Journal of Machine Learning Research 11 (2010) 2361-2386 Submitted 8/08; Revised 8/09; Published 8/10

High-dimensional Variable Selection with Sparse Random Projections: Measurement Sparsity and Statistical Efficiency

Dapo Omidiran DAPO@EECS.BERKELEY.EDU
Martin J. Wainwright[∗] WAINWRIG@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
UC Berkeley
Berkeley, CA 94720

Editor: Sanjoy Dasgupta

Abstract

We consider the problem of high-dimensional variable selection: given n noisy observations of a k-sparse vector $\beta^* \in \mathbb{R}^p$, estimate the subset of non-zero entries of β^*. A significant body of work has studied behavior of ℓ_1-relaxations when applied to random measurement matrices that are dense (e.g., Gaussian, Bernoulli). In this paper, we analyze *sparsified* measurement ensembles, and consider the trade-off between measurement sparsity, as measured by the fraction γ of non-zero entries, and the statistical efficiency, as measured by the minimal number of observations n required for correct variable selection with probability converging to one. Our main result is to prove that it is possible to let the fraction on non-zero entries $\gamma \to 0$ at some rate, yielding measurement matrices with a vanishing fraction of non-zeros per row, while retaining the same statistical efficiency as dense ensembles. A variety of simulation results confirm the sharpness of our theoretical predictions.

Keywords: variable selection, sparse random projections, high-dimensional statistics, Lasso, consistency, ℓ_1-regularization

1. Introduction

Recent years have witnessed a flurry of research on the recovery of high-dimensional models satisfying some type of sparsity constraint. These types of sparse recovery problems arise in a variety of domains, including variable selection in regression (Tibshirani, 1996), graphical model selection (Meinshausen and Buhlmann, 2006; Ravikumar et al., 2010), sparse principal components analysis (Johnstone and Lu, 2009; Paul, 2007), sparse approximation (Tropp, 2006), and compressed sensing (Candes and Tao, 2005; Donoho, 2006). In all of these settings, the basic problem is to recover information about a high-dimensional signal $\beta^* \in \mathbb{R}^p$, based on a set of n observations. The signal β^* is assumed *a priori* to be sparse: either exactly k-sparse, or lying within some ℓ_q-ball with $q < 1$.

A particular instance of high-dimensional sparse recovery involves the linear regression model $Y = X\beta^* + W$, where $Y \in \mathbb{R}^n$ is the observation vector, $W \in \mathbb{R}^n$ is observation noise, and $X \in \mathbb{R}^{n \times p}$ is the measurement matrix. In this context, high-dimensional scaling means that the sample size n can be of the same order of magnitude, or substantially smaller than the ambient dimension p. A particularly simple version of this model, studied extensively within the compressed sensing com-

[∗] Also in the Department of Statistics.

munity (Candes and Tao, 2005; Donoho, 2006), is the *noiseless version* in which $W = 0$, so that the model reduces to the under-determined linear system $Y = X\beta^*$. Here the model is only interesting when $n \ll p$, so that the linear system is not fully determined. Other work in machine learning and statistics (e.g., Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; Zhou et al., 2007) has focused on the noisy version of the model, say with $W \sim N(0, \sigma^2 I_p)$ for some noise variance σ^2. The problem of high-dimensional regression can be studied either for deterministic designs, for which the measurement matrix X is fixed, or for random designs in which X is drawn randomly from some ensemble. Past work on random designs has focused on the behavior of various ℓ_1-relaxations when applied to measurement matrices drawn from the standard Gaussian ensemble (e.g., Donoho, 2006; Candes and Tao, 2005), or more general random ensembles satisfying mutual incoherence conditions (Meinshausen and Buhlmann, 2006; Wainwright, 2009). A measurement matrix X drawn from such a standard random ensemble is dense, in that each row of X has p non-zero entries with high probability. In various applications of the sparse regression problem, the measurement matrix is itself a design variable, and dense measurement matrices are undesirable. For instance, in applications such as sensor networks (Wang et al., 2007), digital imaging (Wakin et al., 2006) or database management (Achlioptas, 2001; Li et al., 2006), it would be preferable to take measurements of the signal β^* based on sparse inner products, using measurement matrices X in which each row has a relatively small fraction of non-zero entries. Furthermore, sparse measurement matrices require significantly less storage space, and have the potential for reduced algorithmic complexity for signal recovery, since many algorithms for linear programming and conic programming more generally (Boyd and Vandenberghe, 2004), can be accelerated by exploiting problem structure.

In the noiseless instance of the regression problem (with $n \ll p$), the standard approach to estimating β^* is by solving the basis pursuit linear program (Chen et al., 1998). Recent work by Baraniuk et al. (2007) has established connections between the success of this method and the behavior of random projections, as characterized by the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2003). Random projections and their applications have been studied extensively in machine learning and related fields, with applications to dimensionality reduction (Dasgupta, 1999; Li et al., 2007), data stream processing (Alon et al., 1996; Indyk, 2006), databases (Achlioptas, 2001; Li et al., 2006) and compressed sensing (Wang et al., 2007; Baraniuk et al., 2007). One standard proof of the Johnson-Lindenstrauss lemma is based on random Gaussian matrices. Achlioptas (2001) was the first to apply sparse random projections, with each entry distributed on $\{-1, 0, 1\}$ with probabilities $[\frac{1}{6}, \frac{2}{3}, \frac{1}{6}]$, to the Johnson-Lindenstrauss problem setting, and to provide the same guarantees as dense projections. Unfortunately, his proof technique does not allow the non-zero mass to be decreased much beyond $\frac{1}{3}$. Other recent work (Li et al., 2006) provides theoretical and experimental justification for scaling the non-zero mass aggressively to zero. Stemming from different sources than the random projection literature, another line of recent work (e.g., Cormode and Muthukrishnan, 2005; Gilbert et al., 2006; Sarvotham et al., 2006; Xu and Hassibi, 2007) has studied compressed sensing methods based on sparse measurement matrices, using constructions motivated by group testing and coding theory.

