Sharp thresholds for high-dimensional and noisy recovery of sparsity using $\ell_1$-constrained quadratic programming

Martin J. Wainwright
Department of Statistics, and
Department of Electrical Engineering and Computer Sciences,
UC Berkeley, Berkeley, CA 94720

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 signal denoising, sparse approximation, compressed sensing, and model selection. We analyze the behavior of $\ell_1$-constrained quadratic programming (QP), also referred to as the Lasso, for recovering the sparsity pattern. Our main result is to establish precise conditions on the problem dimension $p$, the number $k$ of non-zero elements in $\beta^*$, and the number of observations $n$ that are necessary and sufficient for subset selection using the Lasso. For a broad class of Gaussian ensembles satisfying mutual incoherence conditions, we establish existence and compute explicit values of thresholds $0 < \theta_\ell \leq 1 \leq \theta_u < +\infty$ with the following properties: for any $\delta > 0$, if $n > 2 (\theta_u - \delta) k \log(p - k)$, then the Lasso succeeds in recovering the sparsity pattern with probability converging to one for large problems, whereas for $n < 2 (\theta_\ell - \delta) k \log(p - k)$, 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 precise threshold $n = 2 k \log(p - k)$ is exactly determined.

Keywords: Convex relaxation; $\ell_1$-constraints; sparse approximation; signal denoising; subset selection; compressed sensing; model selection; high-dimensional inference; thresholds.

1 Introduction

The problem of recovering the sparsity pattern of an unknown vector $\beta^*$—that is, the positions of the non-zero entries of $\beta^*$—based on noisy observations arises in a broad variety of contexts, including subset selection in regression \[27\], compressed sensing \[9, 4\], structure estimation in graphical models \[26\], sparse approximation \[8\], and signal denoising \[7\]. A natural optimization-theoretic formulation of this problem is via $\ell_0$-minimization, where the $\ell_0$ “norm” of a vector corresponds to the number of non-zero elements. Unfortunately, however, $\ell_0$-minimization problems are known to be NP-hard in general \[28\], so that the existence of polynomial-time algorithms is highly unlikely. This challenge motivates the use of computationally tractable approximations or relaxations to $\ell_0$ minimization. In particular, a great deal of research over the past decade has studied the use of the $\ell_1$-norm as a computationally tractable surrogate to the $\ell_0$-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

$$Y_k = x_k^T \beta^* + W_k, \quad k = 1, \ldots, n, \quad (1)$$

\[1\] This work was posted in June 2006 as Technical Report 709, Department of Statistics, UC Berkeley, and on arxiv.org as CS.IT/0605740. It was presented in part at the Allerton Conference on Control, Communication and Computing in September 2006.
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 \( \beta^* \) is sparse, in that its support

\[
S := \{ i \in \{1, \ldots, p\} \mid \beta^*_i \neq 0 \}
\]

has relatively small cardinality \( k = |S| \). Given the observation model (1) and sparsity assumption (2), a reasonable approach to estimating \( \beta^* \) is by solving the \( \ell_1 \)-constrained quadratic program (QP), known as the Lasso in the statistics literature [30], given by

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

where \( \rho_n > 0 \) is a regularization parameter. Equivalently, the convex program (3) can be reformulated as the \( \ell_1 \)-constrained quadratic program

\[
\min_{\beta \in \mathbb{R}^p} \sum_{k=1}^n \| Y_k - x_k^T \beta \|_2^2, \quad \text{such that} \quad \| \beta \|_1 \leq C_n
\]

where the regularization parameter \( \rho_n \) and constraint level \( C_n \) are in one-to-one correspondence via Lagrangian duality. Of interest are conditions on the ambient dimension \( p \), the sparsity index \( k \), and the number of observations \( n \) for which it is possible (or impossible) to recover the support set \( S \) of \( \beta^* \) using this type of \( \ell_1 \)-constrained quadratic programming.

1.1 Overview of previous work

Recent years have witnessed a great deal of work on the use of \( \ell_1 \) constraints for subset selection and/or estimation in the presence of sparsity constraints. Given this substantial literature, we provide only a brief (and hence necessarily incomplete) overview here, with emphasis on previous work most closely related to our results. In the noiseless version \( (\sigma^2 = 0) \) of the linear observation model (1), one can imagine estimating \( \beta^* \) by solving the problem

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

This problem is in fact a linear program (LP) in disguise, and corresponds to a method in signal processing known as basis pursuit, pioneered by Chen et al. [7]. For the noiseless setting, the interesting regime is the underdetermined setting (i.e., \( n < p \)). With contributions from a broad range of researchers [3, 7, 15, 16, 25, 31], there is now a fairly complete understanding of the conditions on deterministic vectors \( \{x_k\} \) and sparsity indices \( k \) that ensure that the true solution \( \beta^* \) can be recovered exactly using the LP relaxation (5).

Most closely related to the current paper—as we discuss in more detail in the sequel—are recent
results by Donoho \cite{11}, as well as Candes and Tao \cite{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 $k \leq \alpha p$ can be recovered with high probability. These initial results have been sharpened in subsequent work by Donoho and Tanner \cite{13}, who show that the basis pursuit LP \cite{5} exhibits phase transition behavior, and provide precise information on the location of the threshold. The results in this paper are similar in spirit but applicable to the case of noisy observations: for a class of Gaussian measurement ensembles including the standard one $x_k \sim N(0, I_p)$ as a special case, we show that the Lasso quadratic program \cite{3} also exhibits a phase transition in its failure/success, and provide precise information on the location of the threshold.

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. The $\ell_1$-constrained quadratic program \cite{3}, known as the Lasso in the statistics literature \cite{30, 14}, has been the focus of considerable research in recent years. Knight and Fu \cite{22} analyze the asymptotic behavior of the optimal solution, not only for $\ell_1$ regularization but for $\ell_q$-regularization with $q \in (0, 2]$. Other work focuses more specifically on the recovery of sparse vectors in the high-dimensional setting. In contrast to the noiseless setting, there are various error metrics that can be considered in the noisy case, including:

- various $\ell_p$ norms $\mathbb{E}\|\hat{\beta} - \beta^*\|_p$, especially $\ell_2$ and $\ell_1$;
- some measurement of predictive power, such as the mean-squared error $\mathbb{E}[\|Y_i - \hat{Y}_i\|^2]$, where $\hat{Y}_i$ is the estimate based on $\hat{\beta}$; and
- a model selection criterion, meaning the correct recovery of the subset $S$ of non-zero indices.

One line of work has focused on the analysis of the Lasso and related convex programs for deterministic measurement ensembles. Fuchs \cite{17} investigates optimality conditions for the constrained QP \cite{3}, and provides deterministic conditions, of the mutual incoherence form, under which a sparse solution, which is known to be within $\epsilon$ of the observed values, can be recovered exactly. Among a variety of other results, both Tropp \cite{32} and Donoho et al. \cite{12} also provide sufficient conditions for the support of the optimal solution to the constrained QP \cite{3} to be contained within the true support of $\beta^*$. Also related to the current paper is recent work on the use of the Lasso for model selection, both for random designs by Meinshausen and Buhlmann \cite{26} and deterministic designs by Zhao and Yu \cite{35}. Both papers established that when suitable mutual incoherence conditions are imposed on either random \cite{26} or deterministic design matrices \cite{35}, then the Lasso can recover the sparsity pattern with high probability for a specific regime of $n$, $p$ and $k$. In this paper, we present more general sufficient conditions for both deterministic and random designs, thus recovering these previous scalings as special cases. In addition, we derive a set of necessary conditions for random designs, which allow us to establish a threshold result for the Lasso. Another line of work has ana-
alyzed the use of the Lasso \[3, 10\], as well as other closely related convex relaxations (e.g., the Dantzig selector \[5\]) when applied to random ensembles with measurement vectors drawn from the standard Gaussian ensemble \(x_k \sim N(0, I_p \times p)\). These papers either provide conditions under which estimation of a noise-contaminated signal via the Lasso is stable in the \(\ell_2\) sense \[3, 10\], or bounds on the MSE prediction error \[5\]. However, stability results of this nature do not guarantee exact recovery of the underlying sparsity pattern, according to the model selection criterion that we consider in this paper.

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:

\[
X := \begin{bmatrix}
    x_1^T \\
    x_2^T \\
    \vdots \\
    x_n^T
\end{bmatrix}
= \begin{bmatrix}
    X_1 & X_2 & \cdots & X_p
\end{bmatrix}. \tag{6}
\]

Consider the (random) set \(S(X, \beta^*, W, \rho_n)\) 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

\[
\text{sgn}(\beta_i) := \begin{cases}
+1 & \text{if } \beta_i > 0 \\
-1 & \text{if } \beta_i < 0 \\
0 & \text{if } \beta_i = 0.
\end{cases} \tag{7}
\]

Of interest is the event that the Lasso solution set contains a vector that recovers the sparsity pattern of the fixed underlying \(\beta^*\).

**Property \(R(X, \beta^*, W, \rho_n)\): There exists a solution \(\hat{\beta} \in S(X, \beta^*, W, \rho_n)\) such that \(\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)\).**

This analysis in this paper applies to high-dimensional setting, based on sequences of models indexed by \((p, k)\) whose dimension \(p = p(n)\) and sparsity level \(k = k(n)\) are allowed to grow with the number of observations. In this paper, we allow for completely general scaling of the triplet \((n, p, k)\), so that the analysis applies to different sparsity regimes, including linear sparsity \((k = \alpha p\) for some \(\alpha > 0\)), as well as sublinear sparsity (meaning that \(k/p \to 0\)). We begin by providing sufficient conditions for Lasso-based recovery to succeed with high probability (over the observation noise) when applied to deterministic designs. Moving to the case of random designs, we then sharpen this analysis by proving thresholds in the success/failure probability of the Lasso for various classes of Gaussian random measurement ensembles. In particular, for suitable measurement ensembles,
we prove that there exist fixed constants $0 < \theta \ell \leq 1$ and $1 \leq \theta u < +\infty$ such that for all $\delta > 0$, the following properties hold. In terms of sufficiency, we show that it is always possible to choose the regularization parameter $\rho_n$ such that the Lasso recovers the sparsity pattern with probability converging to one (over the choice of noise vector $W$ and random matrix $X$) whenever

$$n > 2(\theta_u + \delta) k \log(p - k).$$

(8)

Conversely, for whatever regularization parameter $\rho_n > 0$ is chosen, the Lasso fails to recover with probability converging to one whenever the number of observations $n$ satisfies

$$n < 2(\theta \ell - \delta) k \log(p - k).$$

(9)

Although negative results of this type have been established for the basis pursuit LP in the noiseless setting [13], to the best of our knowledge, the condition (9) is the first result on necessary conditions for exact sparsity recovery in the noisy setting.

