gaussian random vector with covariance matrix Σ. In particular, we prove the existence of thresholds that provide a sharp description of the failure/success of the Lasso as a function of (n, p, s). We begin by setting up and providing a precise statement of the main result, and then discussing its connections to previous work. In the later part of this section, we provide the proof.

3.1 Statement of main result

Consider a covariance matrix Σ with unit diagonal, and with its minimum and maximum eigenvalues (denoted Λ_{min} and Λ_{max} respectively) bounded as

$$\Lambda_{min}(\Sigma_{SS}) \geq C_{min}, \qquad \text{and} \qquad \Lambda_{max}(\Sigma) \leq C_{max} \tag{17}$$

for constants $C_{min} > 0$ and $C_{max} < +\infty$. Given a vector $\beta^* \in \mathbb{R}^p$, define its support $S = \{i \in \{1, \ldots, p\} \mid \beta_i^* \neq 0\}$, as well as the complement S^c of its support. Suppose that Σ and S satisfy the conditions $\|(\Sigma_{SS})^{-1}\|_\infty \leq D_{\max}$ for some $D_{\max} < +\infty$, and

$$\|\Sigma_{S^c S}(\Sigma_{SS})^{-1}\|_\infty \quad \leq \quad (1 - \epsilon) \tag{18}$$

for some $\epsilon \in (0, 1]$. Under these conditions, we consider the observation model

$$Y_k \quad = \quad x_k^T \beta^* + W_k, \qquad k = 1, \ldots, n, \tag{19}$$

where $x_k \sim N(0, \Sigma)$ and $W_k \sim N(0, \sigma^2)$ are independent Gaussian variables for $k = 1, \ldots, n$. Furthermore, we define $\rho_n := \min_{i \in S} |\beta_i^*|$, and the sparsity index $s = |S|$.

Theorem 1. *Consider a sequence of covariance matrices $\{\Sigma[p]\}$ and solution vectors $\{\beta^*[p]\}$ satisfying conditions* (17) *and* (18)*. Under the observation model* (19)*, consider a sequence $(n, p(n), s(n))$ such that s, $(n - s)$ and $(p - s)$ tend to infinity. Define the thresholds*

$$\theta_\ell \;:=\; \frac{(\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}})^2}{C_{max}\,(2 - \epsilon)^2} \leq 1, \qquad \textit{and} \qquad \theta_u \;:=\; \frac{C_{max}}{\epsilon^2 C_{min}} \geq 1. \tag{20}$$

Then for any constant $\nu > 0$, we have the following *(a) If $n < 2(\theta_\ell - \nu)\, s \log(p-s) + s + 1$, then $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \to 0$ for any non-increasing sequence $\lambda_n > 0$.*

(b) Conversely, if $n > 2(\theta_u + \nu)\, s \log(p-s) + s$, and $\lambda_n \to 0$ is chosen such that

$$\frac{n\lambda_n^2}{\log(p-s)} \to +\infty, \qquad \text{and} \qquad \frac{1}{\rho_n}\left[\lambda_n + \sqrt{\frac{\log s}{n}}\right] \;\to\; 0, \tag{21}$$

then $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \to 1$.

Remark: Suppose for simplicity that ρ_n remains bounded away from 0. In this case, the requirements on λ_n reduce to $\lambda_n \to 0$, and $\lambda_n^2 n / \log(p-s) \to +\infty$. One suitable choice is $\lambda_n^2 = \frac{\log(s)\, \log(p-s)}{n}$, with which we have

$$\lambda_n^2 \;=\; \left(\frac{s\,\log(p-s)}{n}\right)\frac{\log(s)}{s} \;=\; O\left(\frac{\log s}{s}\right) \;\to\; 0,$$

and

$$\frac{n\lambda_n^2}{\log(p-s)} \;=\; \log(s) \to +\infty.$$

Without a bound on ρ_n, the second condition in equation (21) constrains the rate of decrease of the minimum $\rho_n = \min_{i \in S} |\beta_i^*|$.

3.2 Some consequences

To develop intuition for this result, we begin by stating certain special cases as corollaries, and discussing connections to previous work.

3.2.1 Uniform Gaussian ensembles

First, we consider the special case of the uniform Gaussian ensemble, in which $\Sigma = I_{p \times p}$. Previous work by Donoho [9] as well as Candes and Tao [4] has focused on the uniform Gaussian ensemble in the the noiseless ($\sigma^2 = 0$) and underdetermined setting ($n = \gamma p$ for some $\gamma \in (0,1)$). Analyzing the asymptotic behavior of the linear program (4) for recovering β^*, the basic result is that there exists some $\alpha > 0$ such that all sparsity patterns with $s \leq \alpha p$ can be recovered with high probability.

Applying Theorem 1 to the noisy version of this problem, the uniform Gaussian ensemble means that we can choose $\epsilon = 1$, and $C_{min} = C_{max} = 1$, so that the threshold constants reduce

$$\theta_\ell \;=\; \frac{(\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}})^2}{C_{max}\,(2-\epsilon)^2} \;=\; 1 \qquad \text{and} \qquad \theta_u \;=\; \frac{C_{max}}{\epsilon^2 C_{min}} \;=\; 1.$$

Consequently, Theorem 1 provides a sharp threshold for the behavior of the Lasso, in that failure/success is entirely determined by whether or not $n > 2s\,\log(p-s) + s + 1$. Thus, if we consider the particular linear scaling analyzed in previous work on the noiseless case [9, 4], we have:

Corollary 1 (Linearly underdetermined setting). *Suppose that $n = \gamma p$ for some $\gamma \in (0,1)$. Then* *(a) If $s = \alpha p$ for any $\alpha \in (0,1)$, then $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \to 0$ for any positive sequence $\lambda_n > 0$.*

(b) On the other hand, if $s = O(\frac{p}{\log p})$, then $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \to 1$ for any sequence $\{\lambda_n\}$ satisfying the conditions of Theorem 1(a).

Conversely, suppose that the size s of the support of β^* scales linearly with the number of parameters p. The following result describes the amount of data required for the ℓ_1-constrained QP to recover the sparsity pattern in the noisy setting ($\sigma^2 > 0$):

Corollary 2 (Linear fraction support). *Suppose that $s = \alpha p$ for some $\alpha \in (0,1)$. Then we require $n > 2\alpha p \log[(1-\alpha)\,p] + \alpha p$ in order to obtain exact recovery with probability converging to one for large problems.*

These two corollaries establish that there is a significant difference between recovery using basis pursuit (4) in the noiseless setting versus recovery using the Lasso (3) in the noisy setting. When the amount of data n scales only linearly with ambient dimension p, then the presence of noise means that the recoverable support size drops from a linear fraction (i.e., $s = \alpha p$ as in the work [9, 4]) to a sublinear fraction (i.e., $s = O(\frac{\log p}{p})$, as in Corollary 1).

3.2.2 Non-uniform Gaussian ensembles

We now consider more general (non-uniform) Gaussian ensembles that satisfy conditions (17) and (18). As mentioned earlier, previous papers by both Meinshausen and Buhlmann [28] as well as Zhao and Yu [35] treat model selection with the high-dimensional Lasso. For suitable covariance matrices (e.g., satisfying conditions (17) and (18)), both sets of authors proved that the sparsity pattern can be recovered exactly under scaling conditions of the form

$$s = O(n^{c_1}), \qquad \text{and} \quad p = O(e^{n^{c_2}}), \qquad \text{where} \quad c_1 + c_2 < 1. \tag{22}$$

Applying Theorem 1 in this scenario, we have the following:

Corollary 3. *Under the scaling (22), the Lasso will recover the sparsity pattern with probability converging to one.*

Proof. Substituting the conditions (22) into the threshold condition (7), we obtain that the RHS takes the form

$$\begin{aligned} 2s\log(p-s) + s + 1 &= O(n^{c_1}) \log\left[O(e^{n^{c_2}}) - O(n^{c_1})\right] + O(n^{c_1}) \\ &= O(n^{c_1+c_2}) \;\ll\; n, \end{aligned}$$

since $c_1 + c_2 < 1$ by assumption. Thus, we see that under these conditions, our threshold condition (7) is satisfied *a fortiori*. ☐

In fact, under this stronger scaling (22), both papers [28, 35] proved that the probability of exact recovery converges to one at a rate exponential in some polynomial function of n. Interestingly, our results show that the Lasso can recover the sparsity pattern for a much broader range of (n, p, s) scaling.

3.3 Proof of Theorem 1(b)

We now turn to the proof of part (b) of our main result. As with the proof of Proposition 1, the proof is based on analyzing the collections of random variables $\{V_j \mid j \in S^c\}$ and $\{U_i \mid i \in S\}$, as defined in equations (10a) and (10b) respectively. We begin with some preliminary results that serve to set up the argument.