Whereas this past work focuses on the noiseless recovery problem, our primary interest in this paper is the noisy linear observation model which, as we show, exhibits qualitatively different behavior than the noiseless case. At a high level, our primary goal in this paper is not to design sparse measurement matrices, but rather to gain a theoretical understanding of the *trade-off between the* *degree of measurement sparsity, and statistical efficiency*. We assess measurement sparsity in terms of the fraction γ of non-zero entries in any particular row of the measurement matrix, and we define statistical efficiency in terms of the minimal number of measurements n required to recover the correct support with probability converging to one. Our interest can be viewed in terms of experimental design: more precisely we ask, what degree of measurement sparsity can be permitted without increasing the number of observations required for correct variable selection or subset recovery?

To bring sharp focus to the issue, we analyze this question for exact subset recovery using ℓ_1-constrained quadratic programming, also known as the Lasso in the statistics literature (Chen et al., 1998; Tibshirani, 1996), where past work on dense Gaussian measurement ensembles (Wainwright, 2009) provides a precise characterization of its success/failure. We characterize the density of our measurement ensembles with a positive parameter $\gamma \in (0,1]$, corresponding to the fraction of non-zero entries per row. We first show that for all fixed $\gamma \in (0,1]$, the statistical efficiency of the Lasso remains the same as with dense measurement matrices. We then prove that it is possible to let $\gamma \to 0$ at some rate, as a function of the sample size n, signal length p and signal sparsity k, yielding measurement matrices with a vanishing fraction of non-zeroes per row while requiring exactly the same number of observations as dense measurement ensembles. In general, in contrast to the noiseless setting (Xu and Hassibi, 2007), our theory still requires that the average number of non-zeroes per column of the measurement matrix (i.e., γn) tend to infinity, however, under the loss function considered here (exact signed support recovery), we prove that no method can succeed with probability one if this condition does not hold.

The remainder of this paper is organized as follows. In Section 2, we set up the problem more precisely, state our main result, and discuss some of its implications. In Section 3, we provide a high-level outline of the proof with more technical aspects of the argument deferred to the appendices. We provide some illustrative simulations in Section 4 that illustrate the sharpness of our theoretical predictions. Work in this paper was presented in part at the International Symposium on Information Theory in Toronto, Canada (July, 2008). We note that concurrent and complementary work (Wang et al., 2010) analyzes the information-theoretic limitations of sparse measurement matrices for exact support recovery.

1.1 Notation

Throughout this paper, we use the following standard asymptotic notation: $f(n) = O(g(n))$ if $f(n) \leq Cg(n)$ for some constant $C < +\infty$; $f(n) = \Omega(g(n))$ if $f(n) \geq cg(n)$ for some constant $c > 0$; and $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. In addition, we use $\log(x)$ to denote the natural logarithm of x.

2. Problem Set-up and Main Result

We begin by setting up the problem, stating our main result, and discussing some of its consequences.

2.1 Observation Model

Define the *support set* of a signal $\beta \in \mathbb{R}^p$

$$S(\beta) \quad := \quad \{i \in \{1,\dots,p\} \mid \beta_i \neq 0\},$$

and consider the class of k-sparse signals of length p:

$$\mathcal{C}(p,k,\beta^*_{\min}) \quad := \quad \left\{\beta \in \mathbb{R}^p \mid |S(\beta)| = k \leq \frac{p}{2}, \min_{i \in S} |\beta_i| \geq \beta^*_{\min}\right\}. \tag{1}$$

Let $\beta^* \in \mathbb{R}^p$ be a fixed but unknown vector in $\mathcal{C}(p,k,\beta^*_{\min})$, and suppose that we make a set $\{Y_1,\ldots,Y_n\}$ of n independent and identically distributed (i.i.d.) observations of the unknown vector β^*, each of the form

$$Y_i \quad := \quad x_i^T \beta^* + W_i, \tag{2}$$

where $W_i \sim \mathcal{N}(0,1)$ is observation noise, and $x_i \in \mathbb{R}^p$ is a measurement vector. Note that there is no loss in generality in assuming that the noise variance is one, since the observation model with signal class $\mathcal{C}(p,k,\beta^*_{\min})$ and $W_i \sim \mathcal{N}(0,1)$ is equivalent to the observation model with signal class $\mathcal{C}(p,k,\frac{\beta^*_{\min}}{\sigma})$ with noise variance $W_i \sim \mathcal{N}(0,\sigma^2)$.

It is convenient to use $Y = \begin{bmatrix} Y_1 & Y_2 & \ldots & Y_n \end{bmatrix}^T$ to denote the n-vector of measurements, with similar notation for the noise vector $W \in \mathbb{R}^n$, and

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

to denote the $n \times p$ measurement matrix. With this notation, the observation model can be written compactly as $Y = X\beta^* + W$.

