Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programmming (Lasso)

Martin J. Wainwright[1]

***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 ℓ_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 β^*, and the number of observations n that are necessary and sufficient for sparsity pattern recovery using the Lasso. We first analyze the case of observations made using deterministic design matrices and sub-Gaussian additive noise, and provide sufficient conditions for support recovery and ℓ_∞-error bounds, as well as results showing the necessity of incoherence and bounds on the minimum value. We then turn to the case of random designs, in which each row of the design is drawn from a $N(0, \Sigma)$ ensemble. For a broad class of Gaussian ensembles satisfying mutual incoherence conditions, we compute explicit values of thresholds $0 < \theta_\ell(\Sigma) \leq \theta_u(\Sigma) < +\infty$ 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 ($\Sigma = I_{p \times p}$), we show that $\theta_\ell = \theta_u = 1$, so that the precise threshold $n = 2\,k \log(p-k)$ is exactly determined.**

Keywords: Compressed sensing; Convex relaxation; High-dimensional inference; ℓ_1-constraints; Model selection; Phase transitions; Sparse approximation; Signal denoising; Subset selection. [1]

I. Introduction

The area of high-dimensional statistical inference is concerned with the behavior of models and algorithms in which the dimension p is comparable to, or possibly even larger than the sample size n. In the absence of additional structure, it is well-known that many standard procedures—among them linear regression and principal component analysis—are not consistent unless the ratio p/n converges to zero. Since this scaling precludes having p comparable or larger than n, an active line of research is based on imposing structural conditions on the data—for instance, sparsity, manifold constraints, or graphical model structure—and then studying conditions under which various polynomial-time methods are either consistent, or conversely inconsistent.

In this paper, we study the following problem of high-dimensional inference with sparsity constraints: given noisy linear observations of an unknown vector β^*, how to recover the positions of its non-zero entries, otherwise known as its *sparsity pattern* or *support set*? This problem, known variously as sparsity recovery, support recovery, or variable selection, arises in a broad variety of contexts, including subset selection in regression [28], compressed sensing [9], [4], structure estimation in graphical models [27], sparse approximation [8], 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 [29], 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-dimensional observation vector $y \in \mathbb{R}^n$ of the form

$$y \quad = \quad X\beta^* + w, \tag{1}$$

where $w \in \mathbb{R}^n$ is zero-mean additive observation noise, and $X \in \mathbb{R}^{n \times p}$ is the measurement or design matrix. In many settings, it is natural to assume that the vector β^* is sparse, in that the cardinality $k = |S(\beta^*)|$ of its *support*

$$S(\beta^*) := \{i \in \{1, \ldots p\} \;\mid\; \beta^*_i \neq 0\} \;\; \text{satisfies } k \ll p. \tag{2}$$

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), known as the Lasso in the statistics literature [30], given by

$$\min_{\beta \in \mathbb{R}^p} \big\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \big\}, \tag{3}$$

where $\lambda_n > 0$ is a regularization parameter. Equivalently, the convex program (3) can be reformulated as the ℓ_1-constrained quadratic program [6]

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2, \qquad \text{such that} \qquad \|\beta\|_1 \leq C_n, \tag{4}$$

[1] Department of Statistics, and Department of Electrical Engineering and Computer Sciences, UC Berkeley.

[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. The work of MJW supported in part by NSF grants DMS-0528488, CAREER-CCF-0545862, a Microsoft Research Grant, and a Sloan Foundation Fellowship.

where the regularization parameter λ_n and constraint level C_n are in one-to-one correspondence via Lagrangian duality. In this paper, we focus on the following question: what are necessary and sufficient 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 $S(\beta^*)$ using the Lasso?

A. *Overview of previous work*

Recent years have witnessed a great deal of work on the use of ℓ_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 β^* by solving

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \qquad \text{subject to} \quad X\beta = y. \tag{5}$$

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. [6]. For the noiseless setting, the interesting regime is the underdetermined setting (i.e., $n < p$). With contributions from a broad range of researchers [3], [6], [15], [17], [26], [31], there is now a fairly complete understanding of the conditions on the measurement matrices X and sparsity indices k that ensure that the true solution β^* 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 results by Donoho [10], 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., $X_{ij} \sim N(0,1)$, i.i.d.) with the ambient dimension p scaling linearly in terms of the number of observations (i.e., $p = \delta n$, for some $\delta \in (0,1)$), there exists a constant $\alpha \in (0,1)$ 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 [13], who show that the basis pursuit LP (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 ($X_{ij} \sim N(0,1)$, i.i.d.) as a special case, we show that the Lasso quadratic program (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 ℓ_1-constrained quadratic program (3), known as the Lasso in the statistics literature [30], [14], has been the focus of considerable research in recent years. Knight and Fu [22] analyze the asymptotic behavior of the optimal solution, not only for ℓ_1 regularization but for ℓ_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:

- some measurement of predictive power, such as the mean-squared error $\mathbb{E}[\|Y_i - \widehat{Y}_i\|_2^2]$, where $\widehat{Y}_i$ is the estimate based on $\widehat{\beta}$; and
- various ℓ_q norms $\mathbb{E}\|\widehat{\beta} - \beta^*\|_q^q$, especially ℓ_2 and ℓ_1;
- the subset or variable 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 [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 [32] and Donoho et al. [12] 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^*. We discuss connections to this body of work at more length in Section III. Another line of work has analyzed the use of the Lasso [3], [11], as well as other closely related convex relaxations [5] when applied to random ensembles with measurement vectors drawn from the standard Gaussian ensemble. These papers either provide conditions under which estimation of a noise-contaminated signal via the Lasso is stable in the ℓ_2 sense [3], [11], 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. 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 [27] and deterministic designs by Zhao and Yu [37]. Both papers established that when suitable mutual incoherence conditions are imposed on either random [27] or deterministic design matrices [37], 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. We discuss connections to this body of work at more length in Section IV-A.

B. *Our contributions*

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). Consequently, the analysis applies to different sparsity regimes, including *linear sparsity* ($k = \alpha p$ for some $\alpha > 0$), as well as *sublinear sparsity* (meaning that $k/p \to 0$). In this paper, the bulk of our results concern the problem of *signed support recovery*, defined more precisely as follows.

For any vector $\beta \in \mathbb{R}^p$, we define its extended sign vector

$$\mathbb{S}_{\pm}(\beta_i) \quad := \quad \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0, \end{cases} \tag{6}$$

which encodes the *signed support* of the vector. Of interest to us are the following two questions:

- achievability results: under what scalings of (n, p, k) does the Lasso (3) have a *unique solution* $\widehat{\beta}$ that recovers the signed support ($\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)$)?
- converse results: under what scalings of (n, p, k) does *no solution* of the Lasso specify the correct signed support?

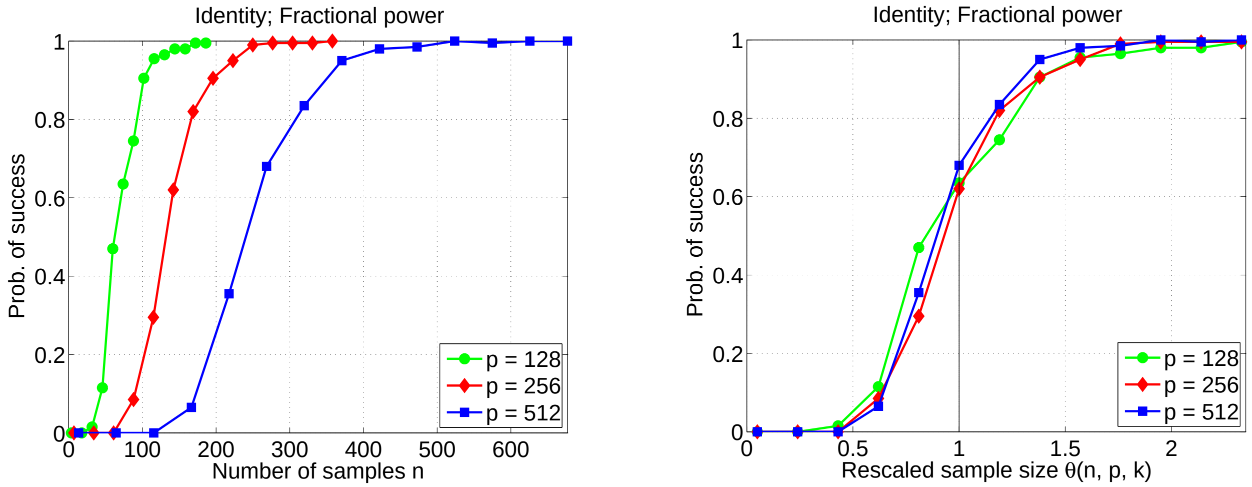
We analyze these questions both for deterministic designs (meaning the measurement matrix X is viewed as fixed) and random designs (X drawn from random ensembles). We begin by providing sufficient conditions for Lasso-based recovery to succeed with high probability over the random observation noise, when applied to deterministic designs. Moving to the case of random designs, we then sharpen this analysis by proving thresholds for the success/failure of the Lasso for various classes of Gaussian random measurement ensembles. Our analysis of the Gaussian random designs can be understood as revealing the *sample complexity* of Lasso-based sparsity recovery, meaning how the sample size n must scale with the problem parameters (p, k) if exact sparsity recovery is to be obtained using the Lasso.

To provide some intuition, panel (a) of Figure 1 plots the probability of successful support recovery $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)]$ versus the sample size n for three different problem sizes $p \in \{128, 256, 512\}$, and $k = \lceil 0.40 p^{0.75} \rceil$ in each case. Each point on each curve corresponds to the average over 200 trials, in each case drawing $X \in \mathbb{R}^{n \times p}$ randomly from the standard Gaussian ensemble ($X_{ij} \sim N(0, 1)$, i.i.d), and drawing $w \sim N(0, \sigma^2 I)$, with $\sigma = 0.50$. Note that each curve starts at $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)] = 0$ for small sample sizes n, and then reach $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)] = 1$ for sufficiently large sample sizes. Of course, the transition point from failure to success depends on the problem size p, with larger problems requiring more samples. This observation raises the natural question: *what is the scaling law that links the problem size p and sparsity index k to the sample size n?* One contribution of our theory is to provide a precise specification of this scaling law. Panel (b) of Figure 1 shows the same experimental results, with the probability of support recovery now plotted versus an "appropriately rescaled" version of the sample size, where the scaling is predicted by our theory. Note that as predicted by our theory, all of the curves now line up with one another, even though the problem sizes and sparsity indices vary dramatically. In Section VII, we show qualitatively similar results for different sparsity scalings (behavior of k as a function of p), and more general measurement ensembles, thereby showing excellent agreement between theoretical prediction and empirical behavior.

In analytical terms, our main result on Gaussian random ensembles (Theorems 3 and 4) show that there exist a pair of constants $0 < \theta_\ell(\Sigma) \leq \theta_u(\Sigma) < +\infty$, depending on the covariance matrix Σ such that the following properties hold. First, for sequences (n, p, k) such that the rescaled sample size $\frac{n}{2k \log(p-k)} > \theta_u(\Sigma)$, it is always possible to choose the regularization parameter λ_n such that the Lasso has a unique solution $\widehat{\beta}$ with $\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)$ with probability converging to one (over the choice of noise vector w and random matrix X). Conversely, whenever the rescaled sample size satisfies $\frac{n}{2k \log(p-k)} < \theta_\ell(\Sigma)$, then for whatever regularization parameter $\lambda_n > 0$ is chosen, no solution of the Lasso correctly specifies the signed support with probability converging to one. Although inachievability results of this type have been established for the basis pursuit LP in the noiseless setting [13], to the best of our knowledge, our lower bound for the Lasso is the first set of necessary conditions for exact sparsity recovery in the noisy setting. For the special case of the uniform Gaussian ensemble considered in past work (i.e., $\Sigma = I$, so that $X_{ij} \sim N(0, 1)$, i.i.d.), we show that $\theta_\ell(I) = \theta_u(I) = 1$, 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 IV-B, in the special case of *linear sparsity* (i.e., $k/p \rightarrow \alpha$ for some $\alpha > 0$), this theorem provides a noisy analog of results previously established for basis pursuit in the noiseless case [10], [4], [13]. Moreover, our result can also be adapted to an entirely different scaling regime, in which the sparsity index is *sublinear* ($k/p \rightarrow 0$), as considered by a separate body of recent work [27], [37] on the high-dimensional Lasso.

The remainder of this paper is organized as follows. We begin in Section II with some necessary and sufficient conditions, based on standard optimality conditions for convex programs, for the Lasso to have a unique solution that recovers the correct signed support. We then prove a consistency result for the case of deterministic design matrices X. Section IV is devoted to the statements of our main result on the asymptotic behavior of the Lasso for random Gaussian ensembles, and discussion of some of their consequences. Proofs are provided in Sections V and VI. We illustrate our theoretical results via simulation in Section VII, and conclude with a discussion in Section VIII.

Notation: We collect here some standard notation used throughout the paper. Throughout the paper, we use the notation c_1, c_2 etc. to refer to positive constants, whose value may differ from line to line. Given sequences $f(n)$ and $g(n)$, the notation $f(n) = \mathcal{O}(g(n))$ means that there exists a constant $c_1 < \infty$ such that $f(n) \leq c_1 g(n)$; the notation $f(n) = \Omega(g(n))$ means that there exists a constant $c_2 > 0$ such that $f(n) \geq c_2 g(n)$; and the notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. The symbol $f(n) = o(g(n))$ means that $f(n)/g(n) \rightarrow 0$. For parameters $a, b \in [1, \infty]$ and a matrix M, we define the ℓ_a/ℓ_b operator norm $|\!|\!|M|\!|\!|_{a,b} := \max\limits_{\|x\|_a = 1} \|Mx\|_b$. Important special cases include the ℓ_2/ℓ_2 operator norm, also known as the spectral norm, denoted by $|\!|\!|M|\!|\!|_{2,2}$ or $|\!|\!|M|\!|\!|_2$ for short, and the ℓ_∞/ℓ_∞ operator norm, given by $|\!|\!|M|\!|\!|_{\infty,\infty} = \max_i \sum_j |M_{ij}|$, and denoted by $|\!|\!|M|\!|\!|_\infty$ for short.



(a)

(b)

Fig. 1. (a) Plots of the success probability $\mathbb{P}[\mathbb{S}_{\pm}(\widehat{\beta}) = \mathbb{S}_{\pm}(\beta^*)]$ of obtaining the correct signed support versus the sample size n for three different problem sizes p, in all cases with sparsity $k = \lceil 0.40 p^{0.75} \rceil$. (b) Same simulation results with success probability plotted versus the rescaled sample size $\theta(n,p,k) = n/[2k\log(p-k)]$. As predicted by Theorems 3 and 4, all the curves now lie on top of one another. See Section VII for further simulation results.

II. Background and primal-dual witness construction

In this section, we begin by developing the convex-analytic conditions that characterize the optima of the ℓ_1-regularized quadratic program (3). We then specify the construction that underlies the proofs of our main results, and prove some elementary lemmas that show how it characterizes the success (or failure) of the Lasso in recovering the correct support set. We refer to this method as a *primal-dual witness*, since it is based on an explicit construction of a pair of vectors that (when the procedure succeeds) are a primal and dual optimal solutions for the Lasso, and act as a witnesses for the correct recovery of the support.