3.3.1 Some preliminary results

We first note that for $s < n$, the random Gaussian matrix X_S will have rank s with probability one, whence the matrix $X_S^T X_S$ is invertible with probability one. Accordingly, the necessary and sufficient conditions of Lemma 1 are applicable. Our first lemma, proved in Appendix D.1, concerns the behavior of the random vector $V = (V_1, \ldots, V_N)$, when conditioned on X_S and W. Recalling the shorthand notation $\vec{b} := \operatorname{sgn}(\beta^*)$, we summarize in the following

Lemma 2. *Conditioned on X_S and W, the random vector $(V \mid W, X_S)$ is Gaussian. Its mean vector is upper bounded as*

$$|\mathbb{E}[V \mid W, X_S]| \quad \leq \quad \lambda_n (1-\epsilon)\, \mathbf{1}. \tag{23}$$

Moreover, its conditional covariance takes the form

$$\operatorname{cov}[V \mid W, X_S] \quad = \quad M_n \Sigma_{(S^c \,|\, S)} \;=\; M_n \left[\Sigma_{S^c S^c} - \Sigma_{S^c S} (\Sigma_{SS})^{-1} \Sigma_{S S^c}\right], \tag{24}$$

where

$$M_n \quad := \quad \lambda_n^2 \vec{b}^T (X_S^T X_S)^{-1} \vec{b} + \frac{1}{n^2} W^T \left[I_{n \times n} - X_S \left(X_S^T X_S\right)^{-1} X_S^T \right] W \tag{25}$$

is a random scaling factor.

The following lemma, proved in Appendix D.2, captures the behavior of the random scaling factor M_n defined in equation (25):

Lemma 3. *The random variable M_n has mean*

$$\mathbb{E}[M_n] \quad = \quad \frac{\lambda_n^2}{n-s-1} \vec{b}^T (\Sigma_{SS})^{-1} \vec{b} + \frac{\sigma^2\,(n-s)}{n^2}. \tag{26}$$

Moreover, it is sharply concentrated in that for any $\delta > 0$, we have

$$\mathbb{P}\left[\left|M_n - \mathbb{E}[M_n]\right| \geq \delta \mathbb{E}[M_n]\right] \quad \to \quad 0 \qquad\qquad \text{as } n \to +\infty. \tag{27}$$

3.3.2 Main argument

With these preliminary results in hand, we now turn to analysis of the collections of random variables $\{U_i, i \in S\}$ and $\{V_j, j \in S^c\}$.

Analysis of $\mathcal{M}(V)$: We begin by analyzing the behavior of $\max_{j \in S^c} |V_j|$. First, for a fixed but arbitrary $\delta > 0$, define the event $\mathcal{T}(\delta) := \{|M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n]\}$. By conditioning on $\mathcal{T}(\delta)$ and its complement $[\mathcal{T}(\delta)]^c$, we have the upper bound

$$\mathbb{P}[\max_{j \in S^c} |V_j| > \lambda_n] \quad \leq \quad \mathbb{P}\left[\max_{j \in S^c} |V_j| > \lambda_n \;\mid\; [\mathcal{T}(\delta)]^c\right] + \mathbb{P}[\mathcal{T}(\delta)].$$

By the concentration statement in Lemma 3, we have $\mathbb{P}[\mathcal{T}(\delta)] \to 0$, so that it suffices to analyze the first term. Set $\mu_j = \mathbb{E}[V_j | X_S]$, and let Z be a zero-mean Gaussian vector with $\text{cov}(Z) = \text{cov}(V \mid X_S, W)$.

$$\begin{aligned}
\max_{j \in S^c} |V_j| &= \max_{j \in S^c} |\mu_j + Z_j| \\
&\leq \max_{j \in S^c} [|\mu_j| + |Z_j|] \\
&\leq (1 - \epsilon)\lambda_n + \max_{j \in S^c} |Z_j|,
\end{aligned}$$

where we have used the upper bound (23) on the mean. This inequality establishes the inclusion of events

$$\{\max_{j \in S^c} |Z_j| \leq \epsilon \lambda_n\} \quad \subseteq \quad \{\max_{j \in S^c} |V_j| \leq \lambda_n\},$$

thereby showing that it suffices to prove that $\mathbb{P}[\max_{j \in S^c} |Z_j| > \epsilon\lambda_n \;\mid\; [\mathcal{T}(\delta)]^c] \to 0$.

Note that conditioned on $[\mathcal{T}(\delta)]^c$, the maximum value of M_n is $v^* := (1 + \delta)\mathbb{E}[M_n]$. Since Gaussian maxima increase with increasing variance, we have

$$\mathbb{P}\left[\max_{j \in S^c} |Z_j| > \epsilon\lambda_n \;\mid\; [\mathcal{T}(\delta)]^c\right] \quad \leq \quad \mathbb{P}\left[\max_{j \in S^c} |\widetilde{Z}_j| > \epsilon\lambda_n\right],$$

where $\widetilde{Z}$ is zero-mean Gaussian with covariance $v^* \, \Sigma_{(S^c|S)}$.

Using Lemma 11, it suffices to show that $\mathbb{P}[\max_{j \in S^c} \widetilde{Z}_j > \epsilon\lambda_n]$ converges to zero. Accordingly, we complete this part of the proof via the following two lemmas, both of which are proved in Appendix D:

Lemma 4. *Under the stated assumptions of the theorem, we have* $\frac{v^*}{\lambda_n^2} \to 0$ *and*

$$\lim_{n \to +\infty} \frac{1}{\lambda_n} \mathbb{E}[\max_{j \in S^c} \widetilde{Z}_j] \quad \leq \quad \epsilon.$$

Lemma 5. *For any* $\eta > 0$*, we have*

$$\mathbb{P}\left[\max_{j \in S^c} \widetilde{Z}_j > \eta + \mathbb{E}[\max_{j \in S^c} \widetilde{Z}_j]\right] \quad \leq \quad \exp\left(-\frac{\eta^2}{2v^*}\right). \tag{28}$$

Lemma 4 implies that for all $\delta > 0$, we have $\mathbb{E}[\max_{j \in S^c} \widetilde{Z}_j] \leq (1 + \frac{\delta}{2})\epsilon\lambda_n$ for all n sufficiently large. Therefore, setting $\eta = \frac{\delta}{2}\lambda_n\epsilon$ in the bound (28), we have for fixed $\delta > 0$ and n sufficiently large:

$$\begin{aligned}
\mathbb{P}\left[\max_{j \in S^c} \widetilde{Z}_j > (1 + \delta)\lambda_n\epsilon\right] &\leq \mathbb{P}\left[\max_{j \in S^c} \widetilde{Z}_j > \frac{\delta}{2}\lambda_n\epsilon + \mathbb{E}[\max_{j \in S^c} \widetilde{Z}_j]\right] \\
&\leq 2\exp\left(-\frac{\delta^2\lambda_n^2\epsilon^2}{8v^*}\right).
\end{aligned}$$

From Lemma 4, we have $\lambda_n^2/v^* \to +\infty$, which implies that $\mathbb{P}[\max_{j\in S^c} \widetilde{Z}_j > (1+\delta)\lambda_n \epsilon] \to 0$ for all $\delta > 0$. By the arbitrariness of $\delta > 0$, we thus have $\mathbb{P}[\max_{j\in S^c} \widetilde{Z}_j \le \epsilon\lambda_n] \to 1$, thereby establishing that property (9a) of Lemma 1 holds w.p. one asymptotically.

Analysis of $\{U_i\}$: Next we prove that $\max_{i\in S} |U_i| < \rho_n := \min_{i\in S} |\beta_i^*|$ with probability one as $n \to +\infty$. Conditioned on X_S, the only random component in U_i is the noise vector W. A straightforward calculation yields that this conditioned RV is Gaussian, with mean and variance

$$\begin{aligned} Y_i \;:=\; \mathbb{E}[U_i \mid X_S] &\;=\; -\lambda_n e_i^T \left(\frac{1}{n} X_S^T X_S\right)^{-1} \vec{b}, \\ Y_i' \;:=\; \mathrm{var}[U_i \mid X_S] &\;=\; \frac{\sigma^2}{n} e_i^T \left[\frac{1}{n} X_S^T X_S\right]^{-1} e_i, \end{aligned}$$

respectively. The following lemma, proved in Appendix D.5, is key to our proof:

Lemma 6. *(a) The random variables Y_i and Y_i' have means*

$$\mathbb{E}[Y_i] = \frac{-\lambda_n\, n}{n-s-1} e_i^T\, (\Sigma_{SS})^{-1}\, \vec{b}, \qquad \textit{and} \qquad \mathbb{E}[Y_i'] \;=\; \frac{\sigma^2}{n-s-1}\, e_i^T (\Sigma_{SS})^{-1} e_i, \tag{29}$$

respectively, which are bounded as

$$|\mathbb{E}[Y_i]| \;\le\; \frac{2D_{\max} n\lambda_n}{n-s-1}, \qquad \textit{and} \qquad \frac{\sigma^2}{C_{max}\,(n-s-1)} \;\le\; \mathbb{E}[Y_i'] \;\le\; \frac{\sigma^2 D_{\max}}{n-s-1}. \tag{30}$$

(b) Moreover, each pair (Y_i, Y_i') is sharply concentrated, in that we have

$$\mathbb{P}\left[|Y_i| \ge \frac{6D_{\max} n\lambda_n}{n-s-1}, \quad \textit{or} \quad |Y_i'| \ge 2\mathbb{E}[Y_i']\right] \;\le\; \frac{K}{n-s}, \tag{31}$$

where K is a fixed constant independent of n and s.