![Figure 1](image_url)

**Figure 1.** Plots of the number of data samples $n = 2\theta k \log(p - k)$, indexed by the control parameter $\theta$, 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: $k(p) = \alpha p$. (b) Sublinear sparsity index $k(p) = \alpha p / \log(\alpha p)$. (c) Fractional power sparsity index $k(p) = \alpha p^\gamma$ with $\gamma = 0.75$. The threshold in Lasso success probability occurs at $\theta^* = 1$, consistent with Theorem 1. See Section 4 for further details.

For the special case of the uniform Gaussian ensemble (i.e., $x_k \sim N(0, I_p)$) considered in past work, we show that $\theta \ell = \theta u = 1$, so that the threshold is sharp. Figure 1 provides experimental confirmation of the accuracy of this threshold behavior for finite-sized problems, with dimension $p$ ranging from 128 to 512. 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 sparsity (i.e., $k/p \to \alpha$ for some $\alpha > 0$), this theorem provides a noisy analog of results previously established for basis pursuit in the noiseless case [11, 4, 13].
Moreover, our result can also be adapted to an entirely different scaling regime, in which the sparsity index is sublinear \((k/p \to 0)\), as considered by a separate body of recent work \([26, 35]\) 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 \(\mathcal{R}(X, \beta^*, W, \rho_n)\) 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 Preliminary analysis and deterministic designs

In this section, we provide necessary and sufficient conditions for Lasso to successfully recover the sparsity pattern—or in shorthand, for property \(\mathcal{R}(X, \beta^*, W, \rho_n)\) to hold. Based on these conditions, we then define collections of random variables that play a central role in our analysis. In particular, characterizing the event \(\mathcal{R}(X, \beta^*, W, \rho_n)\) is reduced to studying 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 Convex optimality conditions

We begin with a simple set of necessary and sufficient conditions for property \(\mathcal{R}(X, \beta^*, W, \rho_n)\) to hold, which follow in a straightforward manner from optimality conditions for convex programs \([20]\).

We define \(S := \{i \in \{1, \ldots, p\} \mid \beta^*_i \neq 0\}\) to be the support of \(\beta^*\), 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 regularization parameter \(\rho > 0\) and noise vector \(w \in \mathbb{R}^n\), property \(\mathcal{R}(X, \beta^*, W, \rho_n)\) holds if and only if

\[
X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \left[\frac{1}{n} X_S^T w - \rho \text{sgn}(\beta^*_S)\right] - \frac{1}{n} X_{S^c}^T w \leq \rho, \quad \text{and} \quad \left(10a\right)
\]

\[
\text{sgn}\left\{\beta^*_S + \left(\frac{1}{n} X_S^T X_S\right)^{-1} \left[\frac{1}{n} X_S^T w - \rho \text{sgn}(\beta^*_S)\right]\right\} = \text{sgn}(\beta^*_S), \quad \left(10b\right)
\]

where both of these vector relations should be taken elementwise. Moreover, if inequality \((10a)\) holds strictly, then any solution \(\hat{\beta}\) of the Lasso \((3)\) satisfies \(\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)\).

See Appendix A for the proof of this claim. For shorthand, define \(\vec{b} := \text{sgn}(\beta^*_S)\), 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:

\[ U_i := e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} \left[ \frac{1}{n} X_S^T W - \rho_n \tilde{b} \right] \quad (11a) \]

\[ V_j := X_j^T \left\{ X_S \left( \frac{1}{n} X_S^T X_S \right)^{-1} \rho_n \tilde{b} - X_S \left( \frac{1}{n} X_S^T X_S \right)^{-1} X_S^T - I_{n \times n} \right\} \frac{W}{n} \quad (11b) \]

From Lemma 1, the behavior of \( R(X, \beta^*, W, \rho_n) \) is determined by the behavior of \( \max_{j \in S^c} |V_j| \) and \( \max_{i \in S} |U_i| \). In particular, condition (10a) holds if and only if the event

\[ \mathcal{E}(V) := \left\{ \max_{j \in S^c} |V_j| \leq \rho_n \right\} \quad (12) \]

holds. On the other hand, if we define \( \mathcal{M}(\beta^*) := \min_{i \in S} |\beta^*_i| \), then the event

\[ \mathcal{E}(U) := \left\{ \max_{i \in S} |U_i| \leq \mathcal{M}(\beta^*) \right\} \quad (13) \]

is sufficient to guarantee that condition (10b) 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

\[ \left| X_S^T X_S \left( \frac{1}{n} X_S^T X_S \right)^{-1} \text{sgn}(\beta^*_S) \right| \leq 1 \quad (14a) \]

\[ \text{sgn} \left( \beta^*_S - \rho \left( \frac{1}{n} X_S^T X_S \right)^{-1} \text{sgn}(\beta^*_S) \right) = \text{sgn}(\beta^*_S). \quad (14b) \]

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

\[ \left\| X_S^T X_S \left( \frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \leq (1 - \epsilon) \quad \text{for some } \epsilon \in (0, 1], \text{ and} \quad (15a) \]

\[ \lambda_{\text{min}} \left( \frac{1}{n} X_S^T X_S \right) \geq C_{\text{min}} > 0, \quad (15b) \]

where \( \lambda_{\text{min}} \) denotes the minimal eigenvalue. Mutual incoherence conditions of the form (15a) have been considered in previous work on the Lasso, initially by Fuchs [17] and Tropp [32]. Other au-
Assume that $\beta \sim W$, chosen such that $\text{Proposition 1.}$

$deterministic designs: k \text{ order of } n, p, k \text{ constraint. If we consider general scaling of } (n, p, k) \text{ to increase simultaneously. As discussed earlier, past work [26, 33] has considered particular types of high-dimensional scaling. If we specialize Theorem 1 to a particular type of scaling considered in past work [33], then we recover the following corollary:}

Corollary 1. Suppose that $p = O(\exp(n^{c_3}))$ and $k = O(n^{c_1})$, and moreover that $M(\beta^*)^2 > 1/n^{(1-c_2)}$ with $0 < c_1 < c_2 < 1$. If we set $\rho_n^2 = 1/n^{1-c_4}$ for some $c_4 \in (c_3, c_2 - c_1)$, then under the conditions (15), the Lasso recovers the sparsity pattern with probability $1 - \exp(-Cn^{c_4})$.

Proof. We simply need to verify that the conditions of Proposition 1 are satisfied with these particular choices. By construction, we have $n\rho_n^2 = n^{c_4} \rightarrow +\infty$, so that condition (16)(a) holds. Similarly, we have $rac{\log(p-k)}{n\rho_n^2} = O\left(n^{c_3}/(n^{c_4-1}n)\right) = O(n^{c_3-c_4}) \rightarrow 0$ so that condition (16)(b) holds. Lastly, we have $\frac{\rho_n^2}{M(\beta^*)^2} = O\left(n^{c_4-1}n(1-c_2)\right) = O(n^{c_4+c_1-c_2}) \rightarrow 0$, so that condition (16)(c) holds.

The given scaling is interesting since it allows the model dimension $p$ to be exponentially larger than the number of observations ($p \gg n$). On the other hand, it imposes a very strong sparsity constraint. If we consider general scaling of $(n, p, k)$, Proposition 1 suggests that having $n$ on the order of $k \log(p-k)$ is appropriate. In Section 3, we show that this type of scaling is both necessary and sufficient for almost all matrices drawn from suitable Gaussian designs.
2.3 Proof of Proposition 1

Recall the events $\mathcal{E}(V)$ and $\mathcal{E}(U)$ defined in equations (12) and (13) respectively. To establish the claim, we must show that $\mathbb{P}[\mathcal{E}(V)^c] \to 0$, where $\mathcal{E}(V)^c$ and $\mathcal{E}(U)^c$ denote the complements of these events. By union bound, it suffices to show both $\mathbb{P}[\mathcal{E}(V)^c]$ and $\mathbb{P}[\mathcal{E}(U)^c]$ converge to zero, or equivalently that $\mathbb{P}[\mathcal{E}(V)]$ and $\mathbb{P}[\mathcal{E}(U)]$ both converge to one.

**Analysis of $\mathcal{E}(V)$:** We begin by establishing that $\mathbb{P}[\mathcal{E}(V)] \to 1$. Recalling our shorthand $\vec{b} := \text{sgn}(\beta^*)$ and the definition (11b) of the random variables $V_j$, note that $\mathcal{E}(V)$ holds if and only if $\min_{j \in S^c} V_j \geq -1$ and $\frac{\max_{j \in S^c} V_j}{\rho_n} \leq 1$. Moreover, we note that each $V_j$ is Gaussian with mean $\mu_j = \mathbb{E}[V_j] = \rho_n X_j^T X S (X_S^T X_S)^{-1} \vec{b}$. Using condition (15a), we have $|\mu_j| \leq (1 - \epsilon) \rho_n$ for all indices $j = 1, \ldots, (p - k)$, from which we obtain that

$$\frac{\max_{j \in S^c} V_j}{\rho_n} \leq (1 - \epsilon) + \frac{1}{\rho_n} \max_j \tilde{V}_j,$$

and

$$\frac{\min_{j \in S^c} V_j}{\rho_n} \geq -(1 - \epsilon) + \frac{1}{\rho_n} \min_j \tilde{V}_j,$$

where $\tilde{V}_j := X_j^T \left[ I_{n \times n} - X_S (X_S^T X_S)^{-1} X_S^T \right] \frac{W}{n}$ are zero-mean (correlated) Gaussian variables. Hence, in order to establish condition (10a) of Lemma 1, we need to show that

$$\mathbb{P} \left[ \frac{1}{\rho_n} \min_{j \in S^c} \tilde{V}_j < -\epsilon, \quad \text{or} \quad \frac{1}{\rho_n} \max_{j \in S^c} \tilde{V}_j > \epsilon \right] \to 0. \quad (17)$$

By symmetry (see Lemma 11 from Appendix C), it suffices to show that $\mathbb{P}\left[ \frac{\max_{j \in S^c} |\tilde{V}_j|}{\rho_n} > \epsilon \right] \to 0$.