A. Convex optimality and uniqueness

We begin with some basic observations about the Lasso problem (3). First, the minimum in the Lasso is always achieved by at least one vector $\beta \in \mathbb{R}^p$. This fact follows from the Weierstrass theorem, because in its ℓ_1-constrained form (4), the minimization is over a compact set, and the objective function is continuous. Second, although the problem is always convex, it is not always strictly convex, so that the optimum can fail to be unique. Indeed, a little calculation shows that the Hessian of the quadratic component of the objective is the $p \times p$ matrix $X^T X/n$, which is positive definite but not strictly so whenever $p > n$. Nonetheless, as stated below in Lemma 1, strict dual feasibility conditions are sufficient to ensure uniqueness, even under high-dimensional scaling ($p \gg n$).

The objective function is not always differentiable, since the ℓ_1-norm is a piecewise linear function. However, the optima of the Lasso (3) can be characterized by a zero subgradient condition. A vector $z \in \mathbb{R}^p$ is a subgradient for the ℓ_1-norm evaluated at $\beta \in \mathbb{R}^p$, written as $z \in \partial\|\beta\|_1$, if its elements satisfy the relations

$$z_i = \operatorname{sign}(\beta_i) \text{ if } \beta_i \neq 0, \text{ and } z_i \in [-1,+1], \text{ otherwise.} \quad (7)$$

For any subset $A \subseteq \{1,2,\ldots,p\}$, let X_A be the $n \times |A|$ matrix formed by concatenating the columns $\{X_i, i \in A\}$ indexed by A. For any vector $\beta \in \mathbb{R}^p$, we define its *support set* $S(\beta) = \{i \mid \beta_i \neq 0\}$. With these definitions, we state the following:

Lemma 1. (a) *A vector $\widehat{\beta} \in \mathbb{R}^p$ is optimal if and only if there exists a subgradient vector $\widehat{z} \in \partial\|\widehat{\beta}\|_1$ such that*

$$\frac{1}{n} X^T X(\widehat{\beta} - \beta^*) - \frac{1}{n} X^T w + \lambda_n \widehat{z} \quad = \quad 0. \quad (8)$$

(b) *Suppose that the subgradient vector satisfies the strict dual feasibility condition $|\widehat{z}_j| < 1$ for all $j \notin S(\widehat{\beta})$. Then* any *optimal solution $\widetilde{\beta}$ to the Lasso satisfies $\widetilde{\beta}_j = 0$ for all $j \notin S(\widehat{\beta})$.*

(c) *Under the conditions of part (b), if the $k \times k$ matrix $X^T_{S(\widehat{\beta})} X_{S(\widehat{\beta})}$ is invertible, then $\widehat{\beta}$ is the unique optimal solution of the Lasso program.*

The proof is provided in Appendix B.

B. Primal-dual witness construction

We now turn to the proof technique that underlies our main results. Using S as a shorthand for the support set $S(\beta^*)$ of the true vector β^*, we assume throughout that the $k \times k$ matrix $X_S^T X_S$ is invertible. Under this condition, the *primal-dual witness* (PDW) method consists of constructing a pair $(\check{\beta}, \check{z}) \in \mathbb{R}^p \times \mathbb{R}^p$ according to the following steps:

1) First, we obtain $\check{\beta}_S \in \mathbb{R}^k$ by solving the *restricted* Lasso problem,

$$\check{\beta}_S \;=\; \arg \min_{\beta_S \in \mathbb{R}^k} \Big\{ \frac{1}{2n} \|y - X_S \beta_S\|_2^2 + \lambda_n \|\beta_S\|_1 \Big\}. \quad (9)$$

The solution to this restricted convex program is guaranteed to be unique under the invertibility condition on $X_S^T X_S$. We set $\check{\beta}_{S^c} = 0$.

2) Second, we choose $\check{z}_S \in \mathbb{R}^k$ as an element of the subdifferential of the ℓ_1 norm evaluated at $\check{\beta}_S$.
3) Third, we solve for a vector $\check{z}_{S^c} \in \mathbb{R}^{p-k}$ satisfying the zero subgradient condition (8), and check whether or not the *dual feasibility condition* $|\check{z}_j| \leq 1$ for all $j \in S^c$ is satisfied. (For ensuring uniqueness, we check for strict dual feasibility, i.e., $|\check{z}_j| < 1$ for all $j \in S^c$.)
4) Fourth, we check whether the *sign consistency condition* $\check{z}_S = \mathrm{sign}(\beta^*_S)$ is satisfied.

To be clear, this procedure is *not* a practical method for solving the ℓ_1-regularized quadratic program (3), since solving the restricted problem in Step 1 requires knowledge of the unknown support set S. Rather, the utility of this constructive procedure is as a proof technique: it succeeds if and only if the Lasso has a unique optimal solution with the correct signed support. This characterization allows us to certify support consistency properties of the Lasso, as summarized by the following result, proved in Appendix C:

Lemma 2. *Assume that $X_S^T X_S$ is invertible.*

(a) *If Steps 1 through 3 of the PDW method succeed with strict dual feasiblity in Step 3, then the Lasso* (3) *has a unique solution $\widehat{\beta}$ with $S(\widehat{\beta}) \subseteq S(\beta^*)$.*
(b) *If Steps 1 through 4 succeed with strict dual feasibility in Step 3, then Lasso* (3) *has a unique solution $\widehat{\beta}$ with the correct signed support (i.e., $\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)$).*
(c) *Conversely, if either Steps 3 or 4 of the PDW method fail, then the Lasso fails to recover the correct signed support.*

The challenges in the primal-dual witness construction lie in verifying the *dual feasibility condition* in Step 3, and the *sign consistency condition* in Step 4. More specifically, whether or not these steps are successful depends on the behavior of certain random variables, associated with the support S and non-support S^c of the true solution β^*. In particular, we define for each $j \in S^c$, the scalar random variable

$$Z_j := X_j^T \Big\{ X_S \big(X_S^T X_S\big)^{-1} \check{z}_S + \Pi_{X_S^\perp}\big(\frac{w}{\lambda_n n}\big) \Big\} \tag{10}$$

where $\Pi_{X_S^\perp} := I_{n \times n} - X_S \big(X_S^T X_S\big)^{-1} X_S^T$ is an orthogonal projection matrix, and $\check{z}_S$ is the subgradient vector chosen in Step 2 of the PDW method. Moreover, for each $i \in S$, we define the scalar random variable

$$\Delta_i \;\; := \;\; e_i^T \big(\frac{1}{n} X_S^T X_S\big)^{-1} \big[\frac{1}{n} X_S^T w - \lambda_n \,\mathrm{sgn}(\beta^*_S)\big]. \tag{11}$$

As formalized by the following lemma, Z_j is the candidate dual variable solved for in Step 3 of the primal-dual construction. On the other hand, if the Lasso is sign-consistent, the variable Δ_i is equal to the difference $\widehat{\beta}_i - \beta^*_i$ at position i between the Lasso solution $\widehat{\beta}$ and the truth β^*.

Lemma 3. *Assume that the matrix $X_S^T X_S$ is invertible. Then*

(a) *The dual feasibility check in Step 3 of the PDW method succeeds if and only if*

$$|Z_j| \leq 1 \quad \textit{for all } j \in S^c. \tag{12}$$

For strict dual feasibility, these inequalities must hold strictly.

(b) *The sign consistency condition in Step 4 of the PDW method can be satisfied if and only if*

$$\mathrm{sgn}\big\{\beta^*_i + \Delta_i\big\} \;\; = \;\; \mathrm{sgn}(\beta^*_i) \;\; \textit{for all } i \in S. \tag{13}$$

See Appendix D for the proof of this claim.

III. Analysis of deterministic designs

In this section, we show how the primal-dual witness construction can be used to analyze the behavior of the Lasso in the case of a deterministic (non-random) design matrix $X \in \mathbb{R}^{n \times p}$, and observation noise vectors $w \in \mathbb{R}^n$ from a sub-Gaussian distribution (see Section A for background on sub-Gaussian random variables). We begin by stating a positive result (Theorem 1) that provides sufficient conditions for Lasso success with high probability over the noise vectors, and then discuss some of its consequences. Our second result on deterministic designs (Theorem 2) isolates some conditions that are sufficient to guarantee failure of the Lasso. Both of these results are proved using the link between the primal-dual witness (PDW) method and the success/failure of the Lasso, as stated in Lemmas 2 and 3.

A. Sufficient conditions and some consequences

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, and moreover to assume that we are seeking signed support recovery, so that we need $\check{z}_S = \mathrm{sign}(\beta^*_S)$. Under these conditions, assuming that $\lambda_n > 0$, the conditions of Lemma 3 reduce to

$$\begin{aligned} \max_{j \in S^c} \big|X_j^T X_S \big(X_S^T X_S\big)^{-1} \mathrm{sign}(\beta^*_S)\big| \;\; &\leq \;\; 1 \\ \mathrm{sgn}\big(\beta^*_i - e_i^T \lambda_n \big(\frac{1}{n} X_S^T X_S\big)^{-1} \mathrm{sgn}(\beta^*_S)\big) \;\; &= \;\; \mathrm{sgn}(\beta^*_i), \end{aligned}$$

where the latter equality must hold for all $i \in S$. If the primal-dual witness method fails 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: first, there exists some *incoherence parameter* $\gamma \in (0, 1]$ such that

$$|\!|\!| X_{S^c}^T X_S \big(X_S^T X_S\big)^{-1} |\!|\!|_\infty \;\; \leq \;\; (1 - \gamma), \tag{15}$$

and $|\!|\!| \cdot |\!|\!|_\infty$ denotes the ℓ_∞/ℓ_∞ operator norm,[2] and second, there exists some $C_{min} > 0$ such that

$$\Lambda_{min}\big(\frac{1}{n} X_S^T X_S\big) \geq C_{min}, \tag{16}$$

where Λ_{min} denotes the minimal eigenvalue. Mutual incoherence conditions of the form (15) have been considered in previous work on the Lasso, initially by Fuchs [18] and

[2] Recall that for an $m \times n$ matrix M, this norm is given by $|\!|\!|M|\!|\!| := \max_{i=1,\ldots m} \sum_{j=1}^n |M_{ij}|$.

Tropp [32]. Other authors [27], [37] provide examples and results on matrix families that satisfy this type of condition.

Consider the linear observation model (1) with fixed design $X \in \mathbb{R}^{n \times p}$ and additive noise vector $w \in \mathbb{R}^n$ with i.i.d. entries that are zero-mean and sub-Gaussian[3] with parameter $\sigma > 0$. With this set-up, we have the following set of sufficient conditions for sparsity recovery for deterministic designs:

Theorem 1. *Assume that the design matrix X satisfies conditions (15) and (16), and has its n-dimensional columns normalized such that $n^{-1/2} \max_{j \in S^c} \|X_j\| \leq 1$. Suppose that the sequence of regularization parameters $\{\lambda_n\}$ satisfies*

$$\lambda_n \quad > \quad \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log p}{n}}. \tag{17}$$

Then for some constant $c_1 > 0$, the following properties hold with probability greater than $1 - 4\exp(-c_1 n \lambda_n^2) \to 1$:

(a) *The Lasso has a unique solution $\widehat{\beta} \in \mathbb{R}^p$ with its support contained within the true support (i.e., $S(\widehat{\beta}) \subseteq S(\beta^*)$), and satisfies the ℓ_∞ bound:*

$$\|\widehat{\beta}_S - \beta^*_S\|_\infty \quad \leq \quad \underbrace{\lambda_n \big[\big\|(X_S^T X_S/n)^{-1}\big\|_\infty + \frac{4\sigma}{\sqrt{C_{min}}} \big]}_{g(\lambda_n)} . \tag{18}$$

(b) *If in addition the minimum value of the regression vector β^*_S on its support is bounded below as $\beta_{\min} > g(\lambda_n)$, then $\widehat{\beta}$ has the correct signed support (i.e., $\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)$).*

Remarks: (1) First, it is worthwhile to consider Theorem 1 in the classical setting, in which the number of samples n tends to infinity, but the model dimension p and sparsity index k do not depend on n. Knight and Fu [22] established consistency of the Lasso with additive Gaussian noise under the scalings $\lambda_n \to 0$ and $n\lambda_n^2 \to +\infty$, which[4] guarantee that the conditions of Theorem 1 hold, and moreover that the probability of success converges to one. For instance, one could choose $\lambda_n^2 = 1/n^{1-\delta}$ for some $\delta > 0$, which would then guarantee recovering, with probability greater than $1 - 2\exp(-c_1 n^\delta)$, all signals β^* with minimum value $\beta_{\min} = \Omega(n^{-\frac{1-\delta}{2}})$.

(2) Theorem 1 is also related to past work by Fuchs [18], Tropp [32] and Donoho et al. [12], who (among various other results) provided sufficient conditions for Lasso-based support recovery under mutual incoherence constraints. These sufficient conditions were deterministic in nature, based on either assuming certain bounds on the ℓ_2-norm of the noise vector w, or on the ℓ_2-norm of the best k-term approximation to the input signal y (e.g., Thm. 4 [18], Thm. 8 [32], or Thm 4.1 [12]). Theorem 1 can be viewed as a stochastic analog of such results, in which we study the probability of successful support recovery as the problem size p and sparsity index k scale with the sample size n.

(3) Theorem 1 is also related to an independent body of past work by Meinshausen and Buhlmann [27] and Zhao and Yu [37], who studied the probability of correct support recovery as (p, k) scaled in a very specific way with sample size. Specializing Theorem 1 to their particular scaling recovers the following corollary [37]:

Corollary 1. *Suppose that the design matrix X satisfies the conditions of Theorem 1, and that $p = O(\exp(n^{\delta_3}))$, $k = O(n^{\delta_1})$, and moreover that $\beta_{\min}^2 > 1/n^{(1-\delta_2)}$ with*

$$0 \ < \ \delta_1 + \delta_3 \ < \ \delta_2 \ < \ 1.$$

If we set $\lambda_n^2 = 1/n^{1-\delta_4}$ for some $\delta_4 \in (\delta_3, \delta_2 - \delta_1)$, then under the conditions of Theorem 1, the Lasso recovers the sparsity pattern with probability $1 - \exp(-c_1 n^{\delta_4})$.