We exploit this lemma as follows. First define the event

$$\mathcal{T}(\delta) \;:=\; \bigcup_{i=1}^{s} \left\{ |Y_i| \ge \frac{6D_{\max} n\lambda_n}{n-s-1}, \quad \text{or} \quad |Y_i'| \ge 2\mathbb{E}[Y_i'] \right\}.$$

By the union bound and Lemma 6(b), we have

$$\mathbb{P}[\mathcal{T}(\delta)] \;\le\; s\frac{K}{n-s} \;=\; \frac{K}{\frac{n}{s}-1} \;\to\; 0,$$

since $\frac{n}{s} \to +\infty$ as $n \to +\infty$. For convenience in notation, for any $a \in \mathbb{R}$ and $b \in \mathbb{R}_+$, we use $U_i(a,b)$ to denote a Gaussian random variable with mean a and variance b. Conditioning on the event $\mathcal{T}(\delta)$ and its complement, we have

$$\begin{aligned} \mathbb{P}[\max_{i\in S} U_i > \rho_n] &\;\le\; \mathbb{P}[\max_{i\in S} U_i > \rho_n \mid \mathcal{T}(\delta)^c] + \mathbb{P}[\mathcal{T}(\delta)] \\ &\;\le\; \mathbb{P}[\max_{i\in S} U_i(\mu_i^*, v_i^*) > \rho_n] + \frac{K}{\frac{n}{s}-1}, \end{aligned} \tag{32}$$

where each $U_i(\mu_i^*, v_i^*)$ is Gaussian with mean $\mu_i^* := 6D_{\max}\lambda_n \frac{n}{n-s-1}$ and variance $v_i^* := 2\mathbb{E}[Y_i']$ respectively. In asserting the inequality (32), we have used the fact that the probability of the event $\{\max_{i\in S} Y_i > \rho_n\}$ increases as the mean and variance of Y_i increase. Continuing the argument, we have

$$\begin{aligned}
\mathbb{P}[\max_{i\in S} U_i(\mu_i^*, v_i^*) > \rho_n] &\leq \mathbb{P}[\max_{i\in S} |U_i(\mu_i^*, v_i^*)| > \rho_n] \\
&\leq \frac{1}{\rho_n}\mathbb{E}\left[\max_{i\in S} |U_i(\mu_i^*, v_i^*)|\right],
\end{aligned}$$

where the last step uses Markov's inequality. We now decompose $U_i(\mu_i^*, v_i^*) \stackrel{d}{=} 2D_{\max}\,\lambda_n \frac{n}{n-s-1} + \widetilde{U}_i(0, v_i^*)$, and write

$$\mathbb{E}\left[\max_{i\in S} |U_i(\mu_i^*, v_i^*)|\right] \leq 2D_{\max}\,\lambda_n \frac{n}{n-s-1} + \mathbb{E}\left[\max_{i\in S} |\widetilde{U}(0, v_i^*)|\right].$$

With this decomposition, we use the bound (30) on $v_i^* := 2\mathbb{E}[Y_i']$ and Lemma 9 on Gaussian maxima (see Appendix B) to conclude that

$$\frac{1}{\rho_n}\mathbb{E}\left[\max_{i\in S} |U_i(\mu_i^*, v_i^*)|\right] \leq \frac{1}{\rho_n}\left[2D_{\max}\,\lambda_n \frac{n}{n-s-1} + 3\sqrt{\frac{2\sigma^2\, D_{\max}\log s}{n-s-1}}\right],$$

which converges to zero by the second condition (21) in the theorem statement.

3.4 Proof of Theorem 1(a)

We establish the claim by proving that under the stated conditions, $\max_{j\in S^c} |V_j| > \lambda_n$ with probability one, for any positive sequence $\lambda_n > 0$. We begin by writing $V_j = \mathbb{E}[V_j] + \widetilde{V}_j$, where $\widetilde{V}_j$ is zero-mean. Now

$$\begin{aligned}
\max_{j\in S^c} |V_j| &\geq \max_{j\in S^c} |\widetilde{V}_j| - \max_{j\in S^c} |\mathbb{E}[V_j]| \\
&\geq \max_{j\in S^c} |V_j| - (1-\epsilon)\lambda_n
\end{aligned}$$

where have used Lemma 2. Consequently, the event $\{\max_{j\in S^c} |\widetilde{V}_j| > (2-\epsilon)\lambda_n\}$ implies the event $\{\max_{j\in S^c} |V_j| > \lambda_n\}$, so that

$$\mathbb{P}[\max_{j\in S^c} |V_j| > \lambda_n] \geq \mathbb{P}[\max_{j\in S^c} |\widetilde{V}_j| > (2-\epsilon)\,\lambda_n].$$

From the preceding proof of Theorem 1(b), we know that conditioned on X_S and W, the random vector $(V_1, \ldots, V_N)$ is Gaussian with covariance of the form $M_n\,[\Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}]$; thus, the zero-mean version $(\widetilde{V}_1, \ldots, \widetilde{V}_N)$ has the same covariance. Moreover, Lemma 3 guarantees that the random scaling term M_n is sharply concentrated. In particular, defining for any $\delta > 0$ the event $\mathcal{T}(\delta) := \{\,|M_n - \mathbb{E}[M_n]| \geq \delta\mathbb{E}[M_n]\}$, we have $\mathbb{P}[\mathcal{T}(\delta)] \to 0$, and the bound

$$\begin{aligned}
\mathbb{P}[\max_{j\in S^c} |\widetilde{V}_j| > (2-\epsilon)\,\lambda_n] &\geq (1-\mathbb{P}[\mathcal{T}(\delta)])\,\mathbb{P}\left[\max_{j\in S^c} |\widetilde{V}_j| > (2-\epsilon)\,\lambda_n \;\middle|\; \mathcal{T}(\delta)^c\right] \\
&\geq (1-\mathbb{P}[\mathcal{T}(\delta)])\,\mathbb{P}\left[\max_{j\in S^c} |Z_j(v^*)| > (2-\epsilon)\,\lambda_n\right],
\end{aligned}$$

where each $Z_j \equiv Z_j(v^*)$ is the conditioned version of $\widetilde{V}_j$ with the scaling factor M_n fixed to $v^* := (1-\delta)\mathbb{E}[M_n]$. (Here we have used the fact that the probability of Gaussian maxima decreases as the variance decreases, and that $\mathrm{var}(\widetilde{V}_j) \geq v^*$ when conditioned on $\mathcal{T}(\delta)^c$.)

Our proof proceeds by first analyzing the expected value, and then exploiting Gaussian concentration of measure. We summarize the key results in the following:

Lemma 7. *Under the stated conditions, one of the following two conditions must hold:*

(a) either $\frac{\lambda_n^2}{v^*} \to +\infty$, and there exists some $\gamma > 0$ such that $\frac{1}{\lambda_n}\mathbb{E}[\max_{j \in S^c} Z_j] \geq (2-\epsilon)\,[1+\gamma]$ for all sufficiently large n, or

(b) there exist constants $\alpha, \gamma > 0$ such that $\frac{v^*}{\lambda_n^2} \leq \alpha$ and $\frac{1}{\lambda_n}\mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma\sqrt{\log N}$ for all sufficiently large n.

Lemma 8. *For any $\eta > 0$, we have*

$$\mathbb{P}[\max_{j \in S^c} Z_j(v^*) < \mathbb{E}[\max_{j \in S^c} Z_j(v^*)] - \eta] \quad \leq \quad \exp\left(-\frac{\eta^2}{2v^*}\right). \tag{33}$$

Using these two lemmas, we complete the proof as follows. First, if condition (a) of Lemma 7 holds, then we set $\eta = \frac{(2-\epsilon)\,\gamma\lambda_n}{2}$ in equation (33) to obtain that

$$\mathbb{P}[\frac{1}{\lambda_n}\max_{j \in S^c} Z_j(v^*) \geq (2-\epsilon)\,(1+\frac{\gamma}{2})\,] \quad \geq \quad 1 - \exp\left(-\frac{(2-\epsilon)^2\,\gamma^2\lambda_n^2}{8v^*}\right).$$

This probability converges to 1 since $\frac{\lambda_n^2}{v^*} \to +\infty$ from Lemma 7(a).