Using Gaussian comparison results [24] (see Lemma 9 in Appendix B), we have $\frac{\mathbb{E}[\max_{j \in S^c} |\tilde{V}_j|]}{\rho_n} \leq \frac{3 \sqrt{\log(p - k)}}{\rho_n} \max_j \sqrt{\mathbb{E}[\tilde{V}_j^2]}$. Straightforward computation yields that

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

since the projection matrix $\Pi_S := I_{n \times n} - X_S (X_S^T X_S)^{-1} X_S^T$ has maximum eigenvalue equal to one, and $\max_j \|X_j\|^2 \leq n$ by assumption. Consequently, we have established that

$$\mathbb{E}[\max_{j \in S^c} |\tilde{V}_j|] \leq \sqrt{\frac{\sigma^2 \log(p - k)}{n}} \leq \frac{\epsilon \rho_n}{2}, \quad (18)$$

where inequality (i) follows from condition (16b). We conclude the proof by claiming that for all $\delta > 0$, we have

$$\mathbb{P}\left[ \max_{j \in S^c} |\tilde{V}_j| > \mathbb{E}[\max_{j \in S^c} |\tilde{V}_j|] + \delta \right] \leq \exp\left(\frac{-n \delta^2 \sigma^2}{2}\right). \quad (19)$$
To establish this claim, define a function \( f : \mathbb{R}^n \to \mathbb{R} \) by \( f(w) = \max_{j \in S^c} \left| \frac{\sigma^2 X_j}{n} \Pi_S w \right| \). Note that by construction, for a standard normal vector \( W \sim N(0, I_{n \times n}) \), we have \( f(W) \overset{d}{=} \max_{j \in S^c} |\tilde{V}_j| \). Consequently, the claim \((19)\) will follow from measure concentration for Lipschitz functions of Gaussian variates \([23]\), as long as we can suitably upper bound the \( \ell_2 \)-Lipschitz constant \( \| f \|_{\text{Lip}} \). By the triangle and Cauchy-Schwarz inequalities, we have

\[
f(w) - f(v) \leq \max_{j \in S^c} \left| \frac{\sigma^2 X_j}{n} \Pi_S (w - v) \right| = \sigma^2 \max_{j \in S^c} \left| \frac{X_j}{n} \right| \| \Pi_S \|_2 \left\| w - v \right\|_2 = \frac{\sigma^2}{\sqrt{n}} \left\| w - v \right\|_2,
\]

since \( \| \Pi_S \|_2 = 1 \) and \( \max_j \left| X_j \right| \leq \sqrt{n} \) by assumption. Hence, we have established that \( \| f \|_{\text{Lip}} \leq \frac{\sigma^2}{\sqrt{n}} \) so that the concentration \((19)\) follows.

Finally, if we set \( \delta = \frac{\epsilon \rho_n}{2} \), then we are guaranteed that \( \delta + \mathbb{E} \max_{j \in S^c} |\tilde{V}_j| \leq \rho_n \epsilon \), so that \( \mathbb{P} \left( \max_{j \in S^c} |\tilde{V}_j| > \epsilon \rho_n \right) \leq \exp(-\frac{n \rho_n^2 \epsilon^2}{4}) \) follows form the bound \((19)\). Consequently, condition \((16)\a\) in the theorem statement—namely, that \( n \rho_n^2 \to +\infty \)—suffices to ensure that \( \mathbb{P}(E(V)) \to 1 \) at rate \( \exp(-Cn\rho_n^2) \) as claimed.

**Analysis of \( E(U) \):** We now show that \( \mathbb{P}(E(U)) \to 1 \). Beginning with the triangle inequality, we upper bound \( \max_{i \in S} |U_i| := \left\| \left( \frac{1}{n} X_S^T X_S \right)^{-1} \frac{1}{n} X_S^T W - \rho_n \text{sgn}(\beta_S^*) \right\|_\infty \) as

\[
\max_{i \in S} |U_i| \leq \left\| \left( \frac{1}{n} X_S^T X_S \right)^{-1} \frac{1}{n} X_S^T W \right\|_\infty + \left\| \left( \frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \rho_n \quad (20)
\]

The second term in this expression is a deterministic quantity, which we bound as follows

\[
\rho_n \left\| \left( \frac{1}{n} X_S^T X_S \right)^{-1} \right\|_\infty \leq \rho_n \sqrt{k} \lambda_{\max} \left( \left( \frac{1}{n} X_S^T X_S \right)^{-1} \right) = \frac{\rho_n \sqrt{k}}{\lambda_{\min} \left( \frac{1}{n} X_S^T X_S \right)} \leq \frac{\rho_n \sqrt{k}}{C_{\min}}, \quad (21)
\]

where we have use condition \((15b)\) in the final step.

Turning to the first term in the expansion \((20)\), 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 \( Z_i := e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} \frac{1}{n} X_S^T W \). Each such \( Z_i \) is a zero-mean Gaussian; computing its variance yields \( \text{var}(Z_i) = \frac{\sigma^2}{n} e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} e_i \leq \frac{\sigma^2}{C_{\min}} \). Hence, by a standard Gaussian comparison theorem \([24]\) (in particular, see Lemma \(9\) in Appendix \(B\)), we have

\[
\mathbb{E} \max_{i \in S} |Z_i| = \mathbb{E} \left[ \left\| \left( \frac{1}{n} X_S^T X_S \right)^{-1} \frac{1}{n} X_S^T W \right\|_\infty \right] \leq 3 \sqrt{\frac{\sigma^2 \log k}{n C_{\min}}} \quad (22)
\]

Now putting together the pieces, recall the definition \( \mathcal{M}(\beta^*) := \min_{i \in S} |\beta_i^*| \). From the decomposition \((20)\) and the bound \((21)\), in order to have \( \max_{i \in S} |U_i| \leq \mathcal{M}(\beta^*) \), it suffices to have \( \frac{\rho_n \sqrt{k}}{C_{\min} \mathcal{M}(\beta^*)} \) be bounded above by constant (condition \((16)(c)\)), which we may take equal to \( \frac{1}{2} \) by rescaling \( \rho_n \) as
necessary. With this deterministic term bounded, it suffices to have \( P[\max_{i \in S} |Z_i| > \frac{M(\beta^*)}{2}] \) converge to zero. To bound this probability, we first claim that for any \( \delta > 0 \)
\[
P \left[ \max_{i \in S} |Z_i| > \mathbb{E} \left[ \max_{i \in S} |Z_i| \right] + \delta \right] \leq \exp \left( -n\delta^2 \frac{\sigma^2 C_{\min}}{2} \right). \tag{23}
\]
As in the previous argument, this follows from concentration of measure for Lipschitz functions for Gaussian random vectors, since \( \|e^T_i (\frac{1}{n} X_S^T X_S)^{-1} 1 1 X_S^T \|_2 \leq \frac{1}{C_{\min} \sqrt{n}} \).

Using the bound (22) and condition (16)(c), we have
\[
\frac{\mathbb{E}[\max_{i \in S} |Z_i|]}{M(\beta^*)} \leq 3 \sqrt{\frac{\sigma^2 \log k}{C_{\min} n [M(\beta^*)]^2}} = \mathcal{O} \left( \sqrt{\frac{\log k}{C_{\min} n \rho_n^2 k}} \right) = \mathcal{O} \left( \sqrt{\frac{1}{n \rho_n^2}} \right),
\]
which converges to zero using condition (16)(a) from the theorem statement. Hence, we may assume that \( \mathbb{E}[\max_{i \in S} |Z_i|] \leq \frac{M(\beta^*)}{4} \) for sufficiently large \( n \), and then take \( \delta = \frac{M(\beta^*)}{4} \) in the bound (23) to conclude that
\[
P[\mathcal{E}(U)^c] \leq P \left[ \max_{i \in S} |Z_i| > \frac{M(\beta^*)}{2} \right] \leq \exp \left( -nM(\beta^*)^2 \frac{\sigma^2 C_{\min}}{8} \right) = \mathcal{O} \left( \exp \left( -D \frac{n \rho_n^2 \sigma^2 C_{\min}}{8} \right) \right),
\]
for some finite constant \( D \), where the final inequality follows since \( \rho_n^2 = \mathcal{O}(M(\beta^*)^2) \) from condition (16)(c). Thus, we have shown that \( P[\mathcal{E}(U)] \to 1 \) at rate \( \exp(-Dn\rho_n^2) \) for a suitable constant \( D > 0 \) as claimed.

\[ \square \]

### 3 Recovery of sparsity: random Gaussian ensembles

The previous section treated the case of a deterministic design \( X \), which allowed for a relatively straightforward analysis. We now turn to the more complex case of random design matrices \( X \), in which each row \( x_k \) is chosen as an i.i.d. Gaussian random vector with covariance matrix \( \Sigma \). In this setting, we provide precise conditions that govern the success and failure of the Lasso over this ensemble; more specifically, we provide explicit thresholds that provide a sharp description of the failure/success of the Lasso as a function of \( (n, p, k) \). 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 \( \Sigma \) with unit diagonal, and with its minimum and maximum eigenvalues (denoted \( \lambda_{\min} \) and \( \lambda_{\max} \) respectively) bounded as

\[
\lambda_{\min}(\Sigma SS) \geq C_{\min}, \quad \text{and} \quad \lambda_{\max}(\Sigma) \leq C_{\max}
\]  

(24)

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

\[
\|(\Sigma_{S^c} \Sigma SS)^{-1}\|_\infty \leq (1 - \epsilon) \]

(25)

for some \( \epsilon \in (0,1] \). The simplest example of a covariance matrix satisfying these conditions is the identity \( \Sigma = I_{p \times p} \), for which we have \( C_{\min} = C_{\max} = D_{\max} = 1 \), and \( \epsilon = 1 \). Another well-known matrix family satisfying these conditions are Toeplitz matrices (see Appendix D for details).

Under these conditions, we consider the observation model

\[
Y_k = x_k^T \beta^* + W_k, \quad k = 1, \ldots, n,
\]  

(26)

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 \( M(\beta^*) := \min_{i \in S} |\beta^*_i| \), and the sparsity index \( k = |S| \).

**Theorem 1.** Consider a sequence of covariance matrices \( \{\Sigma[p]\} \) and solution vectors \( \{\beta^*[p]\} \) satisfying conditions (24) and (25). Under the observation model (26), consider a sequence \( (n, p(n), k(n)) \) such that \( k, (n-k) \) and \( (p-k) \) tend to infinity. Define the constants

\[
\theta_\ell := \frac{(\sqrt{C_{\max}} - \sqrt{C_{\max} - \frac{1}{C_{\max}}})^2}{C_{\max}(2 - \epsilon)^2} \leq 1, \quad \text{and} \quad \theta_u := \frac{C_{\max}}{\epsilon^2 C_{\min}} \geq 1.
\]  

(27)

Then for any fixed \( \delta > 0 \), we have the following

(a) If \( n < 2(\theta_\ell - \delta) k \log(p-k) \), then \( \mathbb{P}[R(X, \beta^*, W, \rho_n)] \to 0 \) for any monotonic sequence \( \rho_n > 0 \).

(b) Conversely, if \( n > 2(\theta_u + \delta) k \log(p-k) \), and \( \rho_n \to 0 \) is chosen such that

\[
(i) \quad \frac{\eta \rho_n^2}{\log(p-k)} \to +\infty, \quad \text{and} \quad (ii) \quad \frac{\rho_n}{\mathcal{M}(\beta^*)} = O(1)
\]  

(28)

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

**Remark:** Suppose for simplicity that \( \mathcal{M}(\beta^*) \) remains bounded away from 0. In this case, the requirements on \( \rho_n \) reduce to \( \rho_n \to 0 \), and \( \rho_n^2 n / \log(p-k) \to +\infty \). One suitable choice is \( \rho_n^2 = \frac{\log(k) \log(p-k)}{n} \),

12
with which we have
\[ \rho_n^2 = \left( \frac{k \log(p - k)}{n} \right) \frac{\log(k)}{k} = O \left( \frac{\log k}{k} \right) \to 0, \]
and \( \frac{n \rho_n^2}{\log(p - k)} = \log(k) \to +\infty. \) More generally, under the threshold scaling of \((n, p, k)\) given, the theorem allows the minimum value \( M(\beta^*) := \min_{i \in S} |\beta_i^*| \) to decay towards zero at rate \( \Omega(1/\sqrt{k}) \) but not faster.