Proof: We simply need to verify that the conditions of Theorem 1 are satisfied with these particular choices. First, with the given choice of regularizer, we have

$$\frac{\log p}{n\lambda_n^2} = \mathcal{O}\big(n^{\delta_3}/(n^{\delta_4 - 1}n)\big) \ = \ \mathcal{O}(n^{\delta_3 - \delta_4}) \to 0,$$

so that the inequality (17) certainly holds. Since $X_S^T X_S$ is a $k \times k$ matrix, we have

$$\begin{aligned} \lambda_n \, \|\frac{1}{n}(X_S^T X_S)^{-1}\|_\infty \quad &\leq \quad \lambda_n \, \sqrt{k} \|\frac{1}{n}(X_S^T X_S)^{-1}\|_2 \\ &\leq \quad \frac{\lambda_n \sqrt{k}}{C_{min}} \end{aligned}$$

Therefore, in order to check the condition from Theorem 1(b) on the minimum value, it is sufficient to show that $\frac{\lambda_n \sqrt{k}}{\beta_{\min}} = o(1)$. Under the given scalings, we have

$$\frac{\lambda_n^2 k}{\beta_{\min}^2} = \mathcal{O}\big(n^{\delta_4 - 1} n^{\delta_1} n^{(1-\delta_2)}\big) \ = \ \mathcal{O}(n^{\delta_4 + \delta_1 - \delta_2}) \to 0,$$

showing that the condition on $\beta_{\min}$ holds.

□

The scaling given in Corollary 1 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 an extremely strong sparsity constraint, since the ratio $k/p \cong n^{\delta_1} \exp(-n^{\delta_3})$ vanishes exponentially fast. If we consider general scaling of the triplet (n, p, k), Theorem 1 suggests that having n on the order of $k \log p$ is appropriate, since this sample size would allow the minimum value $\beta_{\min}$ to decay as $1/\sqrt{k}$. In Section IV, we show that this type of scaling is both necessary and sufficient for almost all matrices drawn from suitable Gaussian designs.

(4) Of course, Theorem 1 has consequences for estimation of β^* in ℓ_2 norm. In particular, assuming that $\|(X_S^T X_S)/n\|_\infty = \mathcal{O}(1)$ for simplicity, the ℓ_∞ bound (18), in conjunction with the k-sparsity of β^*, implies that

$$\|\widehat{\beta} - \beta^*\|_2 \quad = \quad \mathcal{O}(\lambda_n \sqrt{k}) \ = \mathcal{O}(\sqrt{\frac{k \log p}{n}}),$$

where we have chosen $\lambda_n = \mathcal{O}(\sqrt{\frac{\log p}{n}})$. Of course, given that Theorem 1 guarantees recovery of the support S, one could

[3] See Appendix A for background.

[4] To be clear, our regularization parameter λ_n is related to the choice ρ_n of the paper [22] via the relation $\lambda_n = \rho_n/n$.

obtain a better ℓ_2-estimate by simply performing ordinary regression restricted to the support S.

B. Some necessary conditions

We now turn to some partial inachievability results, providing sufficient conditions for failure of the Lasso (3) in recovering the support set:

Theorem 2 (Inachievability for deterministic design)**.** *Suppose that the eigenvalue condition* (16) *holds, and the noise vector has a distribution symmetric around zero.*

(a) *Suppose that the mutual incoherence condition* (15) *is violated – say*

$$\max_{j \in S^c} |X_j^T X_S (X_S^T X_S)^{-1} \operatorname{sgn}(\beta^*_S)| = 1 + \nu \; > \; 1. \quad (19)$$

Then for any $\lambda_n > 0$ and for any *sample size n, the probability of correct signed support recovery is bounded away from one — namely*

$$\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)] \quad \leq \quad 1/2. \qquad (20)$$

(b) *For each $i \in S$, define the quantity*

$$\widetilde{g}_i(\lambda_n) \quad = \quad \lambda_n e_i^T (\frac{X_S^T X_S}{n})^{-1} \operatorname{sgn}(\beta^*_S). \qquad (21)$$

Suppose that for some $i \in S$, we have the inclusion $\beta_i^* \in (0, \widetilde{g}_i(\lambda_n))$ or the inclusion $\beta_i^* \in (\widetilde{g}_i(\lambda_n), 0)$. Then the probability of correct signed support recovery is bounded away from one:

$$\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)] \quad \leq \quad 1/2. \qquad (22)$$

Theorem 2(a) is a precise statement of the fact that the mutual incoherence condition (15) is an essential requirement for support recovery using the Lasso; see Zhao and Yu [37] for related observations. Theorem 2(b) reveals two factors that are important in signed support consistency: the conditioning of matrix $(X_S^T X_S/n)$, and the magniduate of the regularization parameter λ_n relative to the minimum value $\beta_{\min} = \min_{i \in S} |\beta_i^*|$.

With regards to the former issue, the ideal case is when the columns of X_S are orthogonal, in which case the matrix is simply the identity. More generally, control on ℓ_∞-operator norm $\|(X_S^T X_S/n)^{-1}\|_\infty$ or a related quantity, as needed for the ℓ_∞-bound (18) in Theorem 1 to be reasonably tight, is required for sign consistency. With reference to the latter issue, the quantity $\widetilde{g}_i(\lambda_n)$ corresponds to the amount by which the Lasso estimate in position $i \in S$ is "shrunken", and it imposes the constraint that the value $\beta_{\min}$ cannot decay to zero faster than the regularization parameter λ_n.

C. Proof of Theorem 1

The proof of Theorem 1 consists of two main parts: we first establish that the random variables $\{Z_j, j \in S^c\}$ previously defined (10) satisfy strict dual feasibility with high probability, so that Step 3 of the PDW construction succeeds. We then establish an ℓ_∞ bound on the variables $\{\Delta_i, i \in S\}$ previously defined (11), which (under the assumptions of Theorem 1(b)) ensures that the sign consistency condition in Step 4 of the PDW construction holds.

Establishing strict dual feasibility: We begin by establishing that Step 3 of the primal-dual witness condition succeeds with high probability. Recalling the definition (10), note that we have the decomposition $Z_j = \mu_j + \widetilde{Z}_j$, where

$$\mu_j \quad = \quad X_j^T X_S (X_S^T X_S)^{-1} \check{z}_S, \qquad (23)$$

and $\widetilde{Z}_j := X_j^T \Pi_{X_S^\perp}(\frac{w}{\lambda_n n})$ a zero-mean sub-Gaussian noise variable. Since $\check{z}_S \in \mathbb{R}^k$ is a subgradient vector (chosen in Step 2) for the ℓ_1 norm, we have $\|\check{z}_S\|_\infty \leq 1$. Applying the incoherence condition (15) yields that $|\mu_j| \leq (1-\gamma)$ for all indices $j \in S^c$, from which we obtain that

$$\max_{j \in S^c} |Z_j| \;\leq\; (1-\gamma) + \max_{j \in S^c} |\widetilde{Z}_j|.$$

Since the elements of w are zero-mean and sub-Gaussian with parameter σ^2, it follows from property (49) that the variable $\widetilde{Z}_j$ is sub-Gaussian with parameter at most

$$\frac{\sigma^2}{\lambda_n^2 n^2} \, \|\Pi_{X_S^\perp}(X_j)\|_2^2 \quad \leq \quad \frac{\sigma^2}{\lambda_n^2 n},$$

where we have used the facts that the projection matrix $\Pi_{X_S^\perp}$ has spectral norm one, and the condition $n^{-1} \max_j \|X_j\|_2^2 \leq 1$. Consequently, by the sub-Gaussian tail bound (48) combined with the union bound, we obtain

$$\mathbb{P}[\max_{j \in S^c} |\widetilde{Z}_j| \geq t] \quad \leq \quad 2(p-k) \; \exp\big(-\frac{\lambda_n^2 \, n \, t^2}{2\sigma^2}\big).$$

Setting $t = \frac{\gamma}{2}$ yields that

$$\mathbb{P}[\max_{j \in S^c} |\widetilde{Z}_j| \;\geq\; \frac{\gamma}{2}] \;\leq\; 2 \exp\big\{-\frac{\lambda_n^2 \, n \, \gamma^2}{8\sigma^2} + \log(p-k)\big\}.$$

Putting together the pieces and using our choice (17) of λ_n, we conclude that

$$\mathbb{P}[\max_{j \in S^c} |Z_j| > 1 - \frac{\gamma}{2}] \quad \leq \quad 2 \exp\big\{-c_1 \lambda_n^2 n\big\} \;\rightarrow\; 0.$$

Establishing ℓ_∞ bounds: Next we establish a bound on the ℓ_∞-norm of the random vector Δ_S from equation (11). By the triangle inequality, the quantity $\max_{i \in S} |\Delta_i|$ is upper bounded by

$$\|(\frac{X_S^T X_S}{n})^{-1} X_S^T \frac{w}{n}\|_\infty + \|(\frac{X_S^T X_S}{n})^{-1}\|_\infty \, \lambda_n. \qquad (24)$$

The second term is a deterministic quantity, so that it remains to bound the first term. For each $i = 1, \ldots, k$, consider the random variable

$$V_i \quad := \quad e_i^T (\frac{1}{n} X_S^T X_S)^{-1} \frac{1}{n} X_S^T w.$$

Since the elements of w are zero-mean and i.i.d. sub-Gaussian with parameter σ^2, it follows from property (49) that V_i is zero-mean and sub-Gaussian with parameter at most

$$\frac{\sigma^2}{n} \|(\frac{1}{n} X_S^T X_S)^{-1}\|_2 \quad \leq \quad \frac{\sigma^2}{C_{min} n}.$$

Consequently, by the sub-Gaussian tail bound (48) and the union bound, we have

$$\mathbb{P}[\max_{i=1,\ldots,k} |V_i| > t] \quad \leq \quad 2\exp\big(-\frac{t^2 C_{min} n}{2\sigma^2} + \log k\big).$$

We set $t = 4\sigma\lambda_n/\sqrt{C_{min}}$, and note that by our choice (17) of λ_n, we have the inequality $8n\lambda_n^2 > \log p \geq \log k$. Putting together these pieces, we conclude that the probabability $\mathbb{P}[\max_{i=1,\ldots,k} |Z_i| > 4\sigma\lambda_n/\sqrt{C_{min}}]$ vanishes at rate at least $2\exp(-c_2\lambda_n^2 n)$. Overall, we conclude that

$$\|\widehat{\beta}_S - \beta_S^*\|_\infty \quad \leq \quad \lambda_n\big[\frac{4\sigma}{\sqrt{C_{min}}} + \big\|(X_S^T X_S/n)^{-1}\big\|_\infty\big],$$

with probability greater than $1 - 2\exp(-c_2\lambda_n^2 n)$, as claimed.

D. Proof of Theorem 2

We prove part (a) by showing that either the sign consistency check in Step 4, or the dual feasibility check in Step 3 of the PDW must fail with probability at least $1/2$. We may assume that $\check{z}_S = \operatorname{sign}(\beta_S^*)$; otherwise, the sign consistency condition fails. So it remains to show that under this condition, the dual feasibility condition in Step 3 fails with probability at least $1/2$. Let $j \in S^c$ be an index for which the maximum in the violating condition (19) is achieved. From proof of Theorem 1, we have the decomposition $Z_j = \mu_j + \widetilde{Z}_j$ with $\mu_j = X_j^T X_S (X_S^T X_S)^{-1}\operatorname{sign}(\beta_S^*)$, using the fact that $\check{z}_S = \operatorname{sign}(\beta_S^*)$. Without loss of generality, we may assume that $\mu_j = 1 + \nu$, as the argument with $\mu_j = -1 - \nu$ is entirely analogous by symmetry. Note that since w is symmetric about zero by assumption, the random variable $\widetilde{Z}_j$ from equation (10) is also symmetric. Using this symmetry and the representation $Z_j = (1+\nu) + \widetilde{Z}_j$, we conclude that $\mathbb{P}[Z_j > 1] \geq 1/2$. Applying Lemmas 2 and 3, we conclude that $\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) \neq \mathbb{S}_\pm(\beta^*)] \geq 1/2$ as claimed. Note that this claim holds for any sample size.

We prove the claim (b) by analyzing by using Lemma 3(b). In order for the Lasso to recover the correct signed support, we must have $\check{z}_S = \operatorname{sign}(\beta_S^*)$, and the condition $\operatorname{sign}(\beta_i^* + \Delta_i) = \operatorname{sign}(\beta_i^*)$ must hold for all $i \in S$. Without loss of generality, let us assume that $\beta_i^* \in (0, \widetilde{g}_i(\lambda_n))$. We then have

$$\beta_i^* + \Delta_i \quad = \quad \underbrace{\beta_i^* - \widetilde{g}_i(\lambda_n)}_{D_i} + \underbrace{e_i^T\big(\frac{1}{n}X_S^T X_S\big)^{-1}\frac{1}{n}X_S^T w}_{\widetilde{w}_i},$$

where we have used the definition (11) of Δ. Since the deterministic quantity $D_i < 0$ by assumption, and the noise variable $\widetilde{w}_i$ is symmetric around zero, we have $\mathbb{P}[\operatorname{sign}(\beta_i^* + \Delta_i) \neq \operatorname{sign}(\beta_i^*)] \geq 1/2$, which implies that the probability of success is upper bounded by $1/2$.

IV. Random Gaussian ensembles: Thresholds for sparsity recovery

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 $X \in \mathbb{R}^{n\times p}$, in which each row x_i, $i = 1,\ldots,n$ is chosen as an i.i.d. Gaussian random vector with covariance matrix Σ. In this setting, we specify explicit threshold functions of the triple (n,p,k) and covariance matrix Σ that govern the *success and failure* of the Lasso over a given Gaussian ensemble (Theorems 3 and 4 respectively). Note that our Gaussian ensemble results cover not only the standard Gaussian ensemble ($\Sigma = I_{p\times p}$), but also more general Gaussian designs. We begin by setting up and providing a precise statement of the main results, and then discussing their connections to previous work. In the later part of this section, we provide the proofs.

A. Statement of main results

As before, we consider the noisy linear observation model except that the measurement matrix $X \in \mathbb{R}^{n\times p}$ is now random—namely,

$$y \quad = \quad X\beta^* + w, \quad \text{with i.i.d. rows } x_i \sim N(0,\Sigma). \tag{25}$$

Our results are based on imposing (subsets of) the following conditions on the covariance matrices forming the design:

$$\begin{aligned} \|\Sigma_{S^cS}(\Sigma_{SS})^{-1}\|_\infty &\leq (1-\gamma) \quad \text{for some } \gamma \in (0,1], && \text{(26a)}\\ \Lambda_{min}(\Sigma_{SS}) &\geq C_{min} > 0, \quad \text{and} && \text{(26b)}\\ \Lambda_{max}(\Sigma_{SS}) &\leq C_{max} < +\infty. && \text{(26c)} \end{aligned}$$

Note that conditions (26a) and (26b) are simply the population analogs of the conditions (15) and (16) imposed previously on the deterministic designs. The upper bound (26c) is required only for establishing the inachievability claim—namely, sufficient conditions for failure of the Lasso. The simplest example of a covariance matrix satisfying these conditions is the identity $\Sigma = I_{p\times p}$, for which we have $C_{min} = C_{max} = 1$, and $\gamma = 1$. Another well-known matrix family satisfying these conditions are Toeplitz matrices.