On the other hand, if condition (b) holds, then we use the bound $\frac{1}{\lambda_n}\mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma\sqrt{\log N}$ and set $\eta = \frac{\gamma\lambda_n\sqrt{\log N}}{2}$ in equation (33) to obtain

$$\begin{aligned}\mathbb{P}[\frac{1}{\lambda_n}\max_{j \in S^c} Z_j(v^*) > 2\,(2-\epsilon)\,] \quad &\geq \quad \mathbb{P}[\frac{1}{\lambda_n}\max_{j \in S^c} Z_j(v^*) \geq \frac{\gamma\sqrt{\log N}}{2}\,] \\ &\geq \quad 1 - \exp\left(-\frac{\gamma^2\lambda_n^2\log N}{8v^*}\right).\end{aligned}$$

This probability also converges to 1 since $\frac{\lambda_n^2}{v^*} \geq 1/\alpha$ and $\log N \to +\infty$. Thus, in either case, we have shown that $\lim_{n \to +\infty}\mathbb{P}[\frac{1}{\lambda_n}\max_{j \in S^c} Z_j(v^*) > (2-\epsilon)] = 1$, thereby completing the proof of Theorem 1(a).

4 Illustrative simulations

In this section, we provide some simulations to confirm the threshold behavior predicted by Theorem 1. We consider the following three types of sparsity indices:

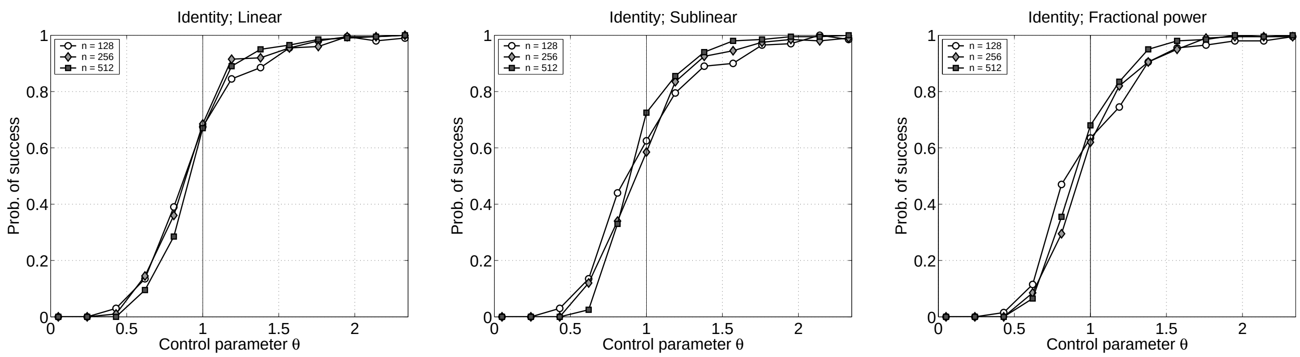
(a) *linear sparsity*, meaning that $s(p) = \alpha p$ for some $\alpha \in (0,1)$;

(b) *sublinear sparsity*, meaning that $s(p) = \alpha p/(\log(\alpha p))$ for some $\alpha \in (0,1)$, and

(c) *fractional power* sparsity, meaning that $s(p) = \alpha p^\gamma$ for some $\alpha, \gamma \in (0,1)$.

For all three types of sparsity indices, we investigate the success/failure of the Lasso in recovering the sparsity pattern, where the number of observations scales as $n = 2\,\theta\, s \log(p-s) + s + 1$. The *control parameter* θ is varied in the interval $(0, 2.4)$. For all results shown here, we fixed $\alpha = 0.40$ for all three ensembles, and set $\gamma = 0.75$ for the fractional power ensemble. In addition, we set $\lambda_n = \sqrt{\frac{\log(p-s)\log(s)}{n}}$ in all cases.

We begin by considering the uniform Gaussian ensemble, in which each row x_k is chosen in an i.i.d. manner from the multivariate $N(0, I_{p\times p})$ distribution. Recall that for the uniform Gaussian ensemble, the critical value is $\theta_u = \theta_\ell = 1$. Figure 1 plots the control parameter θ versus the probability of success, for linear sparsity (a), sublinear sparsity pattern (b), and fractional power sparsity (c), for three different problem sizes ($p \in \{128, 256, 512\}$). Each point represents the average of 200 trials. Note how the probability of success rises rapidly from 0 around the predicted



(a) (b) (c)

Figure 1. Plots of the number of data samples (indexed by the control parameter θ versus the probability of success in the Lasso for the uniform Gaussian ensemble. Each panel shows three curves, corresponding to the problem sizes $p \in \{128, 256, 512\}$, and each point on each curve represents the average of 200 trials. (a) Linear sparsity index: $s(p) = \alpha p$. (b) Sublinear sparsity index $s(p) = \alpha p / \log(\alpha p)$. (c) Fractional power sparsity index $s(p) = \alpha p^\gamma$ with $\gamma = 0.75$.

threshold point $\theta = 1$, with the sharpness of the threshold increasing for larger problem sizes.

We now consider a non-uniform Gaussian ensemble—in particular, one in which the covariance matrices Σ are Toeplitz with the structure

$$\Sigma \;=\; \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} & \rho^p \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^p & \cdots & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}, \tag{34}$$

for some $\rho \in (-1, +1)$. As shown by Zhao and Yu [35], this family of Toeplitz matrices satisfy condition (18). Moreover, the maximum and minimum eigenvalues (C_{min} and C_{max}) can be computed using standard asymptotic results on Toeplitz matrix families [20]. Figure 2 shows representative results for this Toeplitz family with $\rho = 0.10$. Panel (a) corresponds to linear sparsity $s = \alpha p$ with $\alpha = 0.40$), and panel (b) corresponds to sublinear sparsity ($s = \alpha p / \log(\alpha p)$ with $\alpha = 0.40$). Each panel shows three curves, corresponding to the problem sizes $p \in \{128, 256, 512\}$, and each

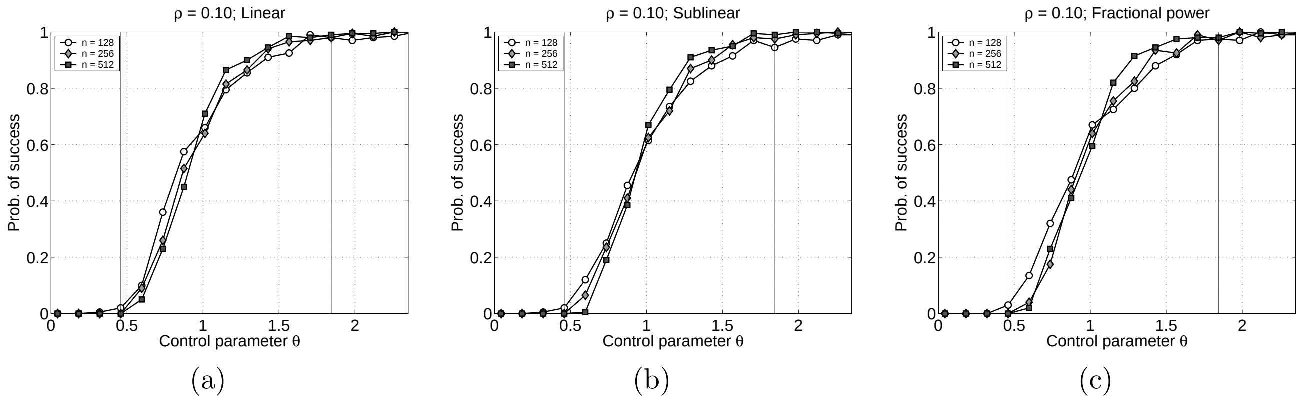


Figure 2. Plots of the number of data samples (indexed by the control parameter θ versus the probability of success in the Lasso for the Toeplitz family (34) with $\rho = 0.10$. Each panel shows three curves, corresponding to the problem sizes $p \in \{128, 256, 512\}$, and each point on each curve represents the average of 200 trials. (a) Linear sparsity index: $s(p) = \alpha p$. (b) Sublinear sparsity index $s(p) = \alpha p / \log(\alpha p)$. (c) Fractional power sparsity index $s(p) = \alpha p^{\gamma}$ with $\gamma = 0.75$.

point on each curve represents the average of 200 trials. The vertical lines to the left and right of $\theta = 1$ represent the theoretical upper and lower bounds on the threshold ($\theta_u \approx 1.84$ and $\theta_\ell \approx 0.46$ respectively in this case). Once again, these simulations show good agreement with the theoretical predictions.

5 Discussion

The problem of recovering the sparsity pattern of a high-dimensional vector β^* from noisy observations has important applications in signal denoising, graphical model selection, sparse approximation, and subset selection. This paper focuses on the behavior of ℓ_1-regularized quadratic programming, also known as the Lasso, for estimating such sparsity patterns in the noisy and high-dimensional setting. The main contribution of this paper is to establish a set of general and sharp conditions on the observations n, the sparsity index s (i.e., number of non-zero entries in β^*), and the ambient dimension p that characterize the success/failure behavior of the Lasso in the high-dimensional setting, in which n, p and s all tend to infinity. For the uniform Gaussian ensemble, our threshold result is sharp, whereas for more general Gaussian ensembles, it should be possible to tighten the analysis given here.