### 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 \([4, 11]\) has focused on the uniform Gaussian ensemble in the noiseless \( (\sigma^2 = 0) \) and underdetermined setting \( (n = \gamma p \text{ for some } \gamma \in (0, 1)). \) Analyzing the asymptotic behavior of the linear program (5) for recovering \( \beta^* \), the basic result is that there exists some \( \alpha > 0 \) such that all sparsity patterns with \( k \leq \alpha p \) 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 \( C_{\text{min}} = C_{\text{max}} = 1, \) so that the threshold constants reduce
\[ \theta_\ell = \frac{(\sqrt{C_{\text{max}}} - \sqrt{C_{\text{max}} - \frac{1}{C_{\text{max}}}})^2}{C_{\text{max}} (2 - \epsilon)^2} = 1 \quad \text{and} \quad \theta_u = \frac{C_{\text{max}}^2}{\epsilon^2 C_{\text{min}}} = 1. \]
Consequently, Theorem 1 provides a sharp threshold for the behavior of the Lasso, in that failure/success is entirely determined by whether or not \( n > 2k \log(p - k). \) Thus, if we consider the linear scaling \( n = \Theta(p) \) analyzed in previous work on the noiseless case \([11, 4]\), we have:

**Corollary 2** (Linearly underdetermined setting). Suppose that \( n = \gamma p \) for some \( \gamma \in (0, 1). \) Then

(a) If \( k = \alpha p \) for any \( \alpha \in (0, 1), \) then \( \mathbb{P}[R(X, \beta^*, W, \rho_n)] \to 0 \) for any positive sequence \( \rho_n > 0. \)

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

Conversely, suppose that the size \( k \) of the support of \( \beta^* \) scales linearly with the number of parameters \( p. \) The following result describes the amount of data required for the \( \ell_1 \)-constrained QP to recover the sparsity pattern in the noisy setting \( (\sigma^2 > 0): \)
Corollary 3 (Linear fraction support). Suppose that $k = \alpha p$ for some $\alpha \in (0, 1)$. Then we require $n > 2\alpha p \log(1 - \alpha)p$ 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 (5) 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., $k = \alpha p$ as in the work [11, 4]) to a sublinear fraction (i.e., $k = O\left(\log p\right)$, as in Corollary 2).

Interestingly, an information-theoretic analysis of this sparsity recovery problem [33] shows that the optimal decoder—namely, an oracle that can search exhaustively over all $\binom{p}{k}$ subsets—can recover linear sparsity ($k = \alpha p$) with the number of observations scaling linearly in the problem size ($n = \Theta(p)$). This behavior, which contrasts dramatically with the Lasso threshold given in Theorem 1, raises an interesting question as to whether there exist computationally tractable methods for achieving this scaling.

3.2.2 Non-uniform Gaussian ensembles

We now consider more general (non-uniform) Gaussian ensembles that satisfy conditions (24) and (25). As mentioned earlier, previous papers treat model selection with the high-dimensional Lasso, both for deterministic designs [35] and random designs [26]. These authors assume eigenvalue and incoherence conditions analogous to (24) and (25); however, in contrast to the results given here, they impose rather specific scaling conditions on the triplet $(n, p, k)$. For instance, Meinshausen and Buhlmann [26] assumed $p(n) = O(n^\gamma)$ for some $\gamma > 0$, $k = O(n^\kappa)$ for some $\kappa \in (0, 1)$, and $M(\beta^*)^2 \geq \Omega(1/n^{1 - \xi})$ for some $\xi \in (0, 1)$, whereas (for deterministic designs) Zhao and Yu [35] allowed faster growth $p = O(\exp(n^{c_3}))$. Here we show that a corollary of Theorem 1 yields the success of Lasso under these particular scalings:

Corollary 4. Suppose that $p = O(\exp(n^{c_3}))$ and $k = O(n^{c_1})$, and $M(\beta^*)^2 \geq \Omega(1/n^{(1-c_2)})$ with $0 < c_1 + c_3 < c_2 < 1$. Then the Lasso recovers the sparsity pattern with probability converging to one, where the probability is taken over both the choice of random design $X$ and noise vector $W$.

Proof. We need to verify that the conditions (28) hold under this particular choice of scaling. First, note that under the given scaling, we have $\frac{n}{k \log(p-k)} > \Omega\left(\frac{n}{n^{c_1} n^{c_3}}\right) \to +\infty$, since $c_1 + c_3 < 1$. Thus, if we set $\rho_n^2 = \frac{\log(k) \log(p-k)}{n} \to 0$, we have $\frac{n^2 \rho_n^2}{\log(p-k)} = \log(k) \to +\infty$, so that condition (28)(i) holds. Secondly, we have

$$\frac{\rho_n^2}{M(\beta^*)^2} < n^{1-c_2} \frac{k \log(p-k)}{n} = \Omega\left(\frac{n^{c_1} n^{c_3}}{n^{c_2}}\right) \frac{\log(k)}{k} \to 0,$$

Under these assumptions, the authors in fact established fast rates for Lasso success.
since $c_1 + c_3 - c_2 < 0$, which shows that condition (28)(ii) holds strongly.

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 (11a) and (11b) 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 $k < n$, the random Gaussian matrix $X_S$ will have rank $k$ 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 E.1, concerns the behavior of the random vector $V = (V_1, \ldots, V_{p-k})$, when conditioned on $X_S$ and $W$. Recalling the shorthand notation $\tilde{b} := \text{sgn}(\beta^*)$, 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

$$|\mathbb{E}[V \mid W, X_S]| \leq \rho_n (1 - \epsilon) \mathbf{1}.$$ (29)

Moreover, its conditional covariance takes the form

$$\text{cov}[V \mid W, X_S] = M_n \Sigma_{(S^c \mid S)} = M_n \left[ \Sigma_{S^c S^c} - \Sigma_{S^c S} (\Sigma_{SS})^{-1} \Sigma_{SS^c} \right],$$ (30)

where

$$M_n := \rho_n^2 \tilde{b}^T (X_S^T X_S)^{-1} \tilde{b} + \frac{1}{n^2} W^T \left[ I_{n \times n} - X_S (X_S^T X_S)^{-1} X_S^T \right] W$$ (31)

is a random scaling factor.

The following lemma, proved in Appendix E.2, captures the behavior of the random scaling factor $M_n$ defined in equation (31):

Lemma 3. The random variable $M_n$ has mean

$$\mathbb{E}[M_n] = \frac{\rho_n^2}{n-k-1} \tilde{b}^T (\Sigma_{SS})^{-1} \tilde{b} + \frac{\sigma^2 (n-k)}{n^2}.$$ (32)

Moreover, it is concentrated: for any $\delta > 0$, we have $\mathbb{P} \left[ |M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n] \right] \to 0$ as $n \to +\infty$. 15
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 \( \mathcal{E}(V) \): 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 \( T(\delta) := \{|M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n]\} \). By conditioning on \( T(\delta) \) and its complement \( [T(\delta)]^c \), we have the upper bound

\[
\Pr[\max_{j \in S^c} |V_j| > \rho_n] \leq \Pr[\max_{j \in S^c} |V_j| > \rho_n | [T(\delta)]^c] + \Pr[T(\delta)].
\]

By the concentration statement in Lemma 3, we have \( \Pr[T(\delta)] \to 0 \), 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 \( \text{cov}(Z) = \text{cov}(V | X_S, W) \). We then have

\[
\max_{j \in S^c} |V_j| = \max_{j \in S^c} |\mu_j + Z_j| \\
\leq (1 - \epsilon)\rho_n + \max_{j \in S^c} |Z_j|,
\]

where we have used the triangle inequality, and the upper bound (29) on the mean. This inequality establishes the inclusion of events \( \{\max_{j \in S^c} |Z_j| \leq \epsilon \rho_n\} \subseteq \{\max_{j \in S^c} |V_j| \leq \rho_n\} \), thereby showing that it suffices to prove that \( \Pr[\max_{j \in S^c} |\tilde{Z}_j| > \epsilon \rho_n | [T(\delta)]^c] \to 0 \).

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

\[
\Pr[\max_{j \in S^c} |\tilde{Z}_j| > \epsilon \rho_n | [T(\delta)]^c] \leq \Pr[\max_{j \in S^c} |\tilde{Z}_j| > \epsilon \rho_n],
\]

where \( \tilde{Z} \) is zero-mean Gaussian with covariance \( M^*_n \Sigma(S^c|S) \).

Using Lemma 11 it suffices to show that \( \Pr[\max_{j \in S^c} \tilde{Z}_j > \epsilon \rho_n] \) converges to zero. Accordingly, we complete this part of the proof via the following two lemmas, both of which are proved in Appendix E.

**Lemma 4.** Under the stated assumptions of the theorem, we have \( \frac{M^*_n}{\rho_n^2} \to 0 \) and

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

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

\[
\Pr[\max_{j \in S^c} \tilde{Z}_j > \eta + \mathbb{E}[\max_{j \in S^c} \tilde{Z}_j]] \leq \exp\left(-\frac{\eta^2}{2M^*_n}\right).
\]  

(33)

Lemma 4 implies that for all \( \delta > 0 \), we have \( \mathbb{E}[\max_{j \in S^c} \tilde{Z}_j] \leq (1 + \frac{\delta}{2})\epsilon \rho_n \) for all \( n \) sufficiently
large. Therefore, setting $\eta = \frac{\delta}{2} \rho_n \epsilon$ in the bound (33), we have for fixed $\delta > 0$ and $n$ sufficiently large:

$$
P \left[ \max_{j \in S'} \tilde{Z}_j > (1 + \delta) \rho_n \epsilon \right] \leq P \left[ \max_{j \in S'} \tilde{Z}_j > \frac{\delta}{2} \rho_n \epsilon + E[ \max_{j \in S'} \tilde{Z}_j ] \right] \leq 2 \exp \left( -\frac{\delta^2 \rho_n^2 \epsilon^2}{8 M_n^*} \right).$$

From Lemma 4, we have $\rho_n^2 / M_n^* \rightarrow +\infty$, which implies that $P[ \max_{j \in S'} \tilde{Z}_j > (1 + \delta) \rho_n \epsilon ] \rightarrow 0$ for all $\delta > 0$. By the arbitrariness of $\delta > 0$, we thus have $P[ \max_{j \in S'} \tilde{Z}_j \leq \epsilon \rho_n ] \rightarrow 1$, thereby establishing that property (10a) of Lemma 1 holds w.p. one asymptotically.