For a positive definite symmetric matrix A, we define

$$\rho_\ell(A) := \frac{1}{2}\min_{i\neq j}\big(A_{ii} + A_{jj} - 2A_{ij}\big), \text{ and } \rho_u(A) := \max_i A_{ii}. \tag{27}$$

We note that $A \succeq 0$ implies that $|A_{ij}| \leq \sqrt{A_{ii}A_{jj}}$, and hence that $\rho_\ell(A) \geq 0$, and moreover

$$\rho_\ell(A) \quad \leq \quad \frac{1}{2}\max_{i\neq j}(\sqrt{A_{ii}} + \sqrt{A_{jj}})^2 \leq \rho_u(A).$$

The threshold constants in our result involve the conditional covariance matrix of $(X_{S^c} \mid X_S)$, namely

$$\Sigma_{S^c|S} := \Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c} \succeq 0. \tag{28}$$

In particular, we define

$$\begin{aligned} \theta_\ell(\Sigma) &:= \frac{\rho_\ell(\Sigma_{S^c|S})}{C_{max}\,(2-\gamma(\Sigma))^2}, \text{ and} && \text{(29a)}\\ \theta_u(\Sigma) &:= \frac{\rho_u(\Sigma_{S^c|S})}{C_{min}\,\gamma^2(\Sigma)}, && \text{(29b)} \end{aligned}$$

where $\gamma(\Sigma) \in (0,1]$ is the incoherence parameter (26a). It is straightforward to verify that we always have the inequalities

$$0 \leq \theta_\ell(\Sigma) \leq \theta_u(\Sigma) < \infty. \tag{30}$$

Equality holds for the standard Gaussian ensemble ($\Sigma = I_{p\times p}$), for which we have $C_{min} = C_{max} = \gamma = 1$, and moreover $\rho_\ell(\Sigma_{S^c|S}) = \rho_u(\Sigma_{S^c|S}) = 1$, so that $\theta_\ell(I_{p\times p}) = \theta_u(I_{p\times p}) = 1$.

Theorem 3 (Achievability)**.** *Consider the linear observation model with random Gaussian design* (25) *and noise* $w \sim N(0, \sigma^2 I_{n\times n})$*. Assume that the covariance matrices* Σ *satisfy conditions* (26a) *and* (26b)*, and consider the family of regularization parameters*

$$\lambda_n(\phi_p) \;=\; \sqrt{\frac{\phi_p\,\rho_u(\Sigma_{S^c|S})}{\gamma^2}\;\frac{2\sigma^2\,\log(p)}{n}}, \qquad (31)$$

for some $\phi_p \geq 2$*. If for some fixed* $\delta > 0$*, the sequence* (n, p, k) *and regularization sequence* $\{\lambda_n\}$ *satisfy*

$$\frac{n}{2k\log(p-k)} \;>\; (1+\delta)\,\theta_u(\Sigma)\big(1 + \frac{\sigma^2 C_{min}}{\lambda_n^2 k}\big), \qquad (32)$$

then the following properties holds with probability greater than $1 - c_1\exp(-c_2\min\{k, \log(p-k)\})$*:*

(i) *The Lasso has a unique solution* $\widehat{\beta}$ *with support contained within* S *(i.e.,* $S(\widehat{\beta}) \subseteq S(\beta^*)$*).*

(ii) *Define the gap*

$$g(\lambda_n) \;:=\; c_3\lambda_n\vert\vert\vert\Sigma_{SS}^{-1/2}\vert\vert\vert_\infty^2 + 20\sqrt{\frac{\sigma^2\log k}{C_{min}\,n}}. \qquad (33)$$

Then if $\beta_{\min} \;:=\; \min_{i\in S}|\beta_i^*| \;>\; g(\lambda_n)$*, the signed support* $\mathbb{S}_\pm(\widehat{\beta})$ *is identical to* $\mathbb{S}_\pm(\beta^*)$*, and moreover* $\|\widehat{\beta}_S - \beta_S^*\|_\infty \leq g(\lambda_n)$*.*

Remarks: It should be noted that the condition (32) couples together the required sample size n and the regularization parameter λ_n. In particular, for the family (31) of regularization parameters, the proof of Theorem 3 shows that it suffices to have

$$\frac{n}{2k\log(p-k)} \;>\; \frac{(1+\delta')}{1-\frac{1}{\phi_p}}\,\theta_u(\Sigma), \qquad (34)$$

for some $\delta' > 0$. Consequently, if we choose a sequence of regularization parameters (31) with $\phi_p \to +\infty$, then Theorem 3 guarantees recovery with $n = 2\theta_u(\Sigma)k\log(p-k)$ observations. More generally, if we choose λ_n in equation (31) with a constant $\phi_p \geq 2$, then we still obtain recovery with $n = \Omega(k\log(p-k))$ samples, although the pre-factor now depends on the precise choice of λ_n through the term ϕ_p.

Note that the decay rate of λ_n imposes limitations for signed support recovery, in particular how quickly the minimum value $\beta_{\min}$ is allowed to decay, since Theorem 3(b) guarantees success only if $\beta_{\min} = \Omega(\lambda_n)$. Consequently, with the choice (31) with a constant ϕ_p and $\vert\vert\vert\Sigma_{SS}^{-1/2}\vert\vert\vert_\infty = \mathcal{O}(1)$, Theorem 3 shows that the Lasso can recover the support of a signal $\beta^* \in \mathbb{R}^p$ for which $\beta_{\min} = \Omega(\sqrt{\frac{\log p}{n}})$.

Theorem 4 (Inachievability)**.** *Consider the linear observation model with random Gaussian design* (25) *and noise* $w \sim N(0, \sigma^2 I_{n\times n})$*. Assume that the covariance matrices satisfy conditions* (26a) *through* (26c)*. If for some fixed* $\delta > 0$*, the sequence* (n, p, k) *satisfies*

$$\frac{n}{2k\log(p-k)} \;<\; (1-\delta)\,\theta_\ell(\Sigma)\,\big(1 + \frac{C_{max}\sigma^2}{\lambda_n^2 k}\big), \qquad (35)$$

then with probability converging to one, no solution of the Lasso (3) *has the correct signed support.*

Remarks: Again, the simplest case is when the regularization parameter is chosen from the family (31) for some $\phi_p \to +\infty$. In this case, the inachievability result (35) is the weakest, in that it asserts only that the Lasso fails with high probability for $n < 2\theta_\ell(\Sigma)k\log(p-k)$.

It is also worth noting that the condition (35) imposes restrictions on the choice of λ_n. In particular, suppose that

$$\begin{aligned} \lambda_n^2 \;&<\; \frac{2\theta_\ell(\Sigma)\sigma^2 C_{max}\,\log(p-k)}{n} \\ &=\; \frac{\rho_\ell(\Sigma_{S^c|S})}{(2-\gamma)^2}\;\frac{2\sigma^2\log(p-k)}{n}. \end{aligned}$$

In this case, the condition (35) is always satisfied, so that the Lasso fails with high probability. The intuition underlying this condition is that λ_n must be sufficiently large to counteract the sampling noise, which aggregates at rate $\Theta(\sqrt{\frac{\log p}{n}})$.

To develop intuition for Theorems 3 and 4, we begin by stating certain special cases as corollaries, and discussing connections to previous work.

B. Some consequences for 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 [10], as well as Candes and Tao [4] has focused on the special case of the uniform Gaussian ensemble (i.e., $X_{ij} \sim N(0,1)$, i.i.d.). in the noiseless ($\sigma^2 = 0$) and underdetermined setting ($n \ll p$). These papers analyze the asymptotic behavior of the basic pursuit linear program (5), in particular when it succeeds in recovering a sparse vector $\beta^* \in \mathbb{R}^p$ based on an n-vector $y = X\beta^*$ of noiseless observations. The basic result is that exists a function $f : (0,1) \to (0,1)$ such that for any vector $\beta^* \in \mathbb{R}^p$ with at most $k = \alpha p$ non-zeros for some $\alpha \in (0,1)$, the basis pursuit LP recovers β^* using $n = f(\alpha)p$ observations, with high probability over choice of the random design matrix $X \in \mathbb{R}^{n\times p}$ from the uniform Gaussian ensemble,

Suppose that we apply our results to analyze support recovery for the noisy version of this problem. For the uniform Gaussian ensemble, we have $\gamma = 1$, $C_{min} = C_{max} = 1$, and $\rho_\ell(I) = \rho_u(I) = 1$, so that the threshold constants are given by $\theta_\ell(I) = \theta_u(I) = 1$. Consequently, Theorems 3 and 4 provide a sharp threshold for the behavior of the Lasso, in that failure/success is entirely determined by whether or not the inequality

$$\frac{n}{2k\log(p-k)} \;>\; 1 + \frac{\sigma^2}{\lambda_n^2 k} \qquad (36)$$

is satisfied or not. Consequently, we have the following corollary for design matrices from the standard Gaussian ensemble.

Corollary 2 (Standard Gaussian designs)**.** *(a) Suppose that* $n = \nu p$ *for some* $\nu \in (0,1)$*. Under this scaling, the Lasso can only recover vectors* β^* *with support* $k \leq (1+o(1))\;\frac{\nu p}{2\log p}$*. It fails with probability converging to one for any vector* $\beta^* \in \mathbb{R}^p$ *with* $k = \Theta(p)$ *non-zero elements.*

(b) Suppose that $k = \alpha p$ *for some* $\alpha \in (0,1)$*. Then the Lasso* (3) *requires a sample size* $n > 2\alpha p \log[(1-\alpha)\,p]$ *in order to obtain exact recovery with probability converging to one for large problems.*

This corollary establishes 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., $k = \nu p$ as in the work [10], [4]) to a sublinear fraction (i.e., $k = \mathcal{O}(\frac{p}{\log p})$, as in Corollary 2).

Interestingly, information-theoretic analysis of this sparsity recovery problem [34], [36] shows that the optimal decoder—namely, an exponential-time algorithm that can search exhaustively over all $\binom{p}{k}$ subsets—has a fundamentally different scaling than the Lasso in some regimes. In particular, if the minimum value $\beta_{\min} = \Omega(\sqrt{\frac{\log k}{k}})$, then the optimal decoder requires only a linear fraction of observations ($n = \Theta(p)$) to recover signals with linear fraction sparsity ($k = \Theta(p)$). This behavior, which contrasts dramatically with the Lasso threshold given in Theorems 3 and 4, raises an interesting question as to whether there exist computationally tractable methods for achieving this scaling.

C. Oracle properties

An interesting question raised by a reviewer is whether the Lasso solution $\widehat{\beta}$ has an "oracle property" [16]. More specifically, consider the oracle that knows a priori the support S of β^*, and then computes the optimal estimate $\breve{\beta}_S$, in the sense of minimizing the expected ℓ_2 error $\mathbb{E}\|\breve{\beta}_S - \beta^*_S\|^2$. A natural question is whether the error $\mathbb{E}\|\widehat{\beta}_S - \beta^*_S\|_2^2$ associated with the Lasso estimate $\widehat{\beta}_S$ has the same scaling as this oracle error. Since the Lasso involves shrinkage (essentially, in order to exclude the variables in S^c), one might expect that the estimate $\widehat{\beta}_S$ would be biased, thereby increasing the mean-squared error relative to an oracle. The following corollary, proved in Appendix E, confirms this intuition:

Corollary 3. *Assume that the covariance matrix satisfies conditions* (26a), (26b)*, and* (26c)*. Under the scaling of Theorem 3, there is a constant* $c_1 > 0$ *such that the Lasso* ℓ_2*-error satisfies*

$$\mathbb{P}\big[\|\widehat{\beta}_S - \beta^*_S\|_2^2 \geq c_1 \lambda_n^2 k\big] \quad = \quad 1 - o(1).$$

Remark: Since $\lambda_n^2 = \Omega(\frac{\log p}{n})$, the ℓ_2-error of the Lasso exceeds the $\mathcal{O}(k/n)$ ℓ_2-error that can be achieved by ordinary least squares restricted to the correct subset S. Consequently, Corollary 3 shows that the one-step Lasso procedure does not have the oracle property, in that its ℓ_2-error is larger than what could be achieved by a method that knew a priori the correct subset. Of course, assuming that the Lasso correctly estimates the subset S, one could estimate β^*_S at oracle rates in ℓ_2-norm by restricting to S; however, the Lasso does not achieve this optimal scaling in a one-step manner.

D. Comparison to information-theoretic limits

A related question is whether some *other algorithm*—whether or not is is computationally feasible—could perform consistent subset selection for scalings (n,p,k) where the Lasso fails. More specifically, Theorem 3 shows that the Lasso can achieve consistent subset selection for sample sizes n that scale with the problem size p and sparsity k as $n = \Omega(k\log(p-k))$. Could an optimal algorithm—namely, one that searches exhaustively over all $\binom{p}{k}$ subsets—-recover the correct one with substantially fewer observations? Since the initial posting of this work [33], our follow-up work [34], [36] has investigated the information-theoretic limitations of the subset selection problem when X is drawn from the standard Gaussian ensemble. This body of work shows that for sub-linear sparsity (i.e., $k/p \to 0$), *any algorithm* requires at least $\Omega(k\log(p-k))$ samples to perform consistent subset selection. Thus, up to constant factors, the Lasso performs as well as any algorithm for sublinear subset selection in the standard Gaussian ensemble. As discussed following Corollary 2, for the regime of linear sparsity ($k/p = \Theta(1)$) and suitably large values of the minimum value $\beta_{\min}$, the Lasso does *not* always achieve the information-theoretically optimal scaling.

V. Proof of Theorem 3

We begin with the achievability result for random Gaussian designs (Theorem 3). As with the proof of Theorem 1, the proof is based on the PDW method, and in particular consists of verifying the strict dual feasibility check in Step 3, and the sign consistency check in Step 4.

Before proceeding, we note that since $\frac{k}{n} = o(1)$ under the scaling of Theorem 3, the random Gaussian matrix X_S has rank k with probability one, whence the matrix $X_S^T X_S$ is invertible with probability one. Accordingly, the conditions of Lemmas 2 and Lemma 3 are applicable.

A. Verifying strict dual feasibility

Recall the definition (28) of the conditional covariance matrix $\Sigma_{S^c|S}$. We begin by conditioning on X_S: since for each $j \in S^c$, the vector $X_j \in \mathbb{R}^n$ is zero-mean Gaussian (and possibly correlated with X_S), we can decompose it into a linear prediction plus prediction error as

$$X_j^T \quad = \quad \Sigma_{jS}(\Sigma_{SS})^{-1} X_S^T + E_j^T,$$

where the elements of the prediction error vector $E_j \in \mathbb{R}^n$ are i.i.d., with $E_{ij} \sim N(0, [\Sigma_{S^c|S}]_{jj})$. Consequently, conditioning on X_S and using the definition (10) of Z_j, we have $Z_j = A_j + B_j$, where

$$A_j \quad := \quad E_j^T \big\{ X_S (X_S^T X_S)^{-1} \check{z}_S + \Pi_{X_S^\perp}\big(\frac{w}{\lambda_n n}\big)\big\}, \text{ and} \tag{37a}$$

$$B_j \quad := \quad \Sigma_{jS}(\Sigma_{SS})^{-1} \check{z}_S. \tag{37b}$$

By the mutual incoherence condition (26a), we have

$$\max_{j \in S^c} |B_j| \quad \leq \quad (1-\gamma). \tag{38}$$

Conditioned on X_S and w, the vector E_j does not depend on the subgradient vector $\check{z}_S$; this subgradient, determined in Step 2 of the PDW method, is a function only of X_S and w, since it is obtained from the solution of the restricted Lasso program (9).