2.2 Sign Consistency and Statistical Efficiency

Given some estimate $\widehat{\beta}$, its error relative to the true β^* can be assessed in various ways, depending on the underlying application of interest. For applications in compressed sensing, various types of ℓ_r norms (i.e., $\mathbb{E}\|\widehat{\beta} - \beta^*\|_r^r$) are well-motivated, whereas for statistical prediction, it is most natural to study a predictive loss (e.g., $\|X(\widehat{\beta} - \beta^*)\|_2^2/n$). For reasons of scientific interpretation or for model selection purposes, the object of primary interest is the support S of β^*. In this paper, we consider a slightly stronger notion of model selection: in particular, our goal is to recover the *signed support* of the unknown β^*, as defined by the p-vector $S_\pm(\beta^*)$ with elements

$$[S_\pm(\beta^*)]_i \quad := \quad \begin{cases} \text{sign}(\beta_i^*) & \text{if } \beta_i^* \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Given some estimate $\widehat{\beta}$, we study the probability $\mathbb{P}[S_\pm(\widehat{\beta}) = S_\pm(\beta^*)]$ that it correctly specifies the signed support. In particular, a sequence of estimates $\widehat{\beta}_{n,p,k}$ is *sign consistent* if

$$\mathbb{P}[S_\pm(\widehat{\beta}_{n,p,k}) = S_\pm(\beta^*)] \to 1, \text{ as } n,p,k \to \infty. \tag{3}$$

The estimator that we analyze is ℓ_1-constrained quadratic programming (QP), also known as the Lasso (Tibshirani, 1996) in the statistics literature. The Lasso generates an estimate $\widehat{\beta}$ by solving the regularized QP

$$\widehat{\beta} \quad = \quad \arg\min_{\beta \in \mathbb{R}^p} \left\{\frac{1}{2n}\|Y - X\beta\|_2^2 + \rho_n \|\beta\|_1\right\}, \tag{4}$$

where $\rho_n > 0$ is a user-defined regularization parameter. A large body of past work has focused on the model selection behavior of the Lasso for both deterministic and random measurement matrices (e.g., Tropp, 2006; Zhao and Yu, 2006; Wainwright, 2009).

Tropp (2006) demonstrates that under some technical conditions the Lasso produces an estimate with support contained in the true support set, while Zhao and Yu (2006) prove the sign consistency of the Lasso for certain sequences (n, k_n, p_n) and measurement ensemble covariances. The main contribution of Wainwright (2009) is to identify the smallest n required for sign consistency of the Lasso when applied to measurement matrices X drawn randomly from Gaussian ensembles, with the standard Gaussian ensemble (i.e., each element $X_{ij} \sim \mathcal{N}(0,1)$ i.i.d.) being one special case. Define the function $n_{crit}(p,k) := 2k\log(p-k)$. The paper (Wainwright, 2009) shows the Lasso undergoes a phase transition as a function of the control parameter

$$\theta(n,p,k) \quad := \quad \frac{n}{n_{crit}(p,k)}. \tag{5}$$

In more detail, for the special case of the standard Gaussian ensemble, for any sequence (n,p,k) such that $\theta(n,p,k) > 1+\varepsilon$ for some $\varepsilon > 0$, the Lasso (with an appropriate choice of regularization parameter ρ_n) is sign consistent with probability converging to one. In contrast, it fails to be sign consistent with high probability, regardless of the choice of ρ_n, for any sequences such that $\theta(n,p,k) < 1-\varepsilon$.

The main contribution of this paper is to show that sparse measurement ensembles are *statistically efficient*: the same sharp threshold (5) holds for γ-sparsified measurement ensembles, including a subset for which $\gamma \to 0$, so that each row of the measurement matrix has a vanishing fraction of non-zero entries.

2.3 Statement of Main Result

A measurement matrix $X \in \mathbb{R}^{n\times p}$ drawn randomly from a Gaussian ensemble is dense, in that each row has $\Theta(p)$ non-zero entries. The main focus of this paper is the observation model (2), using measurement ensembles that are designed to be sparse. To formalize the notion of sparsity, we let $\gamma \in (0,1]$ represent a *measurement sparsity parameter*, corresponding to the (average) fraction of non-zero entries per row. Our analysis allows the sparsity parameter $\gamma(n,p,k)$ to be a function of the triple (n,p,k), but we typically suppress this explicit dependence so as to simplify notation. For a given choice of γ, we consider measurement matrices X with i.i.d. entries of the form

$$X_{ij} \quad \overset{d}{=} \quad \begin{cases} Z \sim \mathcal{N}(0,1) & \text{with probability } \gamma \\ 0 & \text{with probability } 1-\gamma. \end{cases} \tag{6}$$

By construction, the expected number of non-zero entries in each row of X is γp. In fact, the analysis of this paper establishes exactly the same control parameter threshold (5) for γ-sparsified measurement ensembles, for any fixed $\gamma \in (0,1)$, as the completely dense case ($\gamma = 1$). In particular, we state the following result on conditions under which the Lasso applied to sparsified ensembles has the same *sample complexity* as when applied to the dense (standard Gaussian) ensemble:

Theorem 1 *Suppose that the measurement matrix $X \in \mathbb{R}^{n\times p}$ is drawn with i.i.d. entries according to the γ-sparsified distribution* (6). *Then for any $\varepsilon > 0$, if the sample size satisfies*

$$n \quad > \quad (2+\varepsilon)k\log(p-k), \tag{7}$$

then the Lasso is sign consistent as $(n,p,k)\to+\infty$ *so long as*

$$\begin{aligned}\frac{n\rho_n^2\gamma}{\log(p-k)} &\rightarrow \infty, & (8)\\ \frac{\rho_n}{\beta^*_{\min}}\max\{1,\frac{\sqrt{k}}{\gamma}\sqrt{\frac{\log\log(p-k)}{\log(p-k)}}\} &\rightarrow 0, & (9)\\ \gamma^3\min\{k,\frac{\log(p-k)}{\log\log(p-k)}\} &\rightarrow \infty. & (10)\end{aligned}$$

Remark 2 *(a) Note that the sample complexity* (7) *is identical to the Lasso threshold proven in past work (Wainwright, 2009). To gain intuition on the remaining conditions, it is helpful to consider various special cases of the sparsity parameter* γ*. If* γ *is a constant fixed to some value in* $(0,1]$*, then it plays no role in the scaling, and condition* (10) *is always satisfied. Conditions* (8) *and* (9) *are slightly weaker than the corresponding condition from previous work, in that they require that* ρ_n *be slightly larger, and hence that* $\beta^*_{\min}$ *must approach zero more slowly than the requirement of this previous work. Depending on the exact behavior of* $\beta^*_{\min}$*, choosing* ρ_n^2 *to decay slightly more slowly than* $\log p/n$ *is sufficient to guarantee exact recovery with* $n=\Theta(k\log(p-k))$*, meaning that we recover exactly the same statistical efficiency as the dense case (*$\gamma=1$*) for all constant measurement sparsities* $\gamma\in(0,1)$*. At least initially, one might think that reducing* γ *should increase the required number of observations, since it effectively reduces the signal-to-noise ratio by a factor of* γ*. However, under high-dimensional scaling (*$p\to+\infty$*), a major effect limiting the Lasso performance is the number (*$p-k$*) of irrelevant factors, and under the scaling considered here, this effect is dominant.*

(b) However, Theorem 1 also allows for general scalings of the measurement sparsity γ *along with the triplet* (n,p,k)*. More concretely, let us suppose for simplicity that* $\beta^*_{\min}=\Theta(1)$*. Then over a range of signal sparsities—say* $k=\alpha p$*,* $k=\Theta(\sqrt{p})$ *or* $k=\Theta(\log(p-k))$*, corresponding respectively to linear sparsity, polynomial sparsity, and exponential sparsity—we can choose a decaying measurement sparsity, for instance*

$$\gamma \;=\; [\frac{\log\log(p-k)}{\log(p-k)}]^{\frac{1}{6}}\to 0 \tag{11}$$

along with the regularization parameter $\rho_n^2=\frac{\log(p-k)}{n}\sqrt{\frac{\log(p-k)}{\log\log(p-k)}}$ *while maintaining the same sample complexity (required number of observations for support recovery) as the Lasso with dense measurement matrices.*

(c) Of course, the conditions of Theorem 1 do not allow the measurement sparsity γ *to approach zero arbitrarily quickly. Rather, for any* γ *guaranteeing exact recovery, condition* (8) *implies that the average number of non-zero entries per column of X (namely,* γn*) must tend to infinity. (Indeed, with* $n=\Omega(k\log(p-k))$*, our specific choice* (11) *certainly satisfies this constraint.) A natural question is whether exact recovery is possible using measurement matrices, either randomly drawn or deterministically designed, with the average number of non-zeros per column (namely* γn*) remaining bounded. In fact, under the criterion of exactly recovering the signed support* (3)*, as shown by the following result, if* $\beta^*_{\min}=O(1)$*, then no method can succeed with probability converging to one unless* γn *tends to infinity.*

Proposition 1 *If $\gamma n[\beta^*_{\min}]^2$ does not tend to infinity, then no method can recover the signed support with probability one.*

Proof We construct a sub-problem that must be solvable by any method capable of performing exact signed support recovery. Suppose that $\beta^*_1 = \beta^*_{\min} \neq 0$ and that the column X_1 has n_1 non-zero entries, say without loss of generality indices $i = 1, \ldots, n_1$. Now consider the problem of recovering the sign of β^*_1. Let us extract the observations $i = 1, \ldots, n_1$ that explicitly involve β^*_1, writing

$$Y_i \quad = \quad X_{i1}\beta^*_1 + \sum_{j \in T(i)} X_{ij}\beta^*_j + W_i, \qquad i = 1, \ldots, n_1 \tag{12}$$

where $T(i)$ denotes the set of indices in row i for which X_{ij} is non-zero, excluding index 1. Even assuming that $\{\beta^*_j, j \in T(i)\}$ were perfectly known, this observation model (12) is at best equivalent to observing β^*_1 contaminated by constant variance additive Gaussian noise, and our task is to distinguish whether $\beta^*_1 = \beta^*_{\min}$ or $\beta^*_1 = -\beta^*_{\min}$. The average $\bar{Y} = \frac{1}{n_1}\sum_{i=1}^{n_1}[Y_i - \sum_{j \in T(i)} X_{ij}\beta^*_j]$ is a sufficient statistic, following the distribution $\bar{Y} \sim \mathcal{N}(\beta^*_{\min}, \frac{1}{n_1})$. Unless the effective signal-to-noise ratio, which is of the order $n_1[\beta^*_{\min}]^2$, goes to infinity, there will always be a constant probability of error in distinguishing $\beta^*_1 = \beta^*_{\min}$ from $\beta^*_1 = -\beta^*_{\min}$. Under the γ-sparsified random ensemble, we have $n_1 \leq (1 + o(1))\gamma n$ with high probability, so that no method can succeed unless $\gamma n[\beta^*_{\min}]^2$ goes to infinity, as claimed. ■

Note that the conditions in Theorem 1 imply that $n\gamma[\beta^*_{\min}]^2 \to +\infty$. In particular, condition (9) implies that $\rho_n^2 = o([\beta^*_{\min}]^2)$, and condition (8) implies that $n\gamma\rho_n^2 \to +\infty$, which verifies the condition of Proposition 1.