Acknowledgements

We would like to thank Noureddine El Karoui and Bin Yu for helpful comments and pointers. This work was partially supported by an Alfred P. Sloan Foundation Fellowship, and an Intel Corporation Equipment Grant.

A Proof of Lemma 1

By standard conditions for optimality in a convex program [21], the point $\widehat{\beta} \in \mathbb{R}^p$ is optimal if and only if there exists a subgradient $\widehat{z} \in \partial \ell_1(\widehat{\beta})$ such that

$$\frac{1}{n} X^T X \widehat{\beta} - \frac{1}{n} X^T y + \lambda \widehat{z} \;=\; 0. \tag{35}$$

Here the subdifferential of the ℓ_1 norm takes the form

$$\partial \ell_1(\widehat{\beta}) = \left\{ \widehat{z} \in \mathbb{R}^p \;\mid\; \widehat{z}_i = \operatorname{sgn}(\widehat{\beta}_i) \quad \text{for } \widehat{\beta}_i \neq 0, \qquad |\widehat{z}_j| \leq 1 \quad \text{otherwise} \right\}.$$

Substituting our observation model $y = X\beta^* + w$ and re-arranging yields

$$\frac{1}{n} X^T X (\widehat{\beta} - \beta^*) - \frac{1}{n} X^T w + \lambda \widehat{z} \;=\; 0. \tag{36}$$

Now condition $\mathcal{R}(X, \beta^*, w, \lambda)$ holds if and only we have

$$\widehat{\beta}_{S^c} = 0, \quad \widehat{\beta}_S \neq 0, \qquad \text{and} \qquad \widehat{z}_S = \operatorname{sgn}(\beta^*_S), \qquad |\widehat{z}_{S^c}| \leq 1.$$

From these conditions and using equation (36), we conclude that the condition $\mathcal{R}(X, \beta^*, w, \lambda)$ holds if and only if

$$\begin{aligned} \frac{1}{n} X_{S^c}^T X_S \left(\widehat{\beta}_S - \beta^*_S \right) - \frac{1}{n} X_{S^c}^T w \;&=\; -\lambda \widehat{z}_{S^c}. \\ \frac{1}{n} X_S^T X_S \left(\widehat{\beta}_S - \beta^*_S \right) - \frac{1}{n} X_S^T w \;&=\; -\lambda \operatorname{sgn}(\beta^*_S). \end{aligned}$$

Using the invertibility of $X_S^T X_S$, we may solve for $\widehat{\beta}_S$ and $\widehat{z}_{S^c}$ to conclude that

$$\begin{aligned} \lambda\, \widehat{z}_{S^c} \;&=\; X_{S^c}^T X_S \left(X_S^T X_S \right)^{-1} \left[\frac{1}{n} X_S^T w - \lambda \operatorname{sgn}(\beta^*_S) \right] - \frac{1}{n} X_{S^c}^T w \\ \widehat{\beta}_S \;&=\; \beta^*_S + \left(\frac{1}{n} X_S^T X_S \right)^{-1} \left[\frac{1}{n} X_S^T w - \lambda \operatorname{sgn}(\beta^*_S) \right]. \end{aligned}$$

From these relations, the conditions $|\widehat{z}_{S^c}| \leq 1$ and $\widehat{\beta}_S \neq 0$ yield conditions (9a) and (9b) respectively.

B Some Gaussian comparison results

We state here (without proof) some well-known comparison results on Gaussian maxima [25]. We begin with a crude but useful bound:

Lemma 9. *For any Gaussian random vector $(X_1, \ldots, X_n)$, we have*

$$\mathbb{E} \max_{1 \leq i \leq n} |X_i| \;\leq\; 3\sqrt{\log n} \max_{1 \leq i \leq n} \sqrt{\mathbb{E} X_i^2}.$$

Next we state (a version of) the Sudakov-Fernique inequality [25, 5]:

Lemma 10. *Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be Gaussian random vectors such that for all i, j*

$$\mathbb{E}[(Y_i - Y_j)^2] \;\leq\; \mathbb{E}[(X_i - X_j)^2].$$

Then $\mathbb{E}[\max_{1 \leq i \leq n} Y_i] \leq \mathbb{E}[\max_{1 \leq i \leq n} X_i]$.

C Auxiliary lemma

For future use, we state formally the following elementary

Lemma 11. *Given a collection $\{Z_1, Z_2, \ldots, Z_N\}$ of zero-mean random variables, for any constant $a > 0$ we have*

$$\begin{aligned} \mathbb{P}[\max_{1 \le j \le N} |Z_j| \le a] &\le \mathbb{P}[\max_{1 \le j \le N} Z_j \le a], \qquad \textit{and} &\text{(39a)} \\ \mathbb{P}[\max_{1 \le j \le N} |Z_j| > a] &\le 2\mathbb{P}[\max_{1 \le j \le N} Z_j > a]. &\text{(39b)} \end{aligned}$$

Proof. The first inequality is trivial. To establish the inequality (39b), we write

$$\begin{aligned} \mathbb{P}[\max_{1 \le j \le N} |Z_j| > a] &= \mathbb{P}[(\max_{1 \le j \le N} Z_j > a) \text{ or } (\min_{1 \le j \le N} Z_j < -a)] \\ &\le \mathbb{P}[\max_{1 \le j \le N} Z_j > a] + \mathbb{P}[\min_{1 \le j \le N} Z_j < -a] \\ &= 2\mathbb{P}[\max_{1 \le j \le N} Z_j > a], \end{aligned}$$

where we have used the union bound, and the symmetry of the events $\{\max_{1 \le j \le N} Z_j > a\}$ and $\{\min_{1 \le j \le N} Z_j < -a\}$. □

D Lemma for Theorem 1

D.1 Proof of Lemma 2

Conditioned on both X_S and W, the only random component in V_j is the column vector X_j. Using standard LLSE formula [e.g., 2] (i.e., for estimating X_{S^c} on the basis of X_S), the random variable $(X_{S^c} \mid X_S, W) \sim (X_{S^c} \mid X_S)$ is Gaussian with mean and covariance

$$\begin{aligned} \mathbb{E}[X_{S^c}^T \mid X_S,\, W] &= \Sigma_{S^c S}(\Sigma_{SS})^{-1} X_S^T, &\text{(40a)} \\ \operatorname{var}(X_{S^c} \mid X_S) &= \Sigma_{(S^c|S)} \;=\; \Sigma_{S^c S^c} - \Sigma_{S^c S}(\Sigma_{SS})^{-1}\Sigma_{S S^c}. &\text{(40b)} \end{aligned}$$

Consequently, we have

$$\begin{aligned} |\mathbb{E}[V_j \mid X_S, W]| &= \left| \Sigma_{S^c S}(\Sigma_{SS})^{-1} X_S^T \left\{ X_S \left(X_S^T X_S\right)^{-1} \lambda_n \vec{b} - \left[X_S \left(X_S^T X_S\right)^{-1} X_S^T - I_{n \times n} \right] \frac{W}{n} \right\} \right| \\ &= \left| \Sigma_{S^c S}(\Sigma_{SS})^{-1} \lambda_n \vec{b} \right| \\ &\le \lambda_n (1 - \epsilon) \mathbf{1}, \end{aligned}$$

as claimed.

Similarly, we compute the elements of the conditional covariance matrix as follows

$$\begin{aligned} &\operatorname{cov}(V_j, V_k \,|\, X_S, W) = \\ &\qquad \operatorname{cov}(A_{ji}, A_{ki} \mid X_S, W) \; \left\{ \lambda_n^2 \vec{b}^T (X_S^T X_S)^{-1} \vec{b} + \frac{1}{n^2} W^T \left[I_{n \times n} - X_S \left(X_S^T X_S\right)^{-1} X_S^T \right] W \right\}. \end{aligned}$$

D.2 Proof of Lemma 3

We begin by computing the expected value. Since $X_S^T X_S$ is Wishart with matrix Σ_{SS}, the random matrix $(X_S^T X_S)^{-1}$ is inverse Wishart with mean $\mathbb{E}[(X_S^T X_S)^{-1}] = \frac{(\Sigma_{SS})^{-1}}{n-s-1}$ (see Lemma 7.7.1 of Anderson [1]). Hence we have

$$\mathbb{E}\left[\lambda_n^2 \vec{b}^T \left(X_S^T X_S\right)^{-1} \vec{b}\right] = \frac{\lambda_n^2}{n-s-1}\, \vec{b}^T (\Sigma_{SS})^{-1} \vec{b}. \tag{41}$$

Now define the random matrix $R = I_{n\times n} - X_S(X_S^T X_S)^{-1} X_S^T$. A straightforward calculation yields that $R^2 = R$, so that all the eigenvalues of R are either 0 or 1. In particular, for any vector $z = X_S u$ in the range of X_S, we have

$$Rz = \left[I_{n\times n} - X_S(X_S^T X_S)^{-1} X_S^T\right] X_S u \;=\; 0. \tag{42}$$

Hence $\dim(\ker R) = \dim(\operatorname{range} X_S) = s$. Since R is symmetric and positive semidefinite, there exists an orthogonal matrix U such that $R = U^T D U$, where D is diagonal with $(n-s)$ ones, and s zeros. The random matrices D and U are both independent of W, since X_S is independent of W. Hence we have

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[W^T R W \;\mid\; X_S\right] &= \frac{1}{n^2}\mathbb{E}\left[W^T U^T D U W \;\mid X_S\right] \\
&= \frac{1}{n^2}\operatorname{trace} D U U^T \mathbb{E}\left[W W^T \mid\; X_S\right] \\
&= \sigma^2\, \frac{n-s}{n^2}
\end{aligned} \tag{43}$$

since $\mathbb{E}[WW^T] = \sigma^2 I$. Consequently, we have established that $\mathbb{E}[M_n] = \frac{\lambda_n^2}{n-s-1}\; \vec{b}^T (\Sigma_{SS})^{-1} \vec{b} + \frac{\sigma^2\,(n-s)}{n^2}$ as claimed.