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

$$Y_i := \mathbb{E}[U_i \mid X_S] = -\rho_n e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} \bar{b},$$

$$Y'_i := \text{var}[U_i \mid X_S] = \frac{\sigma^2}{n} e_i^T \left[ \frac{1}{n} X_S^T X_S \right]^{-1} e_i,$$

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

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

$$\mathbb{E}[Y_i] = -\rho_n \frac{n}{n-k-1} e_i^T (\Sigma_{SS})^{-1} \bar{b}, \quad \text{and} \quad \mathbb{E}[Y'_i] = \frac{\sigma^2}{n-k-1} e_i^T (\Sigma_{SS})^{-1} e_i,$$

respectively, which are bounded as

$$|\mathbb{E}[Y_i]| \leq \frac{2D_{\max} n \rho_n}{n-k-1}, \quad \text{and} \quad \frac{\sigma^2}{C_{\max} (n-k-1)} \leq \mathbb{E}[Y'_i] \leq \frac{\sigma^2 D_{\max}}{n-k-1}.$$

(b) Moreover, each pair $(Y_i, Y'_i)$ is concentrated, in that we have

$$P \left[ |Y_i| \geq \frac{6D_{\max} n \rho_n}{n-k-1}, \quad \text{or} \quad |Y'_i| \geq 2 \mathbb{E}[Y'_i] \right] = \mathcal{O} \left( \frac{1}{n-k} \right).$$

We exploit this lemma as follows. First define the event

$$T(\delta) := \bigcup_{i=1}^{k} \left\{ |Y_i| \geq \frac{6D_{\max} n \rho_n}{n-k-1}, \quad \text{or} \quad |Y'_i| \geq 2 \mathbb{E}[Y'_i] \right\}.$$ 

By the union bound and Lemma 6(b), we have $P[T(\delta)] = \mathcal{O} \left( \frac{k}{n-k} \right) = \mathcal{O} \left( \frac{1}{\epsilon-1} \right) \rightarrow 0$, since
\( n^{\frac{1}{k}} \rightarrow +\infty \) as \( n \rightarrow +\infty \) under the scaling \( n \geq \Omega(k \log(p - k)) \). 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 \( T(\delta) \) and its complement, we have

\[
\mathbb{P}[\max_{i \in S} |U_i| > \mathcal{M}(\beta^*)] \leq \mathbb{P}[\max_{i \in S} |U_i| > \mathcal{M}(\beta^*) \ | \ T(\delta)^c] + \mathbb{P}[T(\delta)]
\]

where each \( U_i(\mu_i^*, M_i^*) \) is Gaussian with mean \( \mu_i^* := 6D_{\max} \rho_n \frac{n}{n-k-1} \) and variance \( \nu_i := 2\mathbb{E}[Y_i'] \) respectively. In asserting the inequality (37), we have used the fact that the probability of the event \( \{U_i > \mathcal{M}(\beta^*)\} \) increases as the mean and variance of \( U_i \) increase. Continuing the argument assuming the conditioning on \( T(\delta)^c \), we have via

\[
\mathbb{P}\left[ \frac{1}{\mathcal{M}(\beta^*)} \max_{i \in S} |U_i(\mu_i^*, \nu_i)| > 1 \right] \leq \mathbb{P}\left[ \frac{1}{\mathcal{M}(\beta^*)} \max_{i \in S} |U_i(0, \nu_i)| > 1 - \frac{\max_{i \in S} |\mu_i^*|}{\mathcal{M}(\beta^*)} \right]
\]

conditioning on \( T(\delta)^c \) and Lemma 8(b), and inequality (ii) uses condition (b) on \( \rho_n \) in the theorem statement.

Next we use Markov’s inequality, the bound (35) on \( \nu_i := 2\mathbb{E}[Y_i'] \), and Lemma 9 on Gaussian maxima (see Appendix B) to obtain the upper bound

\[
\mathbb{P}\left[ \frac{1}{\mathcal{M}(\beta^*)} \max_{i \in S} |U_i(0, \nu_i)| > \frac{1}{2} \right] \leq \frac{2\mathbb{E}\left[\max_{i \in S} |U_i(0, M_i^*)|\right]}{\mathcal{M}(\beta^*)} \leq \frac{6 \sqrt{2\sigma^2 D_{\max} \log k}}{\mathcal{M}(\beta^*) \sqrt{n-k}}.
\]

Finally, using condition (ii) in equation (28), we have

\[
\frac{\log k}{\mathcal{M}(\beta^*)^2 (n-k-1)} \leq \frac{\log k}{\rho_n^2 (n-k-1)} = \mathcal{O}\left(\frac{\log(p-k)}{\rho_n^2 n}\right),
\]

which converges to zero from condition (28)(i) in the theorem statement.

### 3.4 Proof of Theorem 1(a)

We establish the claim by proving that under the stated conditions, \( \max_{j \in S^c} |V_j| > \rho_n \) with probability one, for any positive sequence \( \rho_n > 0 \). We begin by writing \( V_j = \mathbb{E}[V_j] + \tilde{V}_j \), where \( \tilde{V}_j \) is
zero-mean. Now
\[
\max_{j \in S^c} |V_j| \geq \max_{j \in S^c} |\tilde{V}_j| - \max_{j \in S^c} |\mathbb{E}[V_j]| \overset{(i)}{\geq} \max_{j \in S^c} |\tilde{V}_j| - (1 - \epsilon)\rho_n,
\]
where have used Lemma 2 in obtaining the lower bound (i). Consequently, the event \(\{\max_{j \in S^c} |\tilde{V}_j| > (2 - \epsilon)\rho_n\}\) implies the event \(\{\max_{j \in S^c} |V_j| > \rho_n\}\), so that
\[
\mathbb{P}[\max_{j \in S^c} |V_j| > \rho_n] \geq \mathbb{P}[\max_{j \in S^c} |\tilde{V}_j| > (2 - \epsilon)\rho_n] .
\]

From the preceding proof of Theorem 1(b), we know that conditioned on \(X_S\) and \(W\), the random vector \((V_1, \ldots, V_{(p - k)})\) is Gaussian with covariance of the form \(M_n [\Sigma_{SS'} - \Sigma_{SS'}(\Sigma_{SS})^{-1}\Sigma_{SS'}]\); thus, the zero-mean version \((\tilde{V}_1, \ldots, \tilde{V}_{(p - k)})\) has the same covariance. Moreover, Lemma 3 guarantees that the random scaling term \(M_n\) is concentrated. In particular, defining for any \(\delta > 0\) the event \(T(\delta) := \{ |M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n] \}\), we have \(\mathbb{P}[T(\delta)] \to 0\), and the bound
\[
\mathbb{P}[\max_{j \in S^c} |\tilde{V}_j| > (2 - \epsilon)\rho_n] \geq (1 - \mathbb{P}[T(\delta)]) \mathbb{P}\left[\max_{j \in S^c} |\tilde{V}_j| > (2 - \epsilon)\rho_n \mid T(\delta)^c\right] \overset{(ii)}{\geq} (1 - \mathbb{P}[T(\delta)]) \mathbb{P}\left[\max_{j \in S^c} |Z_j(M_n^*)| > (2 - \epsilon)\rho_n\right],
\]
where each \(Z_j \equiv Z_j(M_n^*)\) is the conditioned version of \(\tilde{V}_j\) with the scaling factor \(M_n\) fixed to \(M_n^* := (1 - \delta)\mathbb{E}[M_n]\). (In inequality (ii) above, we used the fact that Gaussian tail probabilities decrease as the variance decreases, and the fact that \(\text{var}(\tilde{V}_j) \geq M_n^*\) when conditioned on \(T(\delta)^c\).)

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

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

\(\text{(a)}\) either \(\frac{\rho_n^2}{M_n^2} \to +\infty\), and there exists some \(\gamma > 0\) such that \(\frac{1}{\rho_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq (2 - \epsilon) [1 + \gamma]\) for all sufficiently large \(n\), or
\(\text{(b)}\) there exist constants \(\alpha, \gamma > 0\) such that \(\frac{M_n^*}{\rho_n} \leq \alpha\) and \(\frac{1}{\rho_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma \sqrt{\log (p - k)}\) for all sufficiently large \(n\).

**Lemma 8.** For any \(\eta > 0\), we have
\[
\mathbb{P}\left[\max_{j \in S^c} Z_j(M_n^*) < \mathbb{E}[\max_{j \in S^c} Z_j(M_n^*)] - \eta\right] \leq \exp\left(-\frac{\eta^2}{2M_n^*}\right), \quad (38)
\]

Using these two lemmas, we complete the proof as follows. First, if condition (a) of Lemma 7
holds, then we set $\eta = \frac{(2-\epsilon)\gamma \rho_n}{2}$ in equation (38) to obtain that

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

This probability converges to 1 since $\frac{\rho_n^2}{M_n^*} \to +\infty$ from Lemma 7(a). On the other hand, if condition (b) holds, then we use the bound $\frac{1}{\rho_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma \sqrt{\log (p-k)}$ and set $\eta = \frac{\gamma \rho_n \sqrt{\log (p-k)}}{2}$ in equation (38) to obtain

$$P\left[ \frac{1}{\rho_n} \max_{j \in S^c} Z_j(M_n^*) \geq (2-\epsilon) \right] \geq P\left[ \frac{1}{\rho_n} \max_{j \in S^c} Z_j(M_n^*) \geq \frac{\gamma \sqrt{\log (p-k)}}{2} \right] \geq 1 - \exp \left( -\frac{\gamma^2 \rho_n^2 \log (p-k)}{8M_n^*} \right).$$

This probability also converges to 1 since $\frac{\rho_n^2}{M_n^*} \geq 1/\alpha$ and $\log (p-k) \to +\infty$. Thus, in either case, we have shown that $\lim_{n \to +\infty} P\left[ \frac{1}{\rho_n} \max_{j \in S^c} Z_j(M_n^*) > (2-\epsilon) \right] = 1$, 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 $k(p) = \alpha p$ for some $\alpha \in (0,1)$;

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

(c) fractional power sparsity, meaning that $k(p) = \alpha p^\gamma$ 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 $n = 2(\theta k \log(p-k)) + k + 1$. The control parameter $\theta$ 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. We specified the parameter vector $\beta^*$ by choosing the subset $S$ randomly, and for each $i \in S$ setting $\beta^*_i$ equal to $+1$ or $-1$ with equal probability, and $\beta^*_j = 0$ for all indices $j \notin S$. In addition, we fixed the noise level $\sigma = 0.5$, and the regularization parameter $\rho_n = \sqrt{\frac{\log(p-k) \log(k)}{n}}$ 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 $\theta_u = \theta_l = 1$. Figure 1 displayed earlier in Section 2 plots the control parameter $\theta$ 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 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 \( \Sigma \) are Toeplitz with the structure

\[
\Sigma = \begin{bmatrix}
1 & \mu & \mu^2 & \ldots & \mu^{p-2} & \mu^{p-1} \\
\mu & 1 & \mu & \ldots & \mu^{p-2} \\
\mu^2 & \mu & 1 & \ldots & \mu^{p-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mu^{p-1} & \ldots & \mu^3 & \mu^2 & \mu & 1
\end{bmatrix},
\]

for some \( \mu \in (-1, +1) \). Moreover, the maximum and minimum eigenvalues \( (C_{\text{min}} \text{ and } C_{\text{max}}) \) can be computed using standard asymptotic results on Toeplitz matrix families \([19]\). Figure 2 shows representative results for this Toeplitz family with \( \mu = 0.10 \). Panel (a) corresponds to linear sparsity \( k = \alpha p \) with \( \alpha = 0.40 \), panel (b) corresponds to sublinear sparsity \( k = \alpha p / \log(\alpha p) \) with \( \alpha = 0.40 \), whereas panel (c) corresponds to fractional sparsity \( k = \alpha p^{0.75} \). 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.