3. Proof of Theorem 1

This section is devoted to the proof of Theorem 1. We begin with a high-level outline of the proof; as with previous work on dense Gaussian ensembles (Wainwright, 2009), the key is the notion of a *primal-dual witness* for exact signed support recovery. The proof itself involves a number of additional steps not needed in this past work, in order to gain good control on sparse matrices as opposed to generic Gaussian matrices (see Appendix D). The proof is divided into a sequence of separate lemmas, with some of the more technical results deferred to the appendices.

3.1 High-level Overview of Proof

For the purposes of our proof, it is convenient to consider matrices $X \in \mathbb{R}^{n \times p}$ with i.i.d. entries of the form

$$X_{ij} \quad \stackrel{d}{=} \quad \begin{cases} Z \sim \mathcal{N}(0, \frac{1}{\gamma}) & \text{with probability } \gamma \\ 0 & \text{with probability } 1 - \gamma. \end{cases}$$

So as to obtain an equivalent observation model, we also reset the variance of each noise term W_i to be $\frac{1}{\gamma}$. Finally, we can assume without loss of generality that $S_+(\beta^*_S) = \vec{1} \in \mathbb{R}^k$.

Define the *sample covariance matrix*

$$\widehat{\Sigma} \quad := \quad \frac{1}{n}X^T X \; = \; \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T.$$

Of particular importance to our analysis is the $k \times k$ sub-matrix $\widehat{\Sigma}_{SS}$. For future reference, we state the following claim, proved in Appendix D:

Lemma 1 *Under the conditions of Theorem 1, the submatrix $\widehat{\Sigma}_{SS}$ is invertible with probability greater than* $1 - O(\frac{1}{(p-k)^2})$.

The foundation of our proof is the following lemma: it provides sufficient conditions for the Lasso (4) to recover the signed support set.

Lemma 2 (Primal-dual conditions for support recovery) *Suppose that $\widehat{\Sigma}_{SS} \succ 0$, and that we can find a primal vector $\widehat{\beta} \in \mathbb{R}^p$, and a subgradient vector $\widehat{z} \in \mathbb{R}^p$ that satisfy the* zero-subgradient condition

$$\widehat{\Sigma}\big(\widehat{\beta} - \beta^*\big) - \frac{1}{n} X^T W + \rho_n \widehat{z} \quad = \quad 0, \tag{13}$$

and the signed-support-recovery conditions

$$\begin{aligned} \widehat{z}_i &\;=\; \operatorname{sign}(\beta_i^*) \qquad \textit{for all } i \in S, && (14)\\ \widehat{\beta}_j &\;=\; 0 \qquad \textit{for all } j \in S^c, && (15)\\ |\widehat{z}_j| &\;<\; 1 \qquad \textit{for all } j \in S^c, \textit{ and} && (16)\\ \operatorname{sign}(\widehat{\beta}_i) &\;=\; \operatorname{sign}(\beta_i^*) \qquad \textit{for all } i \in S. && (17) \end{aligned}$$

Then $\widehat{\beta}$ is the unique optimal solution to the Lasso (4)*, and recovers the correct signed support.*

See Appendix B.1 for the proof of this claim.

On the basis of Lemmas 1 and 2, it suffices to show that under the specified scaling of (n, p, k), there exists a primal-dual pair $(\widehat{\beta}, \widehat{z})$ satisfying the conditions of Lemma 2. We establish the existence of such a pair with the following constructive procedure:

(a) We begin by setting $\widehat{\beta}_{S^c} = 0$, and $\widehat{z}_S = \operatorname{sign}(\beta_S^*)$.

(b) Next we determine $\widehat{\beta}_S$ by solving the linear system

$$\widehat{\Sigma}_{SS}\big(\widehat{\beta}_S - \beta_S^*\big) - \frac{1}{n} X_S^T W + \rho_n \operatorname{sign}(\beta_S^*) \quad = \quad 0.$$

(c) Finally, we determine $\widehat{z}_{S^c}$ by solving the linear system:

$$-\rho_n \widehat{z}_{S^c} \quad = \quad \widehat{\Sigma}_{S^c S}\big(\widehat{\beta}_S - \beta_S^*\big) - \frac{1}{n} X_{S^c}^T W.$$

By construction, this procedure satisfies the zero sub-gradient condition (13), as well as auxiliary conditions (14) and (15); it remains to verify conditions (16) and (17).