We now compute the expected value of the squared variance

$$M_n^2 \;=\; \underbrace{\lambda_n^4 \left[\vec{b}^T \left(X_S^T X_S\right)^{-1} \vec{b}\right]^2}_{T_1} + \underbrace{2\frac{\lambda_n^2}{n^2}\left[\vec{b}^T \left(X_S^T X_S\right)^{-1} \vec{b}\right]\left(W^T R W\right)}_{T_2} + \underbrace{\frac{1}{n^4}\left(W^T R W\right)^2}_{T_3}$$

First, conditioning on X_S and using the eigenvalue decomposition D of R, we have

$$\begin{aligned}
\mathbb{E}[T_3 | X_S] &= \frac{1}{n^4}\mathbb{E}[(W^T D W)^2] \\
&= \frac{1}{n^4}\mathbb{E}\left[(\sum_{i=1}^{n-s} W_i)^2\right] \\
&= \frac{2(n-s)\sigma^4}{n^4} + \frac{(n-s)^2\sigma^4}{n^4}.
\end{aligned} \tag{44}$$

whence $\mathbb{E}[T_3] = \frac{2(n-s)\sigma^4}{n^4} + \frac{(n-s)^2\sigma^4}{n^4}$ as well.

Similarly, using conditional expectation and our previous calculation (43) of $\mathbb{E}[W^T RW \mid X_S]$, we have

$$
\begin{aligned}
\mathbb{E}[T_2] &= \frac{2\lambda_n^2}{n^2}\mathbb{E}\left[\mathbb{E}\left[\vec{b}^T(X_S^T X_S)^{-1}\vec{b}\ (W^T RW) \mid X_S\right]\right] \\
&= \frac{2\lambda_n^2\,(n-s)\sigma^2}{n^2}\mathbb{E}\left[\vec{b}^T(X_S^T X_S)^{-1}\vec{b}\right] \\
&= \frac{2\lambda_n^2\,(n-s)\,\sigma^2}{n^2\,(n-s-1)}\vec{b}^T\,(\Sigma_{SS})^{-1}\,\vec{b}, \qquad (45)
\end{aligned}
$$

where the final step uses Lemma 7.7.1 of Anderson [1] on the expectation of inverse Wishart matrices.

Lastly, since $(X_S^T X_S)^{-1}$ is inverse Wishart with matrix $(\Sigma_{SS})^{-1}$, we can use formula for second moments of inverse Wishart matrices (see, e.g., Siskind [31]) to write, for all $n > s+3$,

$$
\mathbb{E}[T_1] = \frac{\lambda_n^4}{(n-s)\,(n-s-3)}\left[\vec{b}^T(\Sigma_{SS})^{-1}\vec{b}\right]^2\left\{1+\frac{1}{n-s-1}\right\}.
$$

Consequently, combining our results, we have

$$
\begin{aligned}
\mathrm{var}(M_n) &= \mathbb{E}[M_n^2] - (\mathbb{E}[M_n])^2 \\
&= \sum_{i=1}^{3}\mathbb{E}[T_i] - \left\{\frac{\sigma^4(n-s)^2}{n^4} + 2\frac{\sigma^2(n-s)}{n^2}\frac{\lambda_n^2}{n-s-1}\,\vec{b}^T(\Sigma_{SS})^{-1}\vec{b} + \left(\frac{\lambda_n^2}{n-s-1}\,\vec{b}^T(\Sigma_{SS})^{-1}\vec{b}\right)^2\right\} \\
&= \underbrace{\frac{2(n-s)\sigma^4}{n^4}}_{H_1} + \underbrace{\frac{\lambda_n^4\,[\vec{b}^T(\Sigma_{SS})^{-1}\vec{b}]^2}{(n-s-1)\,(n-s-3)}\left\{\frac{1}{(n-s)}+\frac{n-s-1}{(n-s)}-\frac{(n-s-3)}{(n-s-1)}\right\}}_{H_2}. \qquad (46)
\end{aligned}
$$

Finally, we establish the concentration result. Using Chebyshev's inequality, we have

$$
\mathbb{P}\left[|M_n - \mathbb{E}[M_n]| \geq \delta\mathbb{E}[M_n]\right] \leq \frac{\mathrm{var}(M_n)}{\delta^2(\mathbb{E}[M_n])^2},
$$

so that it suffices to prove that $\mathrm{var}(M_n)/(\mathbb{E}[M_n])^2 \to 0$ as $n \to +\infty$. We deal with each of the two variance terms H_1 and H_2 in equation (46) separately. First, we have

$$
\frac{H_1}{(\mathbb{E}[M_n])^2} \leq \frac{2(n-s)\sigma^4}{n^4}\frac{n^4}{(n-s)^2\sigma^4} = \frac{2}{n-s} \to 0.
$$

Secondly, denoting $A = (\vec{b}^T(X_S^T X_S)^{-1}\vec{b})$ for short-hand, we have

$$
\begin{aligned}
\frac{H_2}{(\mathbb{E}[M_n])^2} &\leq \frac{(n-s-1)^2}{\lambda_n^4 A^2}\frac{\lambda_n^4 A^2}{(n-s-1)\,(n-s-3)}\left\{\frac{1}{(n-s)}+\frac{n-s-1}{(n-s)}-\frac{(n-s-3)}{(n-s-1)}\right\} \\
&= \frac{(n-s-1)}{(n-s-3)}\left\{\frac{1}{(n-s)}+\frac{n-s-1}{(n-s)}-\frac{(n-s-3)}{(n-s-1)}\right\},
\end{aligned}
$$

which also converges to 0 as $(n-s) \to 0$.

D.3 Proof of Lemma 4

Recall that the Gaussian random vector $(Z_1, \ldots, Z_N)$ is zero-mean with covariance $v^* \Sigma_{(S^c|S)}$, where $\Sigma_{(S^c|S)} := \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}$. For any index i, let $e_i \in \mathbb{R}^N$ be equal to 1 in position i, and zero otherwise. For any two indices $i \neq j$, we have

$$\begin{aligned}
\mathbb{E}[(Z_i - Z_j)^2] &= v^*(e_i - e_j)^T \Sigma_{(S^c|S)}(e_i - e_j) \\
&\leq 2v^* \lambda_{max}(\Sigma_{(S^c|S)}) \\
&\leq 2C_{max} v^*,
\end{aligned}$$

since $\Sigma_{(S^c|S)} \preceq \Sigma_{S^cS^c}$ by definition, and $\Lambda_{max}(\Sigma_{S^cS^c}) \leq \Lambda_{max}(\Sigma) \leq C_{max}$.