Since $\mathrm{var}(E_{ij}) = [\Sigma_{S^c|S}]_{jj} \leq \rho_u(\Sigma_{S^c|S})$, conditioned on X_S and w, the quantity A_j is zero-mean Gaussian with variance at most

$$\begin{aligned} \mathrm{var}(A_j) \quad &\leq \quad \rho_u \big\| X_S (X_S^T X_S)^{-1} \check{z}_S + \Pi_{X_S^\perp}(\frac{w}{\lambda_n n}) \|_2^2 \\ &= \quad \rho_u \underbrace{\left\{ \frac{1}{n} \check{z}_S^T (\frac{X_S^T X_S}{n})^{-1} \check{z}_S + \big\| \Pi_{X_S^\perp}(\frac{w}{\lambda_n n}) \big\|_2^2 \right\}}_{M_n}, \end{aligned}$$

where we have used the Pythagorean identity, and introduced the shorthand $\rho_u = \rho_u(\Sigma_{S^c|S})$. The following lemma, proved in Appendix F, controls the random scaling M_n of this variance bound:

Lemma 4. *For any $\epsilon \in (0, 1/2)$, define the event $\overline{\mathcal{T}}(\epsilon) = \{M_n > \overline{M}_n(\epsilon)\}$, where*

$$\overline{M}_n(\epsilon) := \big(1 + \max\{\epsilon, \frac{8}{C_{min}} \sqrt{\frac{k}{n}}\}\big) \, \big(\frac{k}{C_{min}\, n} + \frac{\sigma^2}{\lambda_n^2 n}\big). \tag{39}$$

Then $\mathbb{P}\big[\overline{\mathcal{T}}(\epsilon)\big] \leq 4 \exp(-c_1 \min\{n\epsilon^2, k\})$ for some $c_1 > 0$.

We exploit this lemma by conditioning on $\overline{\mathcal{T}}(\epsilon)$ and its complement, thereby that $\mathbb{P}[\max_{j \in S^c} |A_j| \geq \gamma]$ is upper bounded by

$$\mathbb{P}\big[\max_{j \in S^c} |A_j| \; \geq \; \gamma \;\; | \;\; \overline{\mathcal{T}}^c(\epsilon)\big] + 4 \exp(-c_1 \min\{n\epsilon^2, k\}).$$

Conditioned on $\overline{\mathcal{T}}^c(\epsilon)$, the variance of A_j is at most $\rho_u \overline{M}_n(\epsilon)$, so that by standard Gaussian tail bounds, we obtain the upper bound

$$\mathbb{P}[\max_{j \in S^c} |A_j| \geq \gamma \mid \overline{\mathcal{T}}^c(\epsilon)] \quad \leq \quad 2(p-k) \; \exp\big(-\frac{\gamma^2}{2\rho_u \overline{M}_n(\epsilon)}\big).$$

Since the assumptions of Theorem 3 ensure that $k/n = o(1)$ and $1/(\lambda_n^2 n) = o(1)$, we are guaranteed that $\overline{M}_n(\epsilon) = o(1)$. Therefore, the exponential term is decaying in our tail bound; we need the decay rate to dominate the $(p-k)$ term from the union bound. Using the definition (39) and following some algebra, we find that it is sufficient[5] to have

$$\frac{n}{1+\epsilon} \;\; > \;\; \frac{2\,\rho_u}{C_{min}\,\gamma^2} k \log(p-k) \left\{ 1 + \frac{\sigma^2 C_{min}}{\lambda_n^2 k} \right\}.$$

Thus, we have established the sufficiency of the lower bound (32) given in the theorem statement.

Now let us verify the sufficiency of the alternative bound (34). Using the definition (29b) of θ_u, and the given form (31) of λ_n, it is equivalent to have

$$\begin{aligned} \frac{n}{1+\epsilon} \quad &> \quad 2\theta_u(\Sigma) k \log(p-k) + \frac{2\rho_u \sigma^2}{\gamma^2} \frac{\log(p-k)}{\frac{\phi_p \rho_u}{\gamma^2} \frac{2\sigma^2 \log p}{n}} \\ &= \quad 2\theta_u(\Sigma) k \log(p-k) + \frac{n}{\phi_p} \frac{\log(p-k)}{\log p}. \end{aligned}$$

or after further manipulation, to have

$$\frac{n}{2k \log(p-k)} \, f(\epsilon, \phi_p) \quad > \quad \frac{\theta_u(\Sigma)}{1 - \frac{1}{\phi_p}}. \tag{40}$$

where $f(\epsilon, \phi_p) = \frac{(1+\epsilon)^{-1} - \frac{1}{\phi_p}}{1 - \frac{1}{\phi_p}}$. We note that for any fixed $\epsilon \in (0, 1/2)$, the function f is increasing for $\phi_p \in [2, \infty)$. Therefore, we have

$$f(\epsilon, \phi_p) \; \geq \; f(\epsilon, 2) \; = \; \frac{2}{1+\epsilon} - 1 \; = \; \frac{1-\epsilon}{1+\epsilon}.$$

Recall the lower bound (34) on n, specified by some fixed $\delta' > 0$. By choosing $\epsilon \in (0, 1/2)$ sufficiently small so that $\frac{1-\epsilon}{1+\epsilon} > (1+\delta')^{-1}$, the condition (34) implies that the condition (40) holds.

B. Sign consistency and ℓ_∞ bounds

We have established that with high probability under the conditions of Theorem 3, the Lasso has a unique solution $\widehat{\beta}$ with support $S(\widehat{\beta}) \subseteq S(\beta^*)$. We now turn to establishing the sign consistency and ℓ_∞ bounds. From its definition (11) and applying triangle inequality, the random variable $\max_{i \in S} |\Delta_i|$ is upper bounded by

$$\underbrace{\lambda_n \| \big(\frac{1}{n} X_S^T X_S\big)^{-1} \mathrm{sgn}(\beta_S^*) \|_\infty}_{F_1} \quad + \quad \underbrace{\| (\frac{1}{n} X_S^T X_S)^{-1} \frac{1}{n} X_S^T w \|_\infty}_{F_2}$$

In order to analyze the first term, we require the following lemma.

Lemma 5. *Consider a fixed non-zero vector $z \in \mathbb{R}^k$ and a random matrix $W \in \mathbb{R}^{n \times k}$ with i.i.d. elements $W_{ij} \sim N(0,1)$. Under the scaling $n = \Omega(k \log(p-k))$, there are positive constants c_1 and c_2 such that for all $t > 0$:*

$$\begin{aligned} &\mathbb{P}\big[\| [(\frac{1}{n} W^T W)^{-1} - I_{k \times k}] \, z \, \|_\infty \geq c_1 \|z\|_\infty\big] \\ &\qquad\qquad \leq 4 \exp(-c_2 \min\{k, \log(p-k)\}). \end{aligned}$$

Using this lemma, we can bound F_1 as follows. By triangle inequality, we have the upper bound $\frac{F_1}{\lambda_n} \leq G_1 + G_2$, where

$$G_1 \quad := \quad \| \big(\Sigma_{SS}\big)^{-1} \mathrm{sgn}(\beta_S^*) \|_\infty,$$

and

$$G_2 \quad := \quad \| \big[(\frac{1}{n} X_S^T X_S)^{-1} - \big(\Sigma_{SS}\big)^{-1} \big] \, \mathrm{sgn}(\beta_S^*) \|_\infty.$$

The first term is deterministic, and bounded as $G_1 \leq (|\!|\!| (\Sigma_{SS})^{-1/2} |\!|\!|_\infty)^2$, so that it remains to bound G_2. By definition, we have $X_S = W_S \, (\Sigma_{SS})^{1/2}$, where $W_S \in \mathbb{R}^{n \times k}$ is a standard Gaussian random matrix. Consequently, we can

[5] Here we have used the fact that for any fixed $\epsilon > 0$, we have $8\sqrt{k/n} < \epsilon$ for n sufficiently large.

write

$$\begin{aligned} G_2 &= \|\Sigma_{SS}^{-1/2}\;[(\frac{1}{n}W_S^TW_S)^{-1} - I_{k\times k}]\Sigma_{SS}^{-1/2}\vec{b}\|_\infty \\ &\le |||\Sigma_{SS}^{-1/2}|||_\infty\;\|[(\frac{1}{n}W_S^TW_S)^{-1} - I_{k\times k}]\Sigma_{SS}^{-1/2}\vec{b}\|_\infty, \end{aligned}$$

where we have introduced the shorthand $\vec{b} = \text{sign}(\beta^*_S)$. Applying Lemma 5 with $z = \Sigma_{SS}^{-1/2}\vec{b}$, we obtain that

$$\mathbb{P}\big[G_2 > c_1|||\Sigma_{SS}^{-1/2}|||_\infty\|z\|_\infty\,\big] \\ \le 4\exp(-c_2\min\{k, \log(p-k)\}).$$

Note that since $\|\vec{b}\|_\infty = 1$, we have the upper bound $\|z\|_\infty \le |||\Sigma_{SS}^{-1/2}|||_\infty$. Putting together the pieces, we conclude that

$$\mathbb{P}\big[F_1 > c_3\lambda_n|||\Sigma_{SS}^{-1/2}|||^2_\infty\big] \le 4\exp(-c_2\min\{k, \log(p-k)\}). \tag{41}$$

Turning to the second term F_2, conditioned on X_S, the k-dimensional random vector $\widetilde{w} := (\frac{1}{n}X_S^TX_S)^{-1}\frac{1}{n}X_S^Tw$ is zero-mean Gaussian with variance at most $\widetilde{\sigma}^2_n(X) := \frac{\sigma^2}{n}|||(X_S^TX_S)^{-1}|||_2$. Define the event

$$\mathcal{T}(X_S) \;\;:=\;\; \Big\{\widetilde{\sigma}^2_n(X) \ge \frac{9\sigma^2}{nC_{min}}\Big\}.$$

By the bound (60) from Appendix K, we have $\mathbb{P}[\mathcal{T}(X_S)] \le 2\exp(-n/2)$. By the total probability rule, we have

$$\mathbb{P}[F_2 > t] \;\;\le\;\; \mathbb{P}[F_2 > t \mid \mathcal{T}^c(X_S)] + \mathbb{P}[\mathcal{T}(X_S)].$$

Conditioned on $\mathcal{T}^c(X_S)$, the random variable $\widetilde{w}$ is zero-mean Gaussian with variance at most $\frac{9\sigma^2}{nC_{min}}$ so that by Gaussian tail bounds, we have

$$\mathbb{P}[\|\widetilde{w}\|_\infty \ge t] \;\;\le\;\; 2k\,\exp\big(-\frac{C_{min}nt^2}{162\sigma^2}\big).$$

Setting $t = 20\sqrt{\frac{\sigma^2\log k}{C_{min}n}}$ yields that this probability vanishes at rate $2\exp(-c_1 n)$. Overall, we conclude that

$$\mathbb{P}\Bigg[F_2 \ge 20\sqrt{\frac{\sigma^2\log k}{C_{min}\,n}}\Bigg] \;\;\le\;\; 4\exp(-c_1 n). \tag{42}$$

Finally, combining bounds (41) and (42), we conclude that with probability greater than $1 - c_3'\exp(-c_2\log k)$, we have

$$\max_{i\in S}|\Delta_i| \;\;\le\;\; c_3\lambda_n|||\Sigma_{SS}^{-1/2}|||^2_\infty + 20\sqrt{\frac{\sigma^2\log k}{C_{min}\,n}} \;:= g(\lambda_n).$$

Consequently, we have shown that the candidate dual vector $\check{z}_S = \text{sign}(\beta^*_S)$ leads to a candidate primal solution $\check{\beta}_S$ such that

$$\max_i|\Delta_i| \;=\; \|\check{\beta}_S - \beta^*_S\|_\infty \;\le\; g(\lambda_n),$$

with high probability. As long as $\beta_{\min} > g(\lambda_n)$, the pair $(\check{\beta}_S, \check{z}_S)$ are primal-dual feasible; by Lemma 2, they are the unique Lasso solution, and show that it successfully recovers the signed support.

VI. Proof of Theorem 4

We establish the claim by showing that under the stated conditions, the random variable $\max_{j\in S^c} Z_j$ exceeds 1 with probability approaching one. By Lemmas 2(c) and 3(a), this event implies failure of the Lasso in recovering the support. From the proof of Theorem 3, recall the decomposition (37) $Z_j = A_j + B_j$. Using the bound (38) on the B_j terms, it suffices to show that $\max_{j\in S^c} A_j$ exceeds $(2-\gamma)$ with probability approaching one.

From Lemma 2, in order for the Lasso to achieve correct signed support recovery, we must have $\check{z}_S = \text{sgn}(\beta^*_S)$. Given this equality and under conditioning on (X_S, w), the vector A_{S^c} is zero-mean Gaussian with covariance matrix $\widetilde{M}_n\,\Sigma_{S^c|S}$, where the random scaling factor $\widetilde{M}_n$ has the form

$$\Big\{\frac{1}{n}\,\text{sign}(\beta^*_S)^T(\frac{X_S^TX_S}{n})^{-1}\,\text{sign}(\beta^*_S) + \big\|\Pi_{X_S^\perp}(\frac{w}{\lambda_n n})\big\|_2^2\Big\}.$$

The following lemma, proved in Appendix 6, provides control on this random scaling:

Lemma 6. *For any $\epsilon \in (0, 1/2)$, define the event $\underline{\mathcal{T}}(\epsilon) = \{\widetilde{M}_n > \underline{M}_n(\epsilon)\}$, where for some positive constant c_2*

$$\underline{M}_n(\epsilon) := \begin{cases} c_2\,\frac{k}{n}, \quad \textit{if } k/n = \Theta(1), \\ (1 - \max\{\epsilon, \frac{8}{C_{min}}\sqrt{\frac{k}{n}}\})\Big(\frac{k}{C_{max}\;n} + \frac{\sigma^2}{\lambda_n^2 n}\Big) \\ \qquad\qquad\qquad\qquad \textit{if } k/n = o(1). \end{cases}$$

Then $\mathbb{P}\big[\underline{\mathcal{T}}(\epsilon)\big] \le 4\exp(-c_1\min\{n\epsilon^2, k\})$ for some $c_1 > 0$.