In order to complete these final two steps, it is helpful to define for $i \in S$ and $j \in S^c$ the following random variables:

$$\begin{aligned} V_j^a &\;:=\; X_j{}^T \big[\frac{1}{n} X_S (\widehat{\Sigma}_{SS})^{-1} \vec{1}\big] \rho_n, && (18)\\ V_j^b &\;:=\; X_j{}^T \big[I_{n \times n} - \frac{1}{n} X_S (\widehat{\Sigma}_{SS})^{-1} X_S^T\big] \frac{W}{n}, \quad \text{and} && (19)\\ U_i &\;:=\; e_i^T \widehat{\Sigma}_{SS}^{-1} \big[\frac{1}{n} X_S^T W - \rho_n \vec{1}\big], && (20) \end{aligned}$$

where $e_i \in \mathbb{R}^k$ is the unit vector with one in position i, and $\vec{1} \in \mathbb{R}^k$ is the all-ones vector.

A little bit of algebra (see Appendix B.2 for details) shows that $\rho_n \widehat{z}_j = V_j^a + V_j^b$, and that $U_i = \widehat{\beta}_i - \beta_i^*$. Consequently, if we define the events

$$\mathcal{E}(V) \;:=\; \big\{\max_{j \in S^c} |V_j^a + V_j^b| < \rho_n\big\}, \tag{21}$$

$$\mathcal{E}(U) \;:=\; \big\{\max_{i \in S} |U_i| \leq \beta^*_{\min}\big\}, \tag{22}$$

where $\beta^*_{\min}$ was defined previously as the minimum value of $|\beta^*|$ on its support, then in order to establish that the Lasso succeeds in recovering the exact signed support, it suffices to show that $\mathbb{P}[\mathcal{E}(V) \cap \mathcal{E}(U)] \to 1$,

We decompose the proof of this final claim in the following three lemmas. As in the statement of Theorem 1, suppose that $n > (2+\varepsilon)k\log(p-k)$, for some fixed $\varepsilon > 0$.

Lemma 3 (Control of V^a) *Under the conditions of Theorem 1, there exists a fixed positive value δ (dependent on ε) such that*

$$\mathbb{P}[\max_{j \in S^c} |V_j^a| \geq (1-\delta)\rho_n] \quad \to \quad 0.$$

Lemma 4 (Control of V^b) *Under the conditions of Theorem 1, there exists a fixed positive value δ (dependent on ε)*

$$\mathbb{P}[\max_{j \in S^c} |V_j^b| \geq \delta\rho_n] \quad \to \quad 0.$$

Lemma 5 (Control of U) *Under the conditions of Theorem 1, we have*

$$\mathbb{P}[(\mathcal{E}(U))^c] \;=\; \mathbb{P}[\max_{i \in S} |U_i| > \beta^*_{\min}] \quad \to \quad 0.$$

3.2 Proof of Lemma 3

We assume throughout that $\widehat{\Sigma}_{SS}$ is invertible, an event which occurs with probability $1-o(1)$ under the stated assumptions (see Lemma 1). If we define the n-dimensional vector

$$h \;:=\; X_S(\widehat{\Sigma}_{SS})^{-1}\vec{1}, \tag{23}$$

then the variable V_j^a can be written compactly as

$$\frac{V_j^a}{\rho_n} \;=\; X_j^T h = \sum_{\ell=1}^n h_\ell X_{\ell j}.$$

Note that each term $X_{\ell j}$ in this sum is distributed as a mixture variable, taking the value 0 with probability $1-\gamma$, and distributed as $\mathcal{N}(0, \frac{1}{\gamma})$ variable with probability γ. For $\ell = 1, \ldots, n$ and each j, define the random vector H^j, with entries

$$H_\ell^j \;\overset{d}{=}\; \begin{cases} h_\ell & \text{with probability } \gamma \\ 0 & \text{with probability } 1-\gamma. \end{cases}$$

For each index $\ell = 1, \ldots, n$, let $Z_{\ell j} \sim \mathcal{N}(0, \frac{1}{\gamma})$. With these definitions, by construction, we have that

$$\frac{V_j^a}{\rho_n} \overset{d}{=} \sum_{\ell=1}^{n} H_\ell^j Z_{\ell j}.$$

To gain some intuition for the behavior of this sum, note that the variables $\{Z_{\ell j}, \ell = 1, \ldots, n\}$ are independent of the vector H^j. (In particular, H^j is a function of X_S, whereas $Z_{\ell j}$ is a function of $X_{\ell j}$, with $j \notin S$.) Consequently, we may condition on H^j without affecting Z, and since Z is Gaussian, we have $(\frac{V_j^a}{\rho_n} \mid H^j) \sim \mathcal{N}(0, \frac{\|H^j\|_2^2}{\gamma})$. Therefore, if we can obtain good control on the norm $\|H^j\|_2$, then we can use standard Gaussian tail bounds (see Appendix A) to control the maximum $\max_{j \in S^c} V_j^a / \rho_n$. The following lemma is proved in Appendix C:

Lemma 6 *Under condition* (10)*, then for any fixed* $\delta > 0$*, we have*

$$\mathbb{P}\big[\|H^j\|_2^2 \leq \frac{\gamma k(1+\delta)}{n}\big] \quad \geq \quad 1 - O\big[\exp(-\min\{2\log(p-k), \frac{n}{2k}\})\big].$$