Letting $(X_1, \ldots, X_N) \sim N(0, C_{max} v^* I_{N \times N})$, we have $\mathbb{E}[(X_i - X_j)^2] = 2C_{max} v^*$. Hence, applying the Sudakov-Fernique inequality [25] yields $\mathbb{E}[\max_j Z_j] \leq \mathbb{E}[\max_j X_j]$. By asymptotic behavior of i.i.d. Gaussians [19, 7], we have $\lim_{N \to \infty} \frac{\mathbb{E}[\max_j X_j]}{\sqrt{2C_{max} v^* \log N}} = 1$. Consequently, for all $\delta' > 0$, there exists an $N(\delta')$ such that for all $N \geq N(\delta')$, we have

$$\begin{aligned}
\frac{1}{\lambda_n} \mathbb{E}[\max_j Z_j(v^*)] &\leq \frac{1}{\lambda_n} \mathbb{E}[\max_j X_j] \\
&\leq (1 + \delta') \sqrt{\frac{2C_{max} v^* \log N}{\lambda_n^2}} \\
&= (1 + \delta') \sqrt{1 + \delta} \sqrt{\frac{2C_{max} \log N}{n - s - 1} \vec{b}^T (\Sigma_{SS})^{-1} \vec{b} + \frac{2C_{max} \sigma^2 (1 - \frac{s}{n}) \log N}{n \lambda_n^2}} \\
&\leq (1 + \delta') \sqrt{1 + \delta} \sqrt{\frac{2C_{max}\, s \log N}{n - s - 1} \frac{1}{C_{min}} + \frac{2C_{max} \sigma^2 \log N}{n \lambda_n^2}}.
\end{aligned}$$

Now, applying our condition bounding n, N via ν and θ_u, we have

$$\frac{1}{\lambda_n} \mathbb{E}[\max_j Z_j(v^*)] < (1 + \delta') \sqrt{1 + \delta} \sqrt{\epsilon^2 \left(1 - \frac{\nu \log N}{n - s - 1}\right) + \frac{2C_{max} \sigma^2 \log N}{n \lambda_n^2}}.$$

Recall that by assumption, as $n, N \to +\infty$, we have that $\frac{\log N}{n \lambda_n^2}$ and $\frac{\log N}{n - s - 1}$ converge to zero. Consequently, the RHS converges to $(1 + \delta') \sqrt{(1 + \delta)} \epsilon$ as $n, N \to \infty$. Hence, we have

$$\lim_{n \to +\infty} \frac{1}{\lambda_n} \mathbb{E}[\max_j Z_j(v^*)] < (1 + \delta') \sqrt{1 + \delta}\, \epsilon.$$

Since $\delta' > 0$ and $\delta > 0$ were arbitrary, the result follows.

D.4 Proof of Lemma 5

Consider the function $f : \mathbb{R}^N \to \mathbb{R}$ given by

$$f(w) := \max_{1 \leq j \leq N} \left[\sqrt{v^* \Sigma_{(S^c|S)}}\, w \right],$$

where $\Sigma_{(S^c|S)} := \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}$. By construction, for a Gaussian random vector $V \sim N(0, I)$, we have $f(V) \stackrel{d}{=} \max_{j \in S^c} \widetilde{Z}_j$.

We now bound the Lipschitz constant of f. Let $R = \sqrt{\Sigma_{(S^c|S)}}$. For each $w, v \in \mathbb{R}^N$ and index $j = 1, \ldots, N$, we have

$$\begin{aligned} \left|[\sqrt{v^*}Rw]_j - [\sqrt{v^*}Rv]_j\right| &\leq \sqrt{v^*}\left|\sum_k R_{jk}[w_k - v_k]\right| \\ &\leq \sqrt{v^*}\sqrt{\sum_k R_{jk}^2}\,\|w - v\|_2 \\ &\leq \sqrt{v^*}\|w - v\|_2, \end{aligned}$$

where the last inequality follows since $\sum_k R_{jk}^2 = [\Sigma_{(S^c|S)}]_{jj} \leq 1$. Therefore, by Gaussian concentration of measure for Lipschitz functions [24, 27], we conclude that for any $\eta > 0$, it holds that

$$\begin{aligned} \mathbb{P}[f(W) \geq \mathbb{E}[f(W)] + \eta] &\leq \exp\left(-\frac{\eta^2}{2v^*}\right), \qquad \text{and} \\ \mathbb{P}[f(W) \leq \mathbb{E}[f(W)] - \eta] &\leq \exp\left(-\frac{\eta^2}{2v^*}\right). \end{aligned}$$

D.5 Proof of Lemma 6

Since the matrix $X_S^T X_S$ is Wishart with n degrees of freedom, using properties of the inverse Wishart distribution, we have $\mathbb{E}[(X_S^T X_S)^{-1}] = \frac{(\Sigma_{SS})^{-1}}{n-s-1}$ (see Lemma 7.7.1 of Anderson [1]). Thus, we compute

$$\begin{aligned} \mathbb{E}[Y_i] &= \frac{-\lambda_n\, n}{n-s-1} e_i^T\,(\Sigma_{SS})^{-1}\,\vec{b}, \qquad \text{and} \\ \mathbb{E}[Y_i'] &= \frac{\sigma^2}{n}\frac{n}{n-s-1}\; e_i^T(\Sigma_{SS})^{-1}e_i \;=\; \frac{\sigma^2}{n-s-1}e_i^T(\Sigma_{SS})^{-1}e_i. \end{aligned}$$

Moreover, using formulae for second moments of inverse Wishart matrices (see, e.g., Siskind [31]), we compute for all $n > s+3$

$$\begin{aligned} \mathbb{E}[Y_i^2] &= \frac{\lambda_n^2\, n^2}{(n-s)\,(n-s-3)} \left[\left(e_i^T(\Sigma_{SS})^{-1}\vec{b}\right)^2 + \frac{1}{n-s-1}\left(\vec{b}^T(\Sigma_{SS})^{-1}\vec{b}\right)\left(e_i^T(\Sigma_{SS})^{-1}e_i\right)\right] \\ \mathbb{E}[(Y_i')^2] &= \frac{\sigma^4 n^2}{(n-s-1)^2\,(n-s)\,(n-s-3)}\;\left(e_i^T(\Sigma_{SS})^{-1}e_i\right)^2\,\left[1 + \frac{1}{n-s-1}\right]. \end{aligned}$$

We now compute and bound the variance of Y_i. Setting $A_i = e_i^T(\Sigma_{SS})^{-1}\vec{b}$ and $B_i = e_i^T(\Sigma_{SS})^{-1}\vec{b}$ for shorthand, we have

$$\begin{aligned} \operatorname{var}(Y_i) &= \frac{\lambda_n^2\, n^2}{(n-s)\,(n-s-3)}\;\left[A_i^2 + \frac{1}{n-s-1}A_iB_i\right] - \frac{\lambda_n^2\, n^2}{(n-s-1)^2}A_i^2 \\ &= \frac{\lambda_n^2 n^2}{(n-s)\,(n-s-3)}\left[A_i^2\,\left(1 - \frac{(n-s)\,(n-s-3)}{(n-s-1)^2}\right) + \frac{1}{n-s-1}A_iB_i\right] \\ &\leq 2\lambda_n^2\left[\frac{3A_i^2}{n-s} + \frac{A_iB_i}{n-s-1}\right] \end{aligned}$$

for n sufficiently large. Using the bound $\|(\Sigma_{SS})^{-1}\|_\infty \leq D_{\max}$, we see that the quantities A_i and B_i are uniformly bounded for all i. Hence, we conclude that, for n sufficiently large, the variance is bounded as

$$\mathrm{var}(Y_i) \;\leq\; \frac{K\lambda_n^2}{n-s} \tag{49}$$

for some fixed constant K independent of s and n.

Now since $|\mathbb{E}[Y_i]| \leq \frac{2D_{\max}\lambda_n n}{n-s-1}$, we have

$$|Y_i - \mathbb{E}[Y_i]| \;\geq\; |Y_i| - |\mathbb{E}[Y_i]| \;\geq\; |Y_i| - \frac{2D_{\max}\lambda_n n}{n-s-1}.$$

Consequently, making use of Chebyshev's inequality, we have

$$\begin{aligned}
\mathbb{P}[|Y_i| \geq \frac{6D_{\max}\lambda_n n}{n-s-1}] &= \mathbb{P}[|Y_i| - \frac{2D_{\max}\lambda_n n}{n-s-1} \geq \frac{4D_{\max}\lambda_n n}{n-s-1}] \\
&\leq \mathbb{P}[|Y_i - \mathbb{E}[Y_i]| \geq \frac{4D_{\max}\lambda_n n}{n-s-1}] \\
&\leq \frac{\mathrm{var}(Y_i)}{16 D_{\max}^2 \lambda_n^2} \\
&\leq \frac{K}{16 D_{\max}\,(n-s)},
\end{aligned}$$

where the final step uses the bound (49).

We now compute and bound the variance of Y_i'. We have

$$\begin{aligned}
\mathrm{var}(Y_i') &= \frac{\sigma^4 n^2}{(n-s-1)^2\,(n-s)\,(n-s-3)} \left(A_i^2 \left[1 + \frac{1}{n-s-1}\right]\right) - \frac{\sigma^4}{(n-s-1)^2} A_i^2 \\
&= \frac{\sigma^4 n^2}{(n-s-1)^2\,(n-s)\,(n-s-3)} \left(A_i^2 \left[1 + \frac{1}{n-s-1} - \frac{(n-s)\,(n-s-3)}{n^2}\right]\right) \\
&\leq \frac{K\sigma^4}{(n-s-1)^3}
\end{aligned}$$

for some constant K independent of s and n. Consequently, applying Chebyshev's inequality, we have

$$\begin{aligned}
\mathbb{P}[Y_i' \geq 2\mathbb{E}[Y_i']] \;=\; \mathbb{P}[Y_i' - \mathbb{E}[Y_i'] \geq \mathbb{E}[Y_i']] &\leq \frac{\mathrm{var}(Y_i')}{(\mathbb{E}[Y_i'])^2} \\
&\leq \frac{K}{(n-s-1)^3} \frac{1}{\frac{\sigma^4}{n^2} e_i^T (\Sigma_{SS})^{-1} e_i} \\
&\leq \frac{K n^2 C_{max}}{\sigma^4 (n-s-1)^3} \\
&\leq \frac{K'}{n-s-1}
\end{aligned}$$

for some constant K' independent of s and n.