**Figure 2.** Plots of the number of data samples \( n = 2 \theta k \log(p - k) \), indexed by the control parameter \( \theta \), versus the probability of success in the Lasso for the Toeplitz family \([39]\) with \( \mu = 0.10 \). 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: \( k(p) = \alpha p \). (b) Sublinear sparsity index \( k(p) = \alpha p / \log(\alpha p) \). (c) Fractional power sparsity index \( k(p) = \alpha p^{\gamma} \) with \( \gamma = 0.75 \).
5 Discussion

The problem of recovering the sparsity pattern of a high-dimensional vector $\beta^*$ from noisy observations has important applications in signal denoising, compressed sensing, graphical model selection, sparse approximation, and subset selection. This paper focuses on the behavior of $\ell_1$-regularized quadratic programming, also known as the Lasso, for estimating such sparsity patterns in the noisy and high-dimensional setting. We first analyzed the case of deterministic designs, and provided sufficient conditions for exact sparsity recovery using the Lasso that allow for general scaling of the number of observations $n$ in terms of the model dimension $p$ and sparsity index $k$. We then turned to the case of random designs, with measurement vectors drawn randomly from certain Gaussian ensembles. The main contribution in this setting was to establish a threshold of the order $n = \Theta(k \log(p-k))$ governing the behavior of the Lasso: in particular, the Lasso succeeds with probability (converging to) one above threshold, and conversely, it fails with probability one below threshold. For the uniform Gaussian ensemble, our threshold result is exactly pinned down to $n = 2k \log(p-k)$ with matching lower and upper bounds, whereas for more general Gaussian ensembles, it should be possible to tighten the constants in our analysis.

There are a number of interesting questions and open directions associated with the work described here. Although the current work focused exclusively on linear regression, it is clear that the ideas and analysis techniques apply to other log-linear models. Indeed, some of our follow-up work [34] has established qualitatively similar results for the case of logistic regression, with application to model selection in binary Markov random fields. Another interesting direction concerns the gap between the performance of the Lasso, and the performance of the optimal (oracle) method for selecting subsets. In this realm, information-theoretic analysis [33] shows that it is possible to recover linear-sized sparsity patterns ($k = \alpha p$) using only a linear fraction of observations ($n = \Theta(p)$). This type of scaling contrasts sharply with the order of the threshold $n = \Theta(k \log(p-k))$ that this paper has established for the Lasso. It remains to determine if a computationally efficient method can achieve or approach the information-theoretic limits in this regime of the triplet $(n, p, k)$.

Acknowledgements

We would like to thank Peter Bickel, Alyson Fletcher, Noureddine El Karoui, Vivek Goyal, John Lafferty, Larry Wasserman and Bin Yu for helpful comments and pointers. This work was partially supported by an Alfred P. Sloan Foundation Fellowship and and NSF Grant DMS-0605165.

A Proof of Lemma [1]

By standard conditions for optimality in a convex program [20], a point $\hat{\beta} \in \mathbb{R}^p$ is optimal for the regularized form of the Lasso (3) if and only if there exists a subgradient $\hat{z} \in \partial \ell_1(\hat{\beta})$ such that
\( \frac{1}{n} X^T X \hat{\beta} - \frac{1}{n} X^T y + \rho \hat{z} = 0 \). Here the subdifferential of the \( \ell_1 \) norm takes the form
\[
\partial \ell_1(\hat{\beta}) = \left\{ \hat{z} \in \mathbb{R}^p \mid \hat{z}_i = \text{sgn}(\hat{\beta}_i) \text{ for } \hat{\beta}_i \neq 0, \quad |\hat{z}_j| \leq 1 \text{ otherwise} \right\}.
\] (40)

Substituting our observation model \( y = X \beta^* + w \) and re-arranging yields
\[
\frac{1}{n} X^T X (\hat{\beta} - \beta^*) - \frac{1}{n} X^T w + \rho \hat{z} = 0.
\] (41)

Now condition \( R(X, \beta^*, w, \rho) \) holds if and only if \( \hat{\beta} \) satisfies \( \hat{\beta}_S = 0 \) and \( \hat{\beta}_S \neq 0 \), and the subgradient satisfies \( \hat{z}_S = \text{sgn}(\beta^*_S) \) and \( |\hat{z}_S| \leq 1 \). From these conditions and using equation (41), we conclude that the condition \( R(X, \beta^*, w, \rho) \) holds if and only if
\[
\frac{1}{n} X^T S X_S (\hat{\beta}_S - \beta^*_S) - \frac{1}{n} X^T S w = -\rho \hat{z}_S.
\]
\[
\frac{1}{n} X^T S X_S (\hat{\beta}_S - \beta^*_S) - \frac{1}{n} X^T S w = -\rho \text{sgn}(\beta^*_S).
\]

Using the invertibility of \( X^T S X_S \), we may solve for \( \hat{\beta}_S \) and \( \hat{z}_S \) to conclude that
\[
\rho \hat{z}_S = X^T S X_S \left( X_S^T X_S \right)^{-1} \left[ \frac{1}{n} X^T S w - \rho \text{sgn}(\beta^*_S) \right] - \frac{1}{n} X^T S w
\]
\[
\hat{\beta}_S = \beta^*_S + \left( \frac{1}{n} X^T S X_S \right)^{-1} \left[ \frac{1}{n} X^T S w - \rho \text{sgn}(\beta^*_S) \right].
\]

From these relations, the requirements \(|\hat{z}_S| \leq 1 \) and \( \text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta^*_S) \) yield conditions (10a) and (10b) respectively. Lastly, note that if the vector inequality \(|z_S| < 1 \) holds strictly, then \( \hat{\beta}_S = 0 \) for all solutions to the Lasso as claimed.

B Some Gaussian comparison results

We state here (without proof) some well-known comparison results on Gaussian maxima [24]. We begin with a crude but useful bound:

**Lemma 9.** For any Gaussian RV \( X_1, \ldots, X_n \), \( E[ \max_{1 \leq i \leq n} |X_i| ] \leq 3 \sqrt{\log n} \max_{1 \leq i \leq n} \sqrt{E X_i^2} \).

Next we state (a version of) the Sudakov-Fernique inequality [24]:

**Lemma 10** (Sudakov-Fernique). Let \( X = (X_1, \ldots, X_n) \) and \( Y = (Y_1, \ldots, Y_n) \) be Gaussian random vectors such that for all \( i, j \) \( E[(Y_i - Y_j)^2] \leq E[(X_i - X_j)^2] \). Then \( E[ \max_{1 \leq i \leq n} Y_i ] \leq E[ \max_{1 \leq i \leq n} X_i ] \).

C Auxiliary lemma

For future use, we state formally the following elementary
Lemma 11. Given a collection \( \{Z_1, Z_2, \ldots, Z_{(p-k)}\} \) of random variables with distribution symmetric around zero, for any constant \( a > 0 \) we have

\[
\mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} |Z_j| \leq a \right] \leq \mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} Z_j \leq a \right], \quad \text{and} \quad (44a)
\]
\[
\mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} |Z_j| > a \right] \leq 2\mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} Z_j > a \right]. \quad (44b)
\]

Proof. The first inequality is trivial. To establish the inequality (44b), we write

\[
\mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} |Z_j| > a \right] = \mathbb{P}\left[ \left( \max_{1 \leq j \leq (p-k)} Z_j > a \right) \text{ or } \left( \min_{1 \leq j \leq (p-k)} Z_j < -a \right) \right]
\]
\[
\leq \mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} Z_j > a \right] + \mathbb{P}\left[ \min_{1 \leq j \leq (p-k)} Z_j < -a \right]
\]
\[
= 2\mathbb{P}\left[ \max_{1 \leq j \leq (p-k)} Z_j > a \right],
\]

where we have used the union bound, and the symmetry of the events \( \max_{1 \leq j \leq (p-k)} Z_j > a \) and \( \min_{1 \leq j \leq (p-k)} Z_j < -a \).

D Toepplitz covariance matrices

In this appendix, we verify that Toepplitz covariance matrices satisfy the conditions of Theorem 1. Standard results on Toepplitz matrices \[19\] show that the eigenvalues are suitably bounded. Previous work \[35\] has shown that Toepplitz families satisfy the mutual incoherence condition (25). It remains to verify the bound \( \| (\Sigma_{SS})^{-1} \|_\infty \leq D_{\text{max}} \). From the block matrix inversion formula \[21\], we have

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

Applying the triangle inequality yields

\[
\| (\Sigma_{SS})^{-1} \|_\infty \leq \| (\Sigma^{-1})_{SS} \|_\infty + \| (\Sigma^{-1})_{SSc} \|_\infty \| \left( (\Sigma^{-1})_{S^cS^c} \right)^{-1} \|_\infty \| (\Sigma^{-1})_{S^cS} \|_\infty \quad (45)
\]

For a Toepplitz matrix \( \Sigma = \text{toep}[1 \ \mu \ \ldots \ \mu^{p-1}] \), the inverse is tridiagonal with bounded diagonal entries \( a_i(\mu) \), immediate off-diagonals \( b_{ij}(\mu) \), and all other entries equal to zero. Hence, it follows immediately that the matrix norms \( \| (\Sigma^{-1})_{SS} \|_\infty \), \( \| (\Sigma^{-1})_{SSc} \|_\infty \), and \( \| (\Sigma^{-1})_{S^cS} \|_\infty \) are all bounded (independently of \( k \) and \( p \)). Finally, the matrix \( (\Sigma^{-1})_{S^cS^c} \) is blockwise tridiagonal, with each block corresponding to a subset of indices \( T \subseteq S^c \) all connected by single hops. Hence, the inverse matrix \( \left[ (\Sigma^{-1})_{S^cS^c} \right]^{-1} \) is a blockwise Toepplitz matrix. Each block can be interpreted as the covariance matrix of a stable autoregressive process, so that the \( \ell_\infty \) norm of each row is bounded independently of \( p \) and the choice of \( S^c \). Thus, we conclude that \( \| \left[ (\Sigma^{-1})_{S^cS^c} \right]^{-1} \|_\infty \) is upper bounded independently of \( p \) and \( S^c \), and hence via equation (45) that the same holds for \( \| (\Sigma_{SS})^{-1} \|_\infty \) as claimed.
E  Lemma for Theorem 1

E.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 (i.e., for estimating $X_{Sc}$ on the basis of $X_S$), the random variable $(X_{Sc} \mid X_S, W) \sim (X_{Sc} \mid X_S)$ is Gaussian with mean and covariance

\[
\mathbb{E}[X_{Sc}^T \mid X_S, W] = \Sigma_{ScS}(\Sigma_{SS})^{-1}X_S^T, \\
\text{var}(X_{Sc} \mid X_S) = \Sigma_{(ScS)} = \Sigma_{ScS} - \Sigma_{ScS}(\Sigma_{SS})^{-1}\Sigma_{SSc}.
\]

Consequently, we have

\[
\mathbb{E}[V_j \mid X_S, W] = \begin{bmatrix} \Sigma_{ScS}(\Sigma_{SS})^{-1}X_S^T \end{bmatrix} X_S (X_S^TX_S)^{-1} \rho_n \bar{b} - [X_S(X_S^TX_S)^{-1}X_S^T - I_{n \times n}] \frac{W}{n}
\]

as claimed.