Conditioning on the complement $\underline{\mathcal{T}}^c(\epsilon)$, we obtain

$$\mathbb{P}\big[\max_{j\in S^c} A_j > 2-\gamma\big] \\ \ge \mathbb{P}\big[\max_{j\in S^c} A_j > 2-\gamma \mid \underline{\mathcal{T}}^c(\epsilon)\big]\big\{1 - 2\exp(-c_1\min\{n\epsilon^2, k\})\big\}.$$

The remainder of our analysis studies the random variable $\max_{j\in S^c} A_j$ conditioned on $\underline{\mathcal{T}}^c(\epsilon)$. We first note that it suffices to show that $\mathbb{P}[\max_{j\in S^c}\widetilde{A}_j > 2-\gamma]$ goes to one, where the vector $\widetilde{A} \in \mathbb{R}^{p-k}$ is zero-mean Gaussian with covariance $\underline{M}_n(\epsilon)\Sigma_{S^c|S}$. Letting $e_i \in \mathbb{R}^{p-k}$ denote a unit vector with 1 in position i, observe that for each $i \ne j$, we have

$$\begin{aligned} \mathbb{E}[(\widetilde{A}_i - \widetilde{A}_j)^2] &= \underline{M}_n(\epsilon)\,(e_i - e_j)^T\Sigma_{S^c|S}(e_i - e_j) \\ &\ge 2\,\underline{M}_n(\epsilon)\,\rho_\ell(\Sigma_{S^c|S}), \end{aligned}$$

where we have used the definition (27) of ρ_ℓ. Consequently, if we let $\{\check{A}_j, j \in S^c\}$ be i.i.d. zero-mean Gaussians with variance $\underline{M}_n(\epsilon)\;\rho_\ell(\Sigma_{S^c|S})$, then we have established the lower bound

$$\mathbb{E}[(\widetilde{A}_i - \widetilde{A}_j)^2] \;\;\ge\;\; \mathbb{E}[(\check{A}_i - \check{A}_j)^2],$$

Therefore, the Sudakov-Fernique inequality [25] implies that the maximum over $\widetilde{A}$ dominates the maximum over $\check{A}$: more precisely, we have $\mathbb{E}[\max_{j\in S^c}\widetilde{A}_j] \ge \mathbb{E}[\max_{j\in S^c}\check{A}_j]$. The $\{\check{A}_j\}$ are i.i.d., so that by standard asymptotics of Gaussian extreme order statistics [25], for all $\nu > 0$, we have

$$\begin{aligned}\mathbb{E}[\max_{j\in S^c}\widetilde{A}_j] &\geq \mathbb{E}[\max_{j\in S^c}\breve{A}_j] \qquad (43)\\ &\geq \sqrt{(2-\nu)\,\underline{M}_n(\epsilon)\,\rho_\ell(\Sigma_{S^c|S})\log(p-k)},\end{aligned}$$

once $p-k$ is large enough.

We now claim that the random variable $\max_{j\in S^c}\widetilde{A}_j$ is sharply concentrated around its expectation.

Lemma 7. *For any $\eta > 0$, we have*

$$\mathbb{P}\Big[\big|\max_{j\in S^c}\widetilde{A}_j - \mathbb{E}[\max_{j\in S^c}\widetilde{A}_j]\big| > \eta\Big] \leq 2\exp\big(-\frac{\eta^2}{2\underline{M}_n(\epsilon)\rho_u}\big),$$

where $\rho_u = \rho_u(\Sigma_{S^c|S})$.

The proof, provided in Appendix I, makes use of concentration results for Lipschitz functions of Gaussian random vectors [25], [24].

Combining the lower bound (43) and the concentration statement from Lemma 7, for all $\nu, \eta, \epsilon > 0$, we have the lower bound

$$\max_{j\in S^c}\widetilde{A}_j \geq \sqrt{(2-\nu)\,\underline{M}_n(\epsilon)\,\rho_\ell(\Sigma_{S^c|S})\log(p-k)} - \eta \qquad (44)$$

with probability greater than $1 - 2\exp\big(-\frac{\eta^2}{2\underline{M}_n(\epsilon)\rho_u}\big)$. Consequently, it suffices to establish the bound

$$2\underline{M}_n(\epsilon)\,\rho_\ell(\Sigma_{S^c|S})\,\log(p-k) \geq \frac{2\,[(2-\gamma)+\eta]^2}{2-\nu}, \qquad (45)$$

using choices of η, ϵ for which $\frac{\eta^2}{\underline{M}_n(\epsilon)\,\rho_u} \to +\infty$ as $(n,p,k) \to +\infty$.

Case 1: If $\underline{M}_n(\epsilon) \to +\infty$ or $\underline{M}_n(\epsilon) = \Theta(1)$, then we may set $\eta^2 = \delta'\underline{M}_n(\epsilon)\log(p-k)$ for some $\delta' > 0$. If $\delta' > 0$ is fixed but chosen sufficiently close to zero (as a function of ν, ϵ and other constants), then from the lower bound (44), there is some constant $c_4 > 0$ such that

$$\mathbb{P}[\max_{j\in S^c}\widetilde{A}_j \geq c_4\sqrt{\log(p-k)})] \to 1.$$

Case 2: The other and more delicate possibility is that $\underline{M}_n(\epsilon) = o(1)$. In this case, we may choose any fixed $\eta > 0$, and have the guarantee that $\eta^2/\underline{M}_n(\epsilon) \to +\infty$. Note $\underline{M}_n(\epsilon) = o(1)$ is possible only if $k/n = o(1)$, so that the second line in the definition of $\underline{M}_n(\epsilon)$ from Lemma 6 applies. Moreover, for any fixed $\epsilon > 0$, we have $\frac{8}{C_{min}}\sqrt{\frac{k}{n}} \leq \epsilon$ once n is sufficiently large, so we include only the terms involving ϵ. Substituting this quantity into inequality (45) and performing some algebra, we find that it suffices to choose fixed $\nu, \eta > 0$ and $\epsilon \in (0, 1/2)$ such that

$$\begin{aligned}(2-\nu)\,(1-\epsilon)\,\big[\frac{1}{C_{max}} + \frac{\sigma^2}{\lambda_n^2 k}\big]\rho_\ell(\Sigma_{S^c|S})\frac{k\log(p-k)}{n}&\\ > [(2-\gamma)+\eta]^2,&\end{aligned}$$

or equivalently, such that

$$\begin{aligned}\frac{\rho_\ell(\Sigma_{S^c|S})}{C_{max}(2-\gamma)^2}\,\big[1 + \frac{C_{max}\sigma^2}{\lambda_n^2 k}\big]\frac{2k\log(p-k)}{n}&\\ > \frac{[(2-\gamma)+\eta]^2}{(2-\gamma)^2(1-\nu/2)\,(1-\epsilon)}.&\end{aligned}$$

Recall that under the assumptions of Theorem 4, the sample size is bounded above as $n < 2\theta_\ell(1 + \frac{C_{max}\sigma^2}{\lambda_n^2 k})(1-\delta)k\log(p-k)$ for some fixed $\delta > 0$, where $\theta_\ell = \frac{\rho_\ell(\Sigma_{S^c|S})}{C_{max}\,(2-\gamma)^2}$. Substituting in these relations, we find that find that after some algebraic manipulation, it suffices to choose $\epsilon \in (0, 1/2)$ and $\nu, \eta > 0$ such that

$$\frac{1}{1-\delta} > \frac{[(2-\gamma)+\eta]^2}{(2-\gamma)^2(1-\nu/2)\,(1-\epsilon)}.$$

Note that the left-hand side is strictly greater than 1. On the right-hand side, the quantity $\gamma \in (0,1]$ is the mutual incoherence constant, whereas ϵ, ν, η are parameters that can be chosen in $(0, 1/2)$. By choosing ϵ, ν, η to be strictly positive but arbitrarily close to 0, we can set the right-hand side arbitrarily close to 1, thereby satisfying the required inequality.

VII. Illustrative simulations

In this section, we provide some simulations to confirm the threshold behavior predicted by Theorems 3 and 4. We consider the following three types of sparsity indices:

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

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

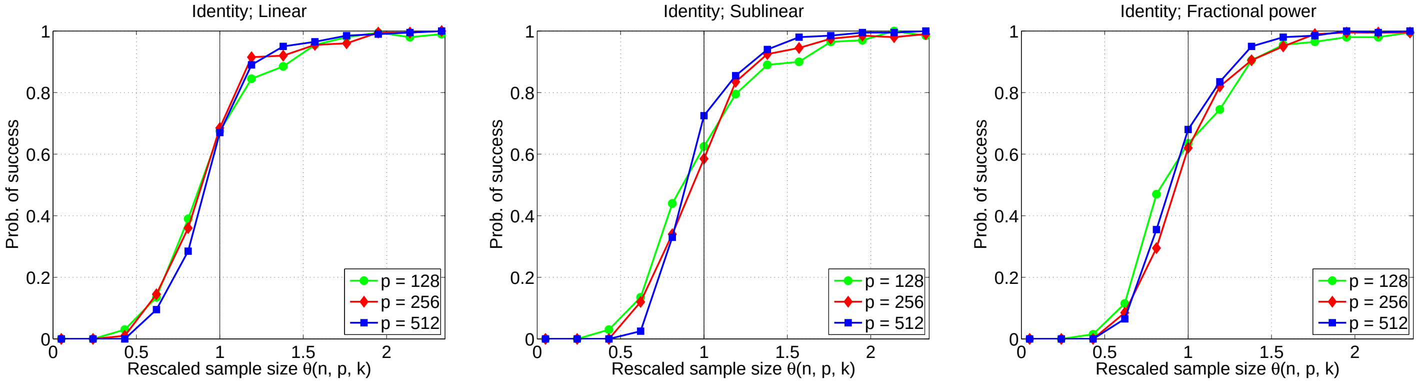
(c) *fractional power* sparsity, meaning that $k(p) = \lceil \alpha p^\delta \rceil$ for some $\alpha, \delta \in (0,1)$.

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

$$n = 2\,\theta\,k\log(p-k),$$

where 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 $\delta = 0.75$ for the fractional power ensemble. We specified the parameter vector β^* by choosing the subset S randomly, and for each $i \in S$ setting β_i^* equal to $+\beta_{\min}$ or $-\beta_{\min}$ with equal probability, and $\beta_j^* = 0$ for all indices $j \notin S$. For the results shown here, we fixed $\beta_{\min} = 0.50$, but have also experimented with decaying choices of the minimum value. In addition, we fixed the noise level $\sigma = 0.5$, and the regularization parameter $\lambda_n = \sqrt{\frac{2\sigma^2\,\log(k)\log(p-k))}{n}}$ in all cases. For this choice of λ_n, Theorem 4 predicts failure with high probability for sequences (n,p,k) such that failure for sequences such that

$$\frac{n}{2k\log(p-k)} < \theta_\ell(\Sigma),$$



(a) (b) (c)

Fig. 2. Plots of the rescaled sample size $\theta = n/[k \log(p-k)]$ versus the probability $\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)]$ of correct signed support recovery using 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: $k = \alpha p$. (b) Sublinear sparsity index $k = \alpha p / \log(\alpha p)$. (c) Fractional power sparsity index $k = \alpha p^\delta$ with $\delta = 0.75$. In all cases, the parameter $\alpha = 0.40$. The threshold in Lasso success probability occurs at $\theta^* = 1$, consistent with the sharp threshold predicted by Theorems 3 and 4.

whereas Theorem 3 predicts success with high probability for sequences such that

$$\frac{n}{2k\log(p-k)} \quad > \quad \theta_u(\Sigma).$$

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 threshold values are $\theta_u(I) = \theta_\ell(I) = 1$. Figure 2 plots the control parameter or rescaled sample size 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 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

$$\Sigma \quad = \quad \begin{bmatrix} 1 & \mu & \mu^2 & \cdots & \mu^{p-2} & \mu^{p-1} \\ \mu & 1 & \mu & \mu^2 & \cdots & \mu^{p-2} \\ \mu^2 & \mu & 1 & \mu & \cdots & \mu^{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu^{p-1} & \cdots & \mu^3 & \mu^2 & \mu & 1 \end{bmatrix}, \tag{46}$$

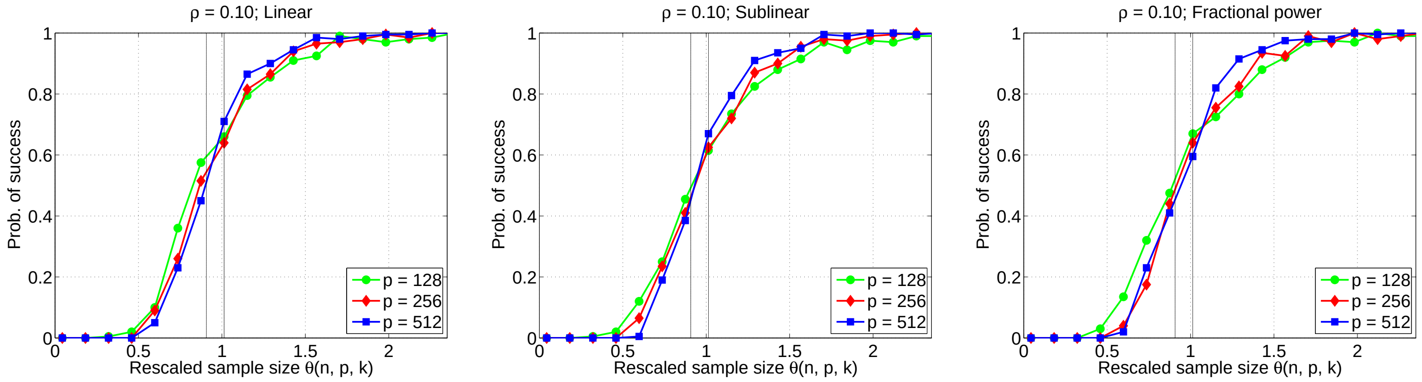
for some $\mu \in (-1, +1)$. The maximum and minimum eigenvalues (C_{min} and C_{max}) can be bounded using standard asymptotic results on Toeplitz matrix families [19].

Figure 3 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. The vertical lines to the left and right of $\theta = 1$ show the numerical values of the theoretical upper and lower bounds on the threshold—that is, $\theta_u(\Sigma)$ and $\theta_\ell(\Sigma)$, as defined in equation (29). Once again, these simulations show good agreement with the theoretical predictions.

VIII. Discussion

The problem of recovering the sparsity pattern of a high-dimensional vector β^* 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 ℓ_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. In addition, we provided some necessary conditions on the design and signal vector for support recovery. 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 $n = 2\,k\log(p-k)$ 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 [35] has established qualitatively similar results for the case of logistic regression, with application to model selection in binary



(a) (b) (c)

Fig. 3. Plots of the rescaled sample size $\theta = n/[k\log(p-k)]$ versus the probability $\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)]$ of correct signed support recovery using the Lasso for the Toeplitz family (46) of random design matrices 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: $k = \alpha p$. (b) Sublinear sparsity index $k = \alpha p/\log(\alpha p)$. (c) Fractional power sparsity index $k = \alpha p^\delta$ with $\delta = 0.75$. The vertical lines to the left and right of $\theta = 1$ show the theoretical upper and lower bounds $\theta_u(\Sigma)$ and $\theta_\ell(\Sigma)$, from equation (29).

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 [34] 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

The author would like to thank Peter Bickel, Alyson Fletcher, Noureddine El Karoui, Vivek Goyal, John Lafferty, Sahand Negahban, Dapo Omidiran, Larry Wasserman and Bin Yu for helpful comments on earlier drafts, as well as the anonymous reviewers for their constructive suggestions. This work was partially supported by an Alfred P. Sloan Foundation Fellowship and NSF Grants DMS-0605165, and CAREER-CCF-0545862.

Appendix

A. Sub-Gaussian variables and tail bounds

Parts of our analysis focus on noise vectors $w \in \mathbb{R}^n$ in the linear observation model (1) that have i.i.d. elements satisfying a sub-Gaussian tail condition.