D.6 Proof of Lemma 7

As in the proof of Lemma 4, we define and bound

$$\Delta_Z(i,j) \;\; := \;\; \mathbb{E}[(Z_i - Z_j)^2] \;\leq\; 2C_{max}v^*.$$

Now let $(X_1, \ldots, X_N)$ be an i.i.d. zero-mean Gaussian vector with $\text{var}(X_i) = C_{max}v^*$, so that $\Delta_X(i,j) := \mathbb{E}[(X_i - X_j)^2] = 2C_{max}v^*$. If we set

$$\Delta^* \;\; := \;\; \max_{i,j \in S^c} |\Delta_X(i,j) - \Delta_Z(i,j)|\,,$$

then, by applying a known error bound for the Sudakov-Fernique inequality [5], we are guaranteed that

$$\mathbb{E}[\max_{j \in S^c} Z_j] \;\; \geq \;\; \mathbb{E}[\max_{j \in S^c} X_j] - \sqrt{\Delta^* \log N}. \tag{50}$$

We now show that the quantity Δ^* is upper bounded by

$$\Delta^* \;\; \leq \;\; 2v^* \, (C_{max} - \frac{1}{C_{max}}).$$

Using the inversion formula for block-partitioned matrices [22], we have

$$\Sigma_{(S^c|S)} \; := \; \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c} \; = \; \left[\Sigma^{-1}\right]_{S^cS^c}.$$

Consequently, we have the lower bound

$$\begin{aligned}
\mathbb{E}[(Z_i - Z_j)^2] \;\; &= \;\; v^*(e_i - e_j)^T \Sigma_{(S^c|S)}(e_i - e_j) \\
&\geq \;\; 2v^* \Lambda_{min}(\Sigma_{(S^c|S)}) \\
&\geq \;\; 2v^* \Lambda_{min}(\Sigma^{-1}) \\
&= \;\; \frac{2v^*}{C_{max}}.
\end{aligned}$$

In turn, this leads to the upper bound

$$\begin{aligned}
\Delta^* \;\; &= \;\; \max_{i,j \in S^c} |\Delta_X(i,j) - \Delta_Z(i,j)| \\
&= \;\; \max_{i,j \in S^c} [2v^* C_{max} - \Delta_Z(i,j)] \\
&\leq \;\; 2v^* \left(C_{max} - \frac{1}{C_{max}} \right).
\end{aligned}$$

We now analyze the behavior of $\mathbb{E}[\max_{j \in S^c} X_j]$. Using asymptotic results on the extrema of i.i.d. Gaussian sequences [19, 7], we have $\lim_{N \to +\infty} \frac{\mathbb{E}[\max_{j \in S^c} X_j]}{\sqrt{2C_{max}v^* \log N}} = 1$. Consequently, for all $\delta' > 0$, there exists an $N(\delta')$ such that for all $N \geq N(\delta')$, we have

$$\mathbb{E}[\max_{j \in S^c} X_j] \;\; \geq \;\; (1 - \delta')\sqrt{2C_{max}v^* \log N}.$$

Applying this lower bound to the bound (50), we have

$$\begin{aligned}
\frac{1}{\lambda_n}\mathbb{E}[\max_{j\in S^c} Z_j] &\geq \frac{1}{\lambda_n}\left[(1-\delta')\sqrt{2C_{max}v^*\log N} - \sqrt{\Delta^* \log N}\right] \\
&\geq \frac{1}{\lambda_n}\left[(1-\delta')\sqrt{2C_{max}v^*\log N} - \sqrt{2\,v^*\,(C_{max} - \frac{1}{C_{max}})\,\log N}\right] \\
&= \left[(1-\delta')\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}\right]\sqrt{2\frac{v^*}{\lambda_n^2}\log N}. \qquad (51)
\end{aligned}$$

First, assume that λ_n^2/v^* does not diverge to infinity. Then, there exists some $\alpha > 0$ such that $\frac{\lambda_n^2}{v^*} \leq \alpha$ for all sufficiently large n. In this case, we have from the bound (51) that

$$\frac{1}{\lambda_n}\mathbb{E}[\max_{j\in S^c} Z_j] \geq \gamma\sqrt{\log N}$$

where $\gamma := \left[(1-\delta')\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}\right]\frac{1}{\sqrt{\alpha}} > 0$. (Note that by choosing $\delta' > 0$ sufficiently small, we can always guarantee that $\gamma > 0$, since $C_{max} \geq 1$.) This completes the proof of condition (b) in the lemma statement.

Otherwise, we may assume that $\lambda_n^2/v^* \to +\infty$. We compute

$$\begin{aligned}
\frac{1}{\lambda_n}\sqrt{2v^*\log N} &= \sqrt{1-\delta}\sqrt{\frac{2\log N}{n-s-1}\vec{b}^T(\Sigma_{SS})^{-1}\vec{b} + \frac{2\sigma^2\,(1-\frac{s}{n})\log N}{n\lambda_n^2}} \\
&\geq \sqrt{1-\delta}\sqrt{\frac{2\log N}{n-s-1}\vec{b}^T(\Sigma_{SS})^{-1}\vec{b}} \\
&\geq \sqrt{\frac{1-\delta}{C_{max}}}\sqrt{\frac{2s\log N}{n-s-1}}.
\end{aligned}$$

We now apply the condition

$$\frac{2s\log N}{n-s-1} > \frac{1}{\theta_\ell - \nu} = C_{max}\,(2-\epsilon)^2/\left[\left[\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}\right]^2 - \nu C_{max}(2-\epsilon)^2\right]$$

to obtain that

$$\frac{1}{\lambda_n}\mathbb{E}[\max_{j\in S^c} Z_j] \geq \sqrt{(1-\delta)}\,\frac{(1-\delta')\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}}{\sqrt{\left[\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}\right]^2 - \nu C_{max}\,(2-\epsilon)^2}}\,(2-\epsilon) \qquad (52)$$

Recall that $\nu C_{max}\,(2-\epsilon)^2 > 0$ is fixed, and moreover that $\delta, \delta' > 0$ are arbitrary. Let $F(\delta, \delta')$ be the lower bound on the RHS (52). Note that F is a continuous function, and moreover that

$$F(0,0) = \frac{\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}}{\sqrt{\left[\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}}\right]^2 - \nu C_{max}\,(2-\epsilon)^2}}\,(2-\epsilon) > (2-\epsilon).$$

Therefore, by the continuity of F, we can choose $\delta, \delta' > 0$ sufficiently small to ensure that for some $\gamma > 0$, we have $\frac{1}{\lambda_n}\mathbb{E}[\max_{j\in S^c} Z_j] \geq (2-\epsilon)\,(1+\gamma)$ for all sufficiently large n.

D.7 Proof of Lemma 8

This claim follows from the proof of Lemma 5.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1984.

[2] P. J. Bickel and K. A. Doksum. *Mathematical statistics: basic ideas and selected topics.* Prentice Hall, Upper Saddle River, N.J., 2001.

[3] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. Technical report, Applied and Computational Mathematics, Caltech, 2004.

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

[5] S. Chatterjee. An error bound in the Sudakov-Fernique inequality. Technical report, UC Berkeley, October 2005. arXiv:math.PR/0510424.

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

[7] H. A. David and H. N. Nagaraja. *Order Statistics.* Wiley Series in Probability and Statistics. Wiley, New York, 2003.

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

[9] D. Donoho. For most large undetermined system of linear equations the minimal l_1-norm near-solution is also the sparsest solution. Technical report, Statistics Department, Stanford University, 2004.

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

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

[12] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Info Theory*, 47(7):2845–2862, 2001.

[13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[14] 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.

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

[16] 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.

[17] J. J. Fuchs. Recovery of exact sparse representations in the presence of noise. In *ICASSP*, volume 2, pages 533–536, 2004.

[18] J. J. Fuchs. Recovery of exact sparse representations in the presence of noise. *IEEE Trans. Info. Theory*, 51(10):3601–3608, October 2005.

[19] J. Galambos. *The Asymptotic Theory of Extreme Order Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1978.

[20] R. M. Gray. Toeplitz and Circulant Matrices: A Review. Technical report, Stanford University, Information Systems Laboratory, 1990.

[21] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*, volume 1. Springer-Verlag, New York, 1993.

[22] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[23] K. Knight and W. J. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

[24] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[25] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[26] 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.

[27] P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

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

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

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

[31] V. Siskind. Second moments of inverse Wishart-matrix elements. *Biometrika*, 59(3):690–691, December 1972.

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

[33] J. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Info Theory*, 50(10):2231–2242, 2004.

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

[35] 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.