The primary implication of the above bound is that each V_j^a/ρ_n variable is (essentially) no larger than a $\mathcal{N}(0, \frac{k}{n})$ variable. We can then use standard techniques for bounding the tails of Gaussian variables to obtain good control over the random variable $\max_{j \in S^c} |V_j^a|/\rho_n$. In particular, by the union bound, we have

$$\mathbb{P}[\max_{j \in S^c} |V_j^a| \geq (1-\delta)\rho_n] \quad \leq \quad (p-k)\, \mathbb{P}[\sum_{\ell=1}^{n} H_\ell^j Z_{\ell j} \geq (1-\delta)].$$

For any $\delta > 0$, define the event $\mathcal{T}^j(\delta) := \{\|H^j\|_2^2 \leq \frac{k\gamma(1+\delta)}{n}\}$. With this definition, we have

$$\begin{aligned}
\mathbb{P}[\max_{j \in S^c} |V_j^a| \geq (1-\delta)\rho_n] \quad &\leq \quad (p-k)\left\{\mathbb{P}[\sum_{\ell=1}^{n} H_\ell^j Z_{\ell j} \geq (1-\delta) \mid \mathcal{T}^j(\delta)] + \mathbb{P}[(\mathcal{T}^j(\delta)^c)]\right\} \\
&\leq \quad (p-k)\, \{2\exp(-\frac{n(1-\delta)^2}{2k(1+\delta)}) + c_1 \exp\big(-\min\big(2\log(p-k), \frac{n}{2k}\big)\big)\},
\end{aligned}$$

where the second inequality uses a standard Gaussian tail bound (see Appendix A), and Lemma 6. Finally, let us assume the condition $n > (2+\varepsilon)k\log(p-k)$ for some fixed $\varepsilon > 0$. Then there exists a numerical constant c_1 such that

$$\begin{aligned}
\mathbb{P}[\max_{j \in S^c} |V_j^a| \geq (1-\delta)\rho_n] \quad &\leq \quad (p-k)\, \big\{2\exp\big(-\frac{n(1-\delta)^2}{2k(1+\delta)}\big) + c_1\exp(-\min(2\log(p-k), \frac{n}{2k}))\big\} \\
&= \quad (p-k)\, \big\{2\exp\big(-\frac{n(1-\delta)^2}{2k(1+\delta)}\big) + c_1\exp(-2\log(p-k))\big\} \\
&\leq \quad (p-k)\, \big\{2\exp\big(-(2+\varepsilon)\log(p-k)\frac{(1-\delta)^2}{2(1+\delta)}\big) \\
&+ \quad c_1\exp(-2\log(p-k))\big\}.
\end{aligned}$$

Note that the above inequality holds for all values of ε. Since $\varepsilon > 0$ is fixed, we can choose a fixed value of δ such that $\frac{(1-\delta)^2}{1+\delta} > \frac{2}{2+\varepsilon/2}$. With this choice, we then have

$$\begin{aligned}
\mathbb{P}[\max_{j \in S^c} |V_j^a| \geq (1-\delta)\rho_n] \quad &\leq \quad (p-k)\, \big\{2\exp\big(-\frac{(2+\varepsilon)}{(2+\varepsilon/2)}\log(p-k)\big) + c_1\exp(-2\log(p-k))\big\} \\
&\leq \quad (p-k)\left\{(2+c_1)\exp\big(-\frac{(2+\varepsilon)}{(2+\varepsilon/2)}\log(p-k)\big)\right\},
\end{aligned}$$

thereby establishing that $\mathbb{P}[\max_{j\in S^c}|V_j^a| \geq (1-\delta)\rho_n] \to 0$, as claimed.

3.3 Proof of Lemma 4

Defining the orthogonal projection matrix $\Pi_S^{\perp} := I_{n\times n} - X_S(X_S^T X_S)^{-1}X_S^T$, we then have

$$\begin{aligned}\mathbb{P}[\max_{j\in S^c}|V_j^b| \geq \delta\rho_n] &= \mathbb{P}[\max_{j\in S^c}\big|X_j^T\Pi_S^{\perp}(\frac{W}{n})\big| \geq \delta\rho_n] \\ &\leq (p-k)\,\mathbb{P}\big[\big|X_j^T\Pi_S^{\perp}(\frac{W}{n})\big| \geq \delta\rho_n\big]. \qquad (24)\end{aligned}$$

Now since $\Pi_S^{\perp}$ is an orthogonal projection matrix, and the column vector X_j, noise vector W and randomness in $\Pi_S^{\perp}$ are all independent, we have the bound

$$\mathbb{P}\big[\big|X_j^T\Pi_S^{\perp}(W/n)\big| \geq \delta\rho_n\big] \leq \mathbb{P}\big[\big|X_j^T(W/n)\big| \geq \delta\rho_n\big], \qquad (25)$$

For each $\ell = 1,\ldots,n$ and $j = 1,\ldots,p$, let $B_{\ell j}$ be a Bernoulli variable with parameter γ, and let $Z_{\ell j} \sim \mathcal{N}(0,\frac{1}{\gamma})$. In terms of these random variables, we have the representation $X_{\ell j} = B_{\ell j}Z_{\ell j}$. Note moreover that the the sum $\sum_{\ell=1}^n B_{\ell j}$ is a binomial random variable, and define the event

$$\mathcal{T} := \{\frac{1}{n}\big|\sum_{\ell=1}^n B_{\ell j} - \gamma n\big| \leq \frac{1}{2\sqrt{k}}\}.$$