Definition 1. A zero-mean random variable Z is *sub-Gaussian* if there exists a constant $\sigma > 0$ such that

$$\mathbb{E}[\exp(tZ)] \quad \leq \quad \exp(\sigma^2\, t^2/2) \qquad \text{for all } t \in \mathbb{R}. \quad (47)$$

By applying the Chernoff bound and optimizing the exponent, this upper bound (47) on the moment-generating function implies a two-sided tail bound of the form

$$\mathbb{P}[|Z| > z] \quad \leq \quad 2\exp\big(-\frac{z^2}{2\sigma^2}\big). \qquad (48)$$

Naturally, any zero-mean Gaussian variable with variance σ^2 satisfies the bounds (47) and (48). In addition to the Gaussian case, the class of sub-Gaussian variates includes any bounded random variable (e.g., Bernoulli, multinomial, uniform), any random variable with strictly log-concave density [2], [24], and any finite mixture of sub-Gaussian variables.

For future use, we also note the following useful property (Lemma 1.7, [2]): if $Z_1, \ldots, Z_n$ are independent and zero-mean sub-Gaussian variables with parameters $\sigma_1^2, \ldots, \sigma_n^2$, then

$$\sum_{i=1}^n Z_i \quad \text{is sub-Gaussian with parameter } \textstyle\sum_{i=1}^n \sigma_i^2. \qquad (49)$$

B. Proof of Lemma 1

From the equivalent constrained form (4), we see that the Lasso involves a continuous objective function over a compact set, and so by Weierstrass' theorem, the minimum is always achieved. By standard conditions for optimality in a convex program [20], a point $\widehat{\beta} \in \mathbb{R}^p$ is optimal for the regularized form of the Lasso (3) if and only if there exists a subgradient $\widehat{z} \in \partial\|\widehat{\beta}\|_1$ such that $\frac{1}{n}X^TX\widehat{\beta} - \frac{1}{n}X^Ty + \lambda\widehat{z} = 0$. Substituting in the observation model $y = X\beta^* + w$ and performing some algebra yields equation (8), thereby establishing Lemma 1(a). By standard duality theory [1], given the subgradient $\widehat{z} \in \mathbb{R}^p$, any optimal solution $\check{\beta} \in \mathbb{R}^p$ of the Lasso must satisfy the complementary slackness condition $\widehat{z}^T\check{\beta} = \|\check{\beta}\|_1$, which can hold only if $\check{\beta}_j = 0$ for all indices j such that $|\widehat{z}_j| < 1$, which establishes Lemma 1(b). Lastly, if $X^T_{S(\widehat{\beta})}X_{S(\widehat{\beta})}$ is strictly positive definite, then when restricted to vectors of the form $(\beta_{S(\widehat{\beta})}, 0)$, the Lasso program is strictly convex, and so its optimum is uniquely attained, as claimed in part (c).

C. Proof of Lemma 2

(a) Suppose that steps 1 through 3 of the PDW construction succeed. Then, we have demonstrated a pair of vectors $\check{\beta} = (\check{\beta}_S, 0) \in \mathbb{R}^p$ and $\check{z} \in \mathbb{R}^p$, such that $\check{z} \in \partial\|\check{\beta}\|_1$. It remains to check that these vectors satisfy the zero subgradient condition (8), so that $\check{\beta}$ is actually an optimal solution to the Lasso. Writing out this condition in block form yields

$$\frac{1}{n}\begin{bmatrix} X_S^T X_S & X_S^T X_{S^c} \\ X_{S^c}^T X_S & X_{S^c}^T X_{S^c} \end{bmatrix}\begin{bmatrix} \beta_S - \beta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} X_S^T \\ X_{S^c}^T \end{bmatrix}\begin{bmatrix} w_S \\ w_{S^c} \end{bmatrix} + \lambda \begin{bmatrix} z_S \\ z_{S^c} \end{bmatrix} \;=\; \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \qquad (50)$$

Since the pair $(\check{\beta}_S, \check{z}_S)$ was obtained by solving the restricted convex program (9), they must satisfy the top block of these equations. Secondly, the bottom block of equations must also be satisfied, since we used these sub-gradient condition to solve for $\check{z}_{S^c}$ in Step 3 of the PDW method. Lastly, the strict dual feasibility guaranteed in Step 3 implies uniqueness, using the assumed invertibility of $X_S^T X_S$, and Lemma 1(c), which completes the proof of Lemma 2(a).

(b) Suppose that in addition, the sign consistency condition in Step 4 is satisfied. Then since $\check{z}_S$ was chosen as an element of the subdifferential $\partial\|\check{\beta}_S\|_1$ in Step 2, we must have $\operatorname{sign}(\check{\beta}_S) = \operatorname{sign}(\beta_S^*)$, from which Lemma 2(b) follows.

(c) Turning to Lemma 2(c), it is equivalent to prove the following assertion: if there exists a Lasso solution $\widehat{\beta} \in \mathbb{R}^p$ with $\widehat{\beta}_{S^c} = 0$ and $\operatorname{sign}(\widehat{\beta}_S) = \operatorname{sign}(\beta_S^*)$, then the PDW method succeeds in producing a dual feasible vector $\check{z}$ with $\check{z}_S = \operatorname{sign}(\widehat{\beta}_S^*)$. Since $X_S^T X_S$ is invertible by assumption, the vector $\widehat{\beta}_S$ must be the unique optimal solution to the restricted program (9), so it will be found in Step 1 of the PDW method. Since $\operatorname{sign}(\widehat{\beta}_S) = \operatorname{sign}(\beta_S^*)$ by assumption, the vector $\widehat{z}_S = \operatorname{sign}(\widehat{\beta}_S^*)$ is only subgradient that can be chosen in Step 2. Since $(\widehat{\beta}_S, 0)$ is an optimal Lasso solution by assumption, then there must exist a dual feasible vector $\widehat{z}_{S^c}$ such that $(\operatorname{sign}(\widehat{\beta}_S), \widehat{z}_{S^c})$ satisfy the zero subgradient condition (8).

D. Proof of Lemma 3

The vectors $(\check{\beta}_S, \check{z}_S)$ determined in Steps 1 and 2 of the PDW must satisfy the top block of equation (50). Using the assumed invertibility of $X_S^T X_S$, we may solve for $\check{\beta}_S - \beta_S^*$ as follows:

$$\check{\beta}_S - \beta_S^* \;=\; \big(\frac{1}{n} X_S^T X_S\big)^{-1}\big[\frac{1}{n} X_S^T w - \lambda \check{z}_S\big]. \qquad (51)$$

Similarly, the vector $\check{z}_{S^c}$ determined in Step 3 of the PDW must satisfy the zero-subgradient conditions (50). Note that it enters only in the bottom block of equations. Consequently, we may solve for $\check{z}_{S^c}$ in terms of $\check{\beta}_S - \beta_S^*$ and $\check{z}_S$ as follows:

$$\widehat{z}_{S^c} \;=\; X_{S^c}^T\left[X_S\big(X_S^T X_S\big)^{-1}\check{z}_S + \Pi_{X_S^\perp}\big(\frac{w}{\lambda_n n}\big)\right]. \qquad (52)$$

The elements of this vector are the variables $\{Z_j\}$ defined in equation (10). Consequently, the claim of Lemma 3(a) follows.

Now suppose that the condition (13) holds. Note that the vector Δ from equation (11) is obtained by solving for $\check{\beta}_S$, as in equation (51), but assuming that $\check{z}_S = \operatorname{sign}(\beta_S^*)$. But if condition (13) holds, then we can conclude that the optimal solution $\check{\beta}_S$ to the restricted program (9) does satisfy $\operatorname{sign}(\check{\beta}_S) = \operatorname{sign}(\beta_S^*)$, so that $\check{z}_S = \operatorname{sign}(\beta_S^*)$ is indeed the only valid choice of subgradient vector. This certifies that the Lasso has a solution with the correct signed support, as claimed. Conversely, if the Lasso has a solution $\check{\beta}$ with the correct signed support, then $\Delta_i = \check{\beta}_i - \beta_i^*$, and condition (13) must hold, thus completing the proof of Lemma 3(b).

E. Proof of Corollary 3

As shown in the proof of Lemma 3 (in particular, equation (51)), when the Lasso correctly recovers the signed support set, its error is given by

$$\widehat{\beta}_S - \beta_S^* \;=\; \big(\frac{1}{n} X_S^T X_S\big)^{-1}\big[\frac{1}{n} X_S^T w - \lambda_n \operatorname{sgn}(\beta_S^*)\big].$$

By triangle inequality, the quantity $\|\widehat{\beta}_S - \beta_S^*\|_2$ is lower bounded by

$$\begin{aligned} &\lambda_n \|\big(\frac{1}{n} X_S^T X_S\big)^{-1}\operatorname{sgn}(\beta_S^*)\|_2 - \|\big(\frac{1}{n} X_S^T X_S\big)^{-1}\frac{1}{n} X_S^T w\|_2 \\ &\geq \underbrace{\lambda_n \sqrt{k}\,\Lambda_{min}\big[\big(\frac{1}{n} X_S^T X_S\big)^{-1}\big]}_{T_1} - \underbrace{\|\big(\frac{1}{n} X_S^T X_S\big)^{-1}\frac{1}{n} X_S^T w\|_2}_{T_2}. \end{aligned}$$

Since $X_S^T X_S / n$ is formed by Gaussian random matrices with $k \leq n$, then the bound (59) in Appendix K implies that $\Lambda_{min}[(\frac{1}{n} X_S^T X_S)^{-1}] \geq \frac{1}{9C_{max}}$ with probability at least $1 - 2\exp(-n/2)$. Consequently, we have

$$\mathbb{P}[T_1 \leq \frac{\lambda_n \sqrt{k}}{9C_{max}}] \;\leq\; 2\exp(-n).$$

As for the second term, conditioned on X_S, the quantity $\big(\frac{1}{n} X_S^T X_S\big)^{-1}\frac{1}{n} X_S^T w$ is zero-mean Gaussian with covariance $\frac{\sigma^2}{n}(X_S^T X_S / n)^{-1}$. Letting $\widetilde{w} \in \mathbb{R}^k$ be a standard Gaussian vector, we can re-express T_2 as

$$\begin{aligned} T_2^2 \;&=\; \widetilde{w}^T \frac{\sigma^2}{n}(X_S^T X_S / n)^{-1}\,\widetilde{w} \\ &\leq\; \frac{\sigma^2\,\|\widetilde{w}\|_2^2}{n} |\!|\!|(X_S^T X_S / n)^{-1}|\!|\!|_2 \end{aligned}$$

Again applying the bound (60) from Appendix K, we have

$$\mathbb{P}[|\!|\!|(X_S^T X_S)^{-1}/n|\!|\!|_2 \geq \frac{9}{C_{min}}] \;\leq\; 2\exp(-n).$$

By χ^2-concentration, we have $\mathbb{P}[\|\widetilde{w}\|_2^2 \geq 2k] \leq 2\exp(-c_2 k)$. Putting together the pieces, we conclude that the second term satisfies $\mathbb{P}[T_2 \geq \sqrt{\frac{18\sigma^2 k}{C_{min} n}}] \leq 4[\exp(-c_1 k)]$.

Consequently, with high probability, we have the lower bound

$$\begin{aligned} \|\widehat{\beta}_S - \beta_S^*\|_2 \;&\geq\; \frac{\lambda_n \sqrt{k}}{9C_{max}} - \sqrt{\frac{18\sigma^2 k}{C_{min} n}} \\ &=\; \frac{\lambda_n \sqrt{k}}{9C_{max}}\big[1 - \frac{c_3}{\lambda_n \sqrt{n}}\big] \\ &\geq\; c_4 \lambda_n \sqrt{k}, \end{aligned}$$

since $\lambda_n \sqrt{n} \to +\infty$.

F. Proof of Lemma 4

Since $\Pi_{X_S^\perp}$ is an orthogonal projection matrix, we have

$$\|\Pi_{X_S^\perp}(\frac{w}{\lambda_n n})\|_2^2 \quad \leq \quad \frac{1}{\lambda_n^2 n}\,\frac{\|w\|_2^2}{n}.$$

Noting that $\|w\|_2^2/\sigma^2$ is χ^2 with n degrees of freedom, by the bound (54a), we have

$$\mathbb{P}\Big[\|\Pi_{X_S^\perp}(\frac{w}{\lambda_n n})\|_2^2 \geq (1+\epsilon)\frac{\sigma^2}{\lambda_n^2 n}\Big] \quad \leq \quad 2\exp(-\frac{3n\epsilon^2}{16}).$$

Turning to the first term defining M_n, by applying Lemma 9) from Appendix K—more specifically, in the form (58a)—we obtain

$$\begin{aligned}\frac{1}{n}\check{z}_S^T(\frac{X_S^TX_S}{n})^{-1}\check{z}_S \quad &\leq \quad \big(1+\frac{8}{C_{min}}\sqrt{\frac{k}{n}}\big)\,\frac{\|\check{z}_S\|_2^2}{nC_{min}} \\ &\leq \quad \big(1+\frac{8}{C_{min}}\sqrt{\frac{k}{n}}\big)\,\frac{k}{nC_{min}},\end{aligned}$$

with probability greater than $1-2\exp(-k/2)$, which completes the proof.

G. Proof of Lemma 5

We begin by diagonalizing the random matrix $(W^TW/n)^{-1}$, writing $(W^TW/n)^{-1} - I_{k\times k} = U^TDU$, where D is diagonal, and U is unitary. Since the distribution of W is invariant to rotations, the matrices D and U are independent. Since $|\!|\!|D|\!|\!|_2 = |\!|\!|(W^TW/n)^{-1} - I_{k\times k}|\!|\!|_2$, the random matrix bound (58a) from Appendix K implies that

$$\mathbb{P}[|\!|\!|D|\!|\!|_2 > 8\sqrt{k/n}] \quad \leq \quad 2\exp(-k/2).$$

We condition on the event $\{|\!|\!|D|\!|\!|_2 < 8\sqrt{k/n}$ throughout the remainder of the analysis.

For a fixed vector $z\in\mathbb{R}^k$, we define, for each $i=1,\ldots,k$, the random variable

$$V_i \quad = \quad e_i^TU^TDUz \; = \; z_iu_i^TDu_i + u_i^TD\big[\sum_{\ell\neq i} z_\ell u_\ell\big],$$

where u_j is the j^{th} column of the unitary matrix U. Observe that the lemma statement concerns the random variable $\max_i |V_i|$. Since the $\{V_i\}$ are identically distributed, it suffices to obtain an exponential tail bound on $\{V_1\geq t\}$.