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

\[
\text{cov}(V_j, V_k \mid X_S, W) = \begin{bmatrix} \Sigma_{ScS}(\Sigma_{SS})^{-1}\rho_n \bar{b} \end{bmatrix} \leq \rho_n(1 - \epsilon)1,
\]

as claimed.

\[
\text{cov}(X_{ji}, X_{ki} \mid X_S, W) = \begin{bmatrix} \rho_n^2X_S^TX_S^{-1} \bar{b} + \frac{1}{n^2}W^T[I_{n \times n} - X_S(X_S^TX_S)^{-1}X_S^TW] \end{bmatrix}.
\]

E.2  Proof of Lemma 3

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

\[
\mathbb{E} \left[ \rho_n^2 \bar{b}^T (X_S^TX_S)^{-1} \bar{b} \right] = \frac{\rho_n^2}{n-s-1} \bar{b}^T (\Sigma_{SS})^{-1} \bar{b}.
\]

Now define the random matrix $R = I_{n \times n} - X_S(X_S^TX_S)^{-1}X_S^T$. 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 $z = X_Su$ in the range of $X_S$, we have $Rz = [I_{n \times n} - X_S(X_S^TX_S)^{-1}X_S^T]X_Su = 0$. Hence dim(ker $R$) = dim(range $X_S$) = $s$. Since $R$ is symmetric and positive semidefinite, there exists an orthogonal matrix $U$ such that $R = U^TDU$, 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

\[
\frac{1}{n^2} \mathbb{E} \left[ W^T R W \mid X_S \right] = \frac{1}{n^2} \mathbb{E} \left[ W^T U^T D U W \mid X_S \right] = \frac{1}{n^2} \text{trace} \left( D U U^T \mathbb{E} \left[ W W^T \mid X_S \right] \right) = \sigma^2 \frac{n-s}{n^2},
\]

25
since $\mathbb{E}[WW^T] = \sigma^2 I$. Consequently, we have established that $\mathbb{E}[M_n] = \frac{\rho_n^2}{n-k-1} \overline{b}^T (\Sigma_{SS})^{-1} \overline{b} + \frac{\sigma^2 (n-k)}{n^2}$ as claimed.

We now compute the expected value of the squared variance

$$M_n^2 = \frac{\rho_n^4}{n^2} \left[ \overline{b}^T (X_S^T X_S)^{-1} \overline{b} \right]^2 + \frac{2 \rho_n^2}{n^2} \left[ \overline{b}^T (X_S^T X_S)^{-1} \overline{b} \right] (W^T RW) + \frac{1}{n^4} (W^T RW)^2$$

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

$$\mathbb{E}[T_3|X_S] = \frac{1}{n^4} \mathbb{E}[(W^T DW)^2] = \frac{1}{n^4} \mathbb{E} \left( \sum_{i=1}^{n-k} W_i \right)^2$$

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

Similarly, using conditional expectation and our previous calculation (48) of $\mathbb{E}[W^T RW \mid X_S]$, we have

$$\mathbb{E}[T_2] = \frac{2 \rho_n^2}{n^2} \mathbb{E} \left[ \overline{b}^T (X_S^T X_S)^{-1} \overline{b} \right],$$

whence $\mathbb{E}[T_2] = \frac{2(n-k)\sigma^4}{n^4} + \frac{(n-k)^2 \sigma^4}{n^4}$ as well.

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 [29]) to write, for all $n > k + 3$,

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

Consequently, combining our results, we have

$$\text{var}(M_n) = \mathbb{E}[M_n^2] - (\mathbb{E}[M_n])^2$$

$$= \sum_{i=1}^{3} \mathbb{E}[T_i] - \left\{ \frac{\sigma^4 (n-k)^2}{n^4} + \frac{2 \sigma^2 (n-k)}{n^2} \frac{\rho_n^2}{n-k-1} \overline{b}^T (\Sigma_{SS})^{-1} \overline{b} + \left( \frac{\rho_n^2}{n-k-1} \overline{b}^T (\Sigma_{SS})^{-1} \overline{b} \right)^2 \right\}$$

$$= \frac{2(n-k) \sigma^4}{n^4} + \frac{\rho_n^4}{(n-k-1)(n-k-3)} \left\{ \frac{1}{(n-k)} + \frac{n-k-1}{(n-k)} - \frac{(n-k-3)}{(n-k-1)} \right\}.$$
Finally, we establish the concentration result. Using Chebyshev’s inequality, we have
\[ P[|M_n - E[M_n]| \geq \delta E[M_n]] \leq \frac{\text{var}(M_n)}{\delta^2(E[M_n])^2}, \]
so that it suffices to prove that \( \text{var}(M_n)/(E[M_n])^2 \to 0 \) as \( n \to +\infty \). We deal with each of the two variance terms \( H_1 \) and \( H_2 \) in equation (51) separately. First, we have
\[ H_1 \left( \frac{E[M_n]^2}{(E[M_n])^2} \right) \leq \frac{2(n-k)\sigma^4}{n^4} = \frac{2}{n-k} \to 0. \]
Secondly, denoting \( A = (\tilde{b}^T(X_S^T X_S)^{-1} \tilde{b}) \) for short-hand, we have
\[ H_2 \left( \frac{E[M_n]^2}{(E[M_n])^2} \right) \leq \frac{(n-k-1)^2}{(n-k-3)} \frac{\rho_n^4 A^2}{(n-k-1)(n-k-3)} \left\{ \frac{1}{(n-k)} + \frac{n-k-1}{(n-k)} - \frac{n-k-3}{(n-k-1)} \right\}, \]
which also converges to 0 as \( (n-k) \to 0 \).

**E.3 Proof of Lemma 4**

Recall that the Gaussian random vector \((Z_1, \ldots, Z_{(p-k)})\) is zero-mean with covariance \( M_n^* \Sigma_{(S^c|S)} \), where \( \Sigma_{(S^c|S)} := \Sigma_{S^cS^c} - \Sigma_{S^cS} \Sigma_{SS}^{-1} \Sigma_{SS^c} \). For any index \( i \), let \( e_i \in \mathbb{R}^{(p-k)} \) be equal to 1 in position \( i \), and zero otherwise. For any two indices \( i \neq j \), we have
\[ E[(Z_i - Z_j)^2] = M_n^* (e_i - e_j)^T \Sigma_{(S^c|S)} (e_i - e_j) \leq 2M_n^* \lambda_{\text{max}}(\Sigma_{(S^c|S)}) \leq 2C_{\text{max}} M_n^*, \]
since \( \Sigma_{(S^c|S)} \preceq \Sigma_{S^cS^c} \) by definition, and \( \lambda_{\text{max}}(\Sigma_{S^cS^c}) \leq \lambda_{\text{max}}(\Sigma) \leq C_{\text{max}} \).

Letting \((X_1, \ldots, X_{(p-k)}) \sim N(0, C_{\text{max}} M_n^* I_{(p-k) \times (p-k)})\), we have \( E[(X_i - X_j)^2] = 2C_{\text{max}} M_n^* \).

Hence, applying the Sudakov-Fernique inequality (see Lemma 10) yields \( E[\max_j Z_j] \leq E[\max_j X_j] \).

From standard results on asymptotic behavior of Gaussian maxima [18], we have
\[ \lim_{(p-k) \to \infty} \frac{E[\max_j X_j]}{\sqrt{2C_{\text{max}} M_n^* \log (p-k)}} = 27 \]
1. Consequently, for all \( \delta' > 0 \), there exists an \( N(\delta') \) such that for all \( (p-k) \geq N(\delta') \), we have

\[
\frac{1}{\rho_n} E[\max_j Z_j(M_n^*)] \leq \frac{1}{\rho_n} E[\max_j X_j] \\
\leq (1 + \delta') \sqrt{\frac{2C_{\max} M_n^* \log(p-k)}{\rho_n^2}} \\
= (1 + \delta') \sqrt{1 + \epsilon} \sqrt{\frac{2C_{\max} \log(p-k)}{n-k-1}} \frac{b^T \Sigma_{SS}^{-1} b + 2C_{\max} \sigma^2 (1 - \frac{k}{n}) \log(p-k)}{n \rho_n^2}.
\]

Now, using our assumption that \( n > 2(\theta_u - \nu)k \log(p-k) - k - 1 \) for some \( \nu > 0 \), where \( \theta_u = \frac{C_{\max}}{\epsilon^2 \log \sigma} \), we have

\[
\frac{1}{\rho_n} E[\max_j Z_j(M_n^*)] < (1 + \delta') \sqrt{1 + \epsilon} \sqrt{\frac{2C_{\max} \log(p-k) - k - 1}{n-k-1}} \frac{\log(p-k)}{n \rho_n^2}.
\]

Recall that by assumption, as \( n, (p-k) \rightarrow +\infty \), we have that \( \frac{\log(p-k)}{n \rho_n^2} \) and \( \frac{\log(p-k)}{n-k-1} \) converge to zero. Consequently, the RHS converges to \( (1 + \delta') \sqrt{(1 + \epsilon) \epsilon} \) as \( n, (p-k) \rightarrow \infty \). Hence, we have \( \lim_{n \rightarrow +\infty} \frac{1}{\rho_n} E[\max_j Z_j(M_n^*)] < (1 + \delta') \sqrt{1 + \epsilon} \). Since \( \delta' > 0 \) and \( \epsilon > 0 \) were arbitrary, the result follows.

### E.4 Proof of Lemma 5

Consider the function \( f : \mathbb{R}^{(p-k)} \rightarrow \mathbb{R} \) given by \( f(w) := \sqrt{M_n^* \max_{j \in S^c} \left[ \sqrt{\Sigma_{(S^c \cup S)}} w \right]} \), where \( \Sigma_{(S^c \cup S)} := \Sigma_{S^c} - \Sigma_{S^c \cup S} \Sigma_{SS}^{-1} \Sigma_{SS^c} \). By construction, for a Gaussian random vector \( V \sim N(0, I) \), we have \( f(V) \overset{d}{=} \max_{j \in S^c} \tilde{Z}_j \).

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

\[
|\sqrt{M_n^* R w}_j - \sqrt{M_n^* R v}_j| \leq \sqrt{M_n^*} \left| \sum_k R_{jk} [w_k - v_k] \right| \\
\leq \sqrt{M_n^*} \\sqrt{\sum_k R_{jk}^2 \|w - v\|_2} \\
\leq \sqrt{M_n^*} \|w - v\|_2,
\]

where the last inequality follows since \( \sum_k R_{jk}^2 = (\Sigma_{(S^c \cup S)})_{jj} \leq 1 \). Therefore, by Gaussian concentra-
tion of measure for Lipschitz functions [23], we conclude that for any $\eta > 0$, it holds that

$$\mathbb{P}(f(W) \geq \mathbb{E}[f(W)] + \eta) \leq \exp\left(-\frac{\eta^2}{2M_n^2}\right), \quad \text{and} \quad \mathbb{P}(f(W) \leq \mathbb{E}[f(W)] - \eta) \leq \exp\left(-\frac{\eta^2}{2M_n^2}\right).$$

E.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 $\mathbb{E}[(X_S^T X_S)^{-1}] = \left(\frac{(SS)^{-1}}{n-k-1}\right)$ (see Lemma 7.7.1, [1]). Thus, we compute

$$\mathbb{E}[Y_i] = -\rho_n \frac{n}{n-k-1} e_i^T (\Sigma SS)^{-1} \vec{b}, \quad \text{and}$$

$$\mathbb{E}[Y_i'] = \frac{\sigma^2}{n} \frac{n}{n-k-1} e_i^T (\Sigma SS)^{-1} e_i = \frac{\sigma^2}{n-k-1} e_i^T (\Sigma SS)^{-1} e_i.$$

Moreover, using formulae for second moments of inverse Wishart matrices (see, e.g., [29]), we compute

$$\mathbb{E}[Y_i^2] = \frac{\rho_n^2 n^2}{(n-k)(n-k-3)} \left[ e_i^T (\Sigma SS)^{-1} e_i \right]^2 + \frac{1}{n-k-1} \left( e_i^T (\Sigma SS)^{-1} \vec{b} \right) \left( e_i^T (\Sigma SS)^{-1} e_i \right)$$

$$\mathbb{E}[(Y_i')^2] = \frac{\sigma^4 n^2}{(n-k-1)^2 (n-k)(n-k-3)} \left( e_i^T (\Sigma SS)^{-1} e_i \right)^2 \left[ 1 + \frac{1}{n-k-1} \right].$$

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

$$\text{var}(Y_i) = \frac{\rho_n^2 n^2}{(n-k)(n-k-3)} \left[ A_i^2 + \frac{1}{n-k-1} A_i B_i \right] - \frac{\rho_n^2 n^2}{(n-k-1)^2} A_i^2$$

$$= \frac{\rho_n^2 n^2}{(n-k)(n-k-3)} \left[ A_i^2 \left( 1 - \frac{(n-k)(n-k-3)}{(n-k-1)^2} \right) + \frac{1}{n-k-1} A_i B_i \right]$$

$$\leq 2 \rho_n^2 \left[ \frac{3 A_i^2}{n-k} + \frac{A_i B_i}{n-k-1} \right]$$

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

$$\text{var}(Y_i) \leq \frac{K \rho_n^2}{n-k}$$

for some fixed constant $K$ independent of $k$ and $n$.

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

$$|Y_i - \mathbb{E}[Y_i]| \geq |Y_i| - |\mathbb{E}[Y_i]| \geq |Y_i| - \frac{2D_{\max} \rho_n n}{n-k-1}.$$
Consequently, making use of Chebyshev’s inequality, we have

\[
P[Y_i \geq \frac{6D_{\text{max}}\rho n}{n - k - 1}] = 2P\left[Y_i - \frac{2D_{\text{max}}\rho n}{n - k - 1} \geq \frac{4D_{\text{max}}\rho n}{n - k - 1}\right] \\
\leq 2P\left[Y_i - \mathbb{E}[Y_i] \geq \frac{4D_{\text{max}}\rho n}{n - k - 1}\right] \\
\leq \frac{\text{var}(Y_i)}{16D_{\text{max}}^2\rho^2_n} \\
\leq \frac{K}{16D_{\text{max}}(n - k)}
\]

where the final step uses the bound (54). Next we compute and bound the variance of \(Y_i'\). We have

\[
\text{var}(Y_i') = \frac{\sigma^4 n^2}{(n - k - 1)^2 (n - k) (n - k - 3)} \left( A_i^2 \left[ 1 + \frac{1}{n - k - 1} \right] - \frac{\sigma^4}{(n - k - 1)^2} A_i^2 \right) \\
\leq \frac{K\sigma^4}{(n - k - 1)^3}
\]

for some constant \(K\) independent of \(k\) and \(n\). Consequently, applying Chebyshev’s inequality, we have

\[
P[Y_i' \geq 2\mathbb{E}[Y_i']] = P[Y_i' - \mathbb{E}[Y_i'] \geq \mathbb{E}[Y_i']] \leq \frac{\text{var}(Y_i')}{(\mathbb{E}[Y_i'])^2} \\
\leq \frac{K}{(n - k - 1)^3} \frac{1}{\frac{\sigma^4}{n^2} \mathbf{e}^T (\Sigma_{SS})^{-1} \mathbf{e}} \\
\leq \frac{Kn^2C_{\text{max}}}{\sigma^4(n - k - 1)^3} \\
\leq \frac{K'}{n - k - 1}
\]

for some constant \(K'\) independent of \(k\) and \(n\).

**E.6 Proof of Lemma 7**

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

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

Now let \((X_1, \ldots, X_{p-k})\) be an i.i.d. zero-mean Gaussian vector with \(\text{var}(X_i) = C_{\text{max}}M_n^*\), so that

\[
\Delta_X(i, j) := \mathbb{E}[(X_i - X_j)^2] = 2C_{\text{max}}M_n^*.
\]

If we set

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

30
then, by applying a known error bound for the Sudakov-Fernique inequality \[6\], we are guaranteed that

\[ E[\max_{j \in S^c} Z_j] \geq E[\max_{j \in S^c} X_j] - \sqrt{\Delta^* \log (p - k)}. \] (55)

We now show that the quantity \( \Delta^* \) is upper bounded as \( \Delta^* \leq 2M_n^* \left( C_{\max} - \frac{1}{C_{\max}} \right) \). Using the inversion formula for block-partitioned matrices \[21\], 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

\[ E[(Z_i - Z_j)^2] = M_n^*(e_i - e_j)^T \Sigma_{(S^c|S)}(e_i - e_j) \geq 2M_n^* \lambda_{\min}(\Sigma_{(S^c|S)}) \geq 2M_n^* \lambda_{\min}(\Sigma^{-1}) = \frac{2M_n^*}{C_{\max}}. \]

In turn, this leads to the upper bound

\[ \Delta^* = \max_{i,j \in S^c} |\Delta_X(i,j) - \Delta_Z(i,j)| = \max_{i,j \in S^c} [2M_n^* C_{\max} - \Delta_Z(i,j)] \leq 2M_n^* \left( C_{\max} - \frac{1}{C_{\max}} \right). \]

We now analyze the behavior of \( E[\max_{j \in S^c} X_j] \). Using asymptotic results on the extrema of i.i.d. Gaussian sequences \[18\], we have \( \lim_{(p - k) \to +\infty} \frac{E[\max_{j \in S^c} X_j]}{\sqrt{2C_{\max}M_n^* \log (p - k)}} = 1 \). Consequently, for all \( \delta' > 0 \), there exists an \((p - k)(\delta')\) such that for all \((p - k) \geq (p - k)(\delta')\), we have

\[ E[\max_{j \in S^c} X_j] \geq (1 - \delta') \sqrt{2C_{\max}M_n^* \log (p - k)}. \]

Applying this lower bound to the bound \((55)\), we have

\[
\frac{1}{\rho_n} E[\max_{j \in S^c} Z_j] \geq \frac{1}{\rho_n} \left[ (1 - \delta') \sqrt{2C_{\max}M_n^* \log (p - k)} - \sqrt{\Delta^* \log (p - k)} \right] \\
\geq \frac{1}{\rho_n} \left[ (1 - \delta') \sqrt{2C_{\max}M_n^* \log (p - k)} - \sqrt{2 M_n^* \left( C_{\max} - \frac{1}{C_{\max}} \right) \log (p - k)} \right] \\
= \left[ (1 - \delta') \sqrt{C_{\max}} - \sqrt{C_{\max} - \frac{1}{C_{\max}}} \right] \sqrt{2 M_n^* \rho_n^2 \log (p - k)}. \] (56)
First, assume that \( \rho_n^2 / M_n^* \) does not diverge to infinity. Then, there exists some \( \alpha > 0 \) such that 
\[
\frac{\rho_n^2}{M_n^* n} \leq \alpha
\]
for all sufficiently large \( n \). In this case, we have from the bound (56) that
\[
\frac{1}{\rho_n^2} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma \sqrt{\log (p - k)}
\]
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 \( \rho_n^2 / M_n^* \to +\infty \). We compute
\[
\frac{1}{\rho_n^2} \sqrt{2M_n^* \log (p - k)} = \sqrt{1 - \delta} \sqrt{\frac{2 \log (p - k) - b^T (\Sigma_{SS})^{-1} b}{n - k - 1}} + \frac{2\sigma^2 (1 - \frac{\varepsilon}{n}) \log (p - k)}{n\rho_n^2}
\]
\[
\geq \sqrt{1 - \delta} \sqrt{\frac{2 \log (p - k)}{n - k - 1}}
\]
\[
\geq \frac{\sqrt{1 - \delta}}{C_{\max}} \sqrt{\frac{2k \log (p - k)}{n - k - 1}}.
\]
We now apply the condition
\[
\frac{2k \log (p - k)}{n - k - 1} > \frac{1}{\theta_k - \nu} = C_{\max} (2 - \epsilon)^2 / \left[ \sqrt{C_{\max}} - \sqrt{C_{\max} - \frac{1}{C_{\max}}} \right]^2 - \nu C_{\max} (2 - \epsilon)^2
\]
to obtain that
\[
\frac{1}{\rho_n^2} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \sqrt{(1 - \delta)} \frac{(1 - \delta') \sqrt{C_{\max}} - \sqrt{C_{\max} - \frac{1}{C_{\max}}}}{\sqrt{\left[ \sqrt{C_{\max}} - \sqrt{C_{\max} - \frac{1}{C_{\max}}} \right]^2 - \nu C_{\max} (2 - \epsilon)^2}} (2 - \epsilon) \tag{57}
\]
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 (57). 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 \( \frac{1}{\rho_n^2} \mathbb{E}[\max_{j \in S^c} Z_j] \geq (2 - \epsilon) (1 + \gamma) \) for all sufficiently large \( n \).

### E.7 Proof of Lemma 8

This claim follows from the proof of Lemma 5.
References