Under our conditioned event on $|\!|\!|D|\!|\!|_2$, we have

$$|V_1| \quad \leq \quad 8\sqrt{k/n}\,|z_1| + u_1^TD\big[\sum_{\ell=2}^k z_\ell u_\ell\big]. \tag{53}$$

Consequently, it suffices to establish a sharp tail bound on the second term. Conditioned on D and the vector $g := \sum_{\ell=2}^k z_\ell u_\ell$, the random vector $u_1\in\mathbb{R}^k$ is uniformly distributed over a sphere in $k-1$ dimensions.[6] Now consider the function $F(u_1) := u_1^TDg$; we claim that it is Lipschitz (with respect to the Euclidean norm) with constant at most $8\sqrt{k/n}\sqrt{k-1}\,\|z\|_\infty$. Indeed, given any pair of vectors $u_1, u_1'\in\mathbb{R}^k$, we have

$$\begin{aligned}|F(u_1)-F(u_1')| \quad &= \quad |(u_1-u_1')^TDg| \\ &\leq \quad \|u_1-u_1'\|_2|\!|\!|D|\!|\!|_2\,\|g\|_2 \\ &\leq \quad 8\sqrt{k/n}\sqrt{\sum_{\ell=2}^k z_\ell^2}\;\|u_1-u_1'\|_2, \\ &= \quad 8\sqrt{k/n}\,\sqrt{k-1}\,\|z\|_\infty\,\|u_1-u_1'\|_2,\end{aligned}$$

where we have used the fact that $\|g\|_2 = \sqrt{\sum_{\ell=2}^k z_\ell^2}$, by the orthonormality of the $\{u_\ell\}$ vectors.

Since $\mathbb{E}[F(u_1)] = 0$, by concentration of measure for Lipschitz functions on the sphere [24], for all $t>0$, we have

$$\begin{aligned}\mathbb{P}[|F(u_1)| > t\|z\|_\infty] \quad &\leq \quad 2\exp\Big(-c_1(k-1)\frac{t^2}{128\frac{k}{n}(k-1)}\Big) \\ &= \quad 2\exp\Big(-c_1\frac{nt^2}{128k}\Big)\end{aligned}$$

Taking union bound, we have

$$\mathbb{P}[\max_{i=1,\ldots,k}|F(u_i)| > t\|z\|_\infty] \quad \leq \quad 2\exp\Big(-c_1\frac{nt^2}{128k} + \log k\Big).$$

Since $\log(p-k) > \log k$, if we set $t = \frac{256k\log(p-k)}{c_1 n}$ then this probability vanishes at rate $2\exp(-c_5\log(p-k))$. But since $n = \Omega(k\log(p-k))$ by assumption, the quantity t is order one, so that the claim follows.

H. Proof of Lemma 6

We begin by proving part (a), which requires only that $k\leq n$ (and not that $k/n = o(1)$). Using only the first term defining $\widetilde{M}_n$, we have

$$\begin{aligned}\widetilde{M}_n \quad &\geq \quad \frac{1}{n}\operatorname{sign}(\beta_S^*)^T(\frac{X_S^TX_S}{n})^{-1}\operatorname{sign}(\beta_S^*) \\ &\geq \quad \frac{k}{n}\,\frac{1}{|\!|\!|X_S^TX_S/n|\!|\!|_2} \\ &\geq \quad \frac{k}{n}\frac{1}{9C_{max}},\end{aligned}$$

where the final bound holds with probability at least $1-2\exp(-n/2)$, using equation (59) from Appendix K.

For part (b), we assume that $k/n = o(1)$. In this case, can apply the concentration bound (58b) from Appendix K to conclude that there are positive constants c_1, c_2 such that

$$\begin{aligned}&\frac{1}{n}\operatorname{sign}(\beta_S^*)^T(\frac{X_S^TX_S}{n})^{-1}\operatorname{sign}(\beta_S^*) \\ &\qquad\qquad \geq \frac{1}{C_{max}}\frac{k}{n}\big(1-\frac{8}{C_{min}}\sqrt{\frac{k}{n}}\big),\end{aligned}$$

with probability greater than $1-2\exp(-k/2)$.

Turning to the second term in $\widetilde{M}_n$, since $\Pi_{X_S^\perp}$ is an orthogonal projection matrix with rank $(n-k)$ and $w\sim N(0,\sigma^2 I)$ is multivariate Gaussian, the variable $\|\Pi_{X_S^\perp}(w)\|_2^2/\sigma^2$ is χ^2 with $d=n-k$ degrees of freedom. Using the tail bound (54b), for

[6] One dimension is lost since u_1 must be orthogonal to $g\in\mathbb{R}^k$.

any $\epsilon \in (0,1/2)$, we have

$$\mathbb{P}\Big[\|\Pi_{X_S^\perp}(\frac{w}{\lambda_n n})\|_2^2 \leq (1-\frac{k}{n})\,(1-\epsilon)\frac{\sigma^2}{\lambda_n^2 n}\Big] \leq 2\exp(-\frac{(n-k)\,\epsilon^2}{4}).$$

Overall, we conclude that with probability greater than $1-4\exp(-c_1\min\{n\epsilon^2, k\})$, we have

$$\widetilde{M}_n \quad \geq \quad (1-\max\{\epsilon, \frac{8}{C_{min}}\sqrt{\frac{k}{n}}\})\Big(\frac{k}{C_{max}\; n}+\frac{\sigma^2}{\lambda_n^2 n}\Big),$$

as claimed.

I. Proof of Lemma 7

Consider the function $f : \mathbb{R}^{p-k} \rightarrow \mathbb{R}$ given by

$$f(u) \; := \; \sqrt{\underline{M}_n(\epsilon)}\; \max_{j \in S^c} \big[e_j^T \sqrt{\Sigma_{S^c|S}}\; u\big],$$

where e_j denotes the unit vector with 1 in position j, and $\sqrt{\Sigma_{S^c|S}}$ is the symmetric matrix square root. By construction, for a Gaussian random vector $u \sim N(0, I)$, we have $f(u) \overset{d}{=} \max_{j \in S^c} \widetilde{A}_j$. We now bound the Lipschitz constant of f. For each $j = 1, \ldots, p-k$ and pairs of vectors $u, v \in \mathbb{R}^{p-k}$, we have

$$\begin{aligned}|f(u)-f(v)| \quad & \leq \quad \sqrt{\underline{M}_n(\epsilon)}\max_{j\in S^c}|e_j^T\sqrt{\Sigma_{S|S^c}}(u-v)| \\ & \overset{(a)}{\leq} \quad \sqrt{\underline{M}_n(\epsilon)}\;\max_{j\in S^c}\big\|e_j^T\sqrt{\Sigma_{S^c|S}}\big\|_2\;\|u-v\|_2 \\ & \overset{(b)}{\leq} \quad \sqrt{\underline{M}_n(\epsilon)\;\rho_u(\Sigma_{S^c|S})}\;\|u-v\|_2,\end{aligned}$$

where inequality (a) follows by Cauchy-Schwartz, and inequality (b) follows since

$$\|e_j^T\sqrt{\Sigma_{S^c|S}}\|_2^2 = e_j^T(\Sigma_{S^c|S})e_j \leq \rho_u(\Sigma_{S^c|S}),$$

using the definition (27) of ρ_u. Therefore, by Gaussian concentration of measure for Lipschitz functions [24], we conclude that for any $\eta > 0$, it holds that

$$\mathbb{P}[|\max_{j\in S^c}\widetilde{A}_j - \mathbb{E}[\max_{j\in S^c}\widetilde{A}_j]| \geq \eta] \leq \; 2\exp\big(-\frac{\eta^2}{2\underline{M}_n(\epsilon)\rho_u(\Sigma_{S^c|S})}\big),$$

as claimed.

J. Tail bounds for χ^2-variates

Given a centralized χ^2-variate X with d degrees of freedom, then for all $t \in (0, 1/2)$, we have

$$\mathbb{P}\big[X \geq d\,(1+t)\big] \quad \leq \quad \exp\big(-\frac{3}{16}\,d\,t^2\big), \quad \text{and} \qquad \text{(54a)}$$

$$\mathbb{P}\big[X \leq (1-t)d\big] \quad \leq \quad \exp(-\frac{1}{4}dt^2). \qquad \text{(54b)}$$

The bound (54a) is taken from Johnstone [21], whereas the bound (54b) follows from Laurent and Massart [23].

K. Spectral norms of random matrices

Here we collect some useful results about concentration of spectral norms and eigenvalues of Gaussian random matrices. We begin with the following basic lemma [7]:

Lemma 8. *For $k \leq n$, let $U \in \mathbb{R}^{n\times k}$ be a random matrix from the standard Gaussian ensemble (i.e., $U_{ij} \sim N(0,1)$, i.i.d.). Then for all $t > 0$, we have*

$$\mathbb{P}\big[\|\frac{1}{n}U^TU - I_{k\times k}\|_2 \geq \delta(n,k,t)\big] \leq 2\exp(-nt^2/2), \qquad (55)$$

where $\delta(n,k,t) := 2(\sqrt{\frac{k}{n}}+t) + (\sqrt{\frac{k}{n}}+t)^2$.

This result can be adapted easily to random matrices X drawn from more general Gaussian ensembles. In particular, for a positive definite matrix $\Lambda \in \mathbb{R}^{k\times k}$, setting $X = U\sqrt{\Lambda}$ yields $n \times k$ matrix with i.i.d. rows, $X_i \sim N(0, \Lambda)$.

Lemma 9. *For $k \leq n$, let $X \in \mathbb{R}^{n\times k}$ have i.i.d. rows $X_i \sim N(0,\Lambda)$.*
(a) If the covariance matrix Λ has maximum eigenvalue $C_{max} < +\infty$, then for all $t > 0$,

$$\mathbb{P}\Big[\|\frac{1}{n}X^TX - \Lambda\|_2 \geq C_{max}\delta(n,k,t)\Big] \quad \leq \quad 2\exp(-nt^2/2). \qquad (56)$$

(b) If the covariance matrix Λ has minimum eigenvalue $C_{min} > 0$, then for all $t > 0$,

$$\mathbb{P}\Big[\|(\frac{X^TX}{n})^{-1} - \Lambda^{-1}\|_2 \geq \frac{\delta(n,k,t)}{C_{min}}\Big] \leq 2\exp(-nt^2/2). \qquad (57)$$

Proof: (a) Letting $\sqrt{\Lambda}$ denote the symmetric matrix square root, we can write $X = U\sqrt{\Lambda}$ where $U \in \mathbb{R}^{n\times k}$ is standard Gaussian ($U_{ij} \sim N(0,1)$, i.i.d.). Thus, we have

$$\|n^{-1}X^TX - \Lambda\|_2 \; = \; \|\sqrt{\Lambda}[n^{-1}U^TU - I]\sqrt{\Lambda}\|_2,$$

which is upper bounded by $C_{max}\,\|n^{-1}\;U^TU - I\|_2$, so that the claim (56) follows from the basic bound (55).

(b) Letting $U \in \mathbb{R}^{n\times k}$ denote a standard Gaussian matrix, we write

$$\begin{aligned}\|(\frac{X^TX}{n})^{-1} - \Lambda^{-1}\|_2 \quad & = \quad \|\Lambda^{-1/2}\big[(\frac{U^TU}{n})^{-1} - I_{k\times k}\big]\Lambda^{-1/2}\|_2 \\ & \leq \quad \|(U^TU/n)^{-1} - I_{k\times k}\|_2\frac{1}{C_{min}},\end{aligned}$$

so that claim (57) follows by applying the basic bound (55). □

Finally, we state some particular choices of t that are useful for future reference. First, if we set $t = \sqrt{k/n}$, then since $k/n \leq 1$, we have

$$\delta(n,k,\sqrt{k/n}) \quad = \quad 4\Big\{\sqrt{\frac{k}{n}} + \frac{k}{n}\Big\} \; \leq \; 8\sqrt{\frac{k}{n}}.$$

Consequently, we obtain specialized versions of the bounds (56), of the form

$$\mathbb{P}[|\!|\!|\frac{X^T X}{n} - \Sigma|\!|\!|_2 \geq 8C_{max}\sqrt{\frac{k}{n}}] \leq 2\exp(-k/2), \text{ and} \quad (58a)$$

$$\mathbb{P}[|\!|\!|(\frac{X^T X}{n})^{-1} - (\Sigma)^{-1}|\!|\!|_2 \geq \frac{8}{C_{min}}\sqrt{\frac{k}{n}}] \leq 2\exp(-k/2). \quad (58b)$$

By setting $t = 1$ and performing some algebra, we obtain another set of very crude but adequate bounds on the spectral norms of random matrices. In particular, by triangle inequality, we have

$$\begin{aligned} |\!|\!|X^T X/n|\!|\!|_2 &\leq |\!|\!|\Sigma|\!|\!|_2 + |\!|\!|X^T X/n - \Sigma|\!|\!|_2 \\ &\leq C_{max} + C_{max}\delta(n,k,t) \end{aligned}$$

with probability greater than $1-2\exp(-nt^2/2)$. Setting $t = 1$, we find (using the bound $k \leq n$) that $\delta(n,k,1) \leq 8$ so that we can conclude that

$$\mathbb{P}[|\!|\!|X^T X/n|\!|\!|_2 \geq 9C_{max}] \quad \leq \quad 2\exp(-n/2). \qquad (59)$$

A similar argument yields that

$$\mathbb{P}[|\!|\!|(X^T X/n)^{-1}|\!|\!|_2 \geq \frac{9}{C_{min}}] \quad \leq \quad 2\exp(-n/2). \qquad (60)$$

References

[1] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[2] V. V. Buldygin and Y. V. Kozachenko. *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI, 2000.

[3] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52(2):489–509, February 2004.

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

[5] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[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] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL, 2001.

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

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

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

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

[12] D. L. 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.

[13] D. L. Donoho and J. M. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc*, July 2008.

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

[15] 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.

[16] J. Fan and R. Li. Variable selection via non-concave penalized likelihood and its oracle properties. *Jour. Amer. Stat. Ass.*, 96(456):1348–1360, December 2001.

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

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

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

[21] I. Johnstone. Chi-square oracle inequalities. In M. de Gunst, C. Klaassen, and A. van der Vaart, editors, *State of the Art in Probability and Statistics*, number 37 in IMS Lecture Notes, pages 399–418. Institute of Mathematical Statistics, 2001.

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

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

[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] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

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

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

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

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

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

[33] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using using ℓ_1-constrained quadratic programs. Technical Report 709, Department of Statistics, UC Berkeley, 2006.

[34] M. J. Wainwright. Information-theoretic bounds for sparsity recovery in the high-dimensional and noisy setting. Technical Report 725, Department of Statistics, UC Berkeley, January 2007. Posted as arxiv:math.ST/0702301; Presented at International Symposium on Information Theory, June 2007.

[35] M. J. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graph selection using ℓ_1-regularized logistic regression. In *NIPS Conference*, December 2006.

[36] W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. Technical Report arXiv:0806.0604, UC Berkeley, June 2008. Presented at ISIT 2008, Toronto, Canada.

[37] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

Biography: Martin Wainwright is currently an assistant professor at University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences. He received his Ph.D. degree in Electrical Engineering and Computer Science (EECS) from Massachusetts Institute of Technology (MIT). His research interests include statistical signal processing, coding and information theory, statistical machine learning, and high-dimensional statistics. He has been awarded an Alfred P. Sloan Foundation Fellowship, an NSF CAREER Award. the George M. Sprowls Prize for his dissertation research (EECS department, MIT), a Natural Sciences and Engineering Research Council of Canada 1967 Fellowship, and several outstanding conference paper awards.