From the Hoeffding bound (see Lemma 7), we have $\mathbb{P}[\mathcal{T}^c] \leq 2\exp(-\frac{n}{2k})$. Using the representation $X_{\ell j} = B_{\ell j}Z_{\ell j}$ and conditioning on $\mathcal{T}$, we obtain

$$\begin{aligned}\mathbb{P}\big[\big|X_j^T W/n\big| \geq \delta\rho_n\big] &\leq \mathbb{P}\big[\big|\frac{1}{n}\sum_{\ell=1}^n B_{\ell j}Z_{\ell j}W_\ell\big| \geq \delta\rho_n \;\mid\; \mathcal{T}\big] + \mathbb{P}[\mathcal{T}^c] \\ &\leq \mathbb{P}\big[\big|\frac{1}{n}\sum_{\ell=1}^{n(\gamma+\frac{1}{2\sqrt{k}})} Z_{\ell j}W_\ell\big| \geq \delta\rho_n\big] + 2\exp(-\frac{n}{2k}),\end{aligned}$$

where we have assumed without loss of generality that the first $n(\gamma+\frac{1}{2\sqrt{k}})$ variables in the collection $\{B_{\ell j}\}_{\ell=1}^n$ are non-zero.

Conditioned on W, the random variable $M_j := \frac{1}{n}\sum_{\ell=1}^{n(\gamma+\frac{1}{2\sqrt{k}})} Z_{\ell j}W_\ell$ is zero-mean Gaussian with variance $\nu(W;\gamma) := \frac{1}{n^2\gamma}\sum_{\ell=1}^{n(\gamma+\frac{1}{2\sqrt{k}})} W_\ell^2$. For some $\delta_1 > 0$, define the event

$$\mathcal{T}_2(\delta_1) := \left\{\nu(W;\gamma) \leq (1+\delta_1)\frac{1}{n\gamma^2}\,(\gamma+\frac{1}{2\sqrt{k}})\right\}.$$

In order to bound the probability of this event, we begin by observing that since $W_\ell \sim N(0,1/\gamma)$, the variable $n^2\gamma^2\nu(W;\gamma)$ is chi-squared with $d = n(\gamma+\frac{1}{2\sqrt{k}})$ degrees of freedom. Consequently, using χ^2-tail bounds (see Appendix A), we have

$$\mathbb{P}[(\mathcal{T}_2(\delta_1))^c] \leq \exp\big(-n(\gamma+\frac{1}{2\sqrt{k}})\frac{3\delta_1^2}{16}\big).$$

Now, by conditioning on $\mathcal{T}_2(\delta_1)$ and its complement and using tail bounds on Gaussian variates (see Appendix A), we obtain

$$\begin{aligned}\mathbb{P}\big[\big|\frac{1}{n}\sum_{\ell=1}^{n(\gamma+\frac{1}{2\sqrt{k}})} Z_{\ell j}W_\ell\big| \geq \delta\rho_n\big] &\leq \mathbb{P}\big[\big|\frac{1}{n}\sum_{\ell=1}^{n(\gamma+\frac{1}{2\sqrt{k}})} Z_{\ell j}W_\ell\big| \geq \delta\rho_n \mid \mathcal{T}_2(\delta_1)\big] + \mathbb{P}[(\mathcal{T}_2(\delta_1))^c] \\ &\leq 2\,\exp\big(-\frac{n\gamma^2(\delta^2\rho_n^2)}{2(1+\delta_1)(\gamma+\frac{1}{2\sqrt{k}})}\big) + \\ &\qquad \exp\big(-n(\gamma+\frac{1}{2\sqrt{k}})\frac{3\delta_1^2}{16}\big). \end{aligned} \tag{26}$$

Finally, putting together the pieces from Equations (26), (25), and (24) we obtain that $\mathbb{P}[\max_{j\in S^c}|V_j^b| \geq \delta\rho_n]$ is upper bounded by

$$(p-k)\,\big\{2\exp(-\frac{n}{2k}) + 2\,\exp\big(-\frac{n\gamma^2(\delta^2\rho_n^2)}{2(1+\delta_1)(\gamma+\frac{1}{2\sqrt{k}})}\big) + \exp\big(-n(\gamma+\frac{1}{2\sqrt{k}})\frac{3\delta_1^2}{16}\big)\big\}. \tag{27}$$

The first term in Equation (27) goes to zero since $n > (2+\varepsilon)k\log(p-k)$. Condition (10) implies that $\gamma\sqrt{k} \to \infty$. In particular, this means that eventually $1 \leq 2\gamma\sqrt{k}$. Once this occurs, we have the inequality $\frac{\gamma^2}{\gamma+\frac{1}{2\sqrt{k}}} > \frac{\gamma}{2}$, and hence

$$\begin{aligned}(p-k)\exp\big(-\frac{n\gamma^2(\delta^2\rho_n^2)}{2(1+\delta_1)(\gamma+\frac{1}{2\sqrt{k}})}\big) &\leq (p-k)\exp\big(-\frac{n\gamma\delta^2\rho_n^2}{4(1+\delta_1)}\big) \\ &= (p-k)\exp\big(-\log(p-k)\frac{n\gamma\delta^2\rho_n^2}{4(1+\delta_1)\log(p-k)}\big).\end{aligned}$$

Recalling that the terms δ and δ_1 are fixed constants, condition (10) implies that eventually

$$(p-k)\exp\big(-\log(p-k)\frac{n\gamma\delta^2\rho_n^2}{4(1+\delta_1)\log(p-k)}\big) \leq (p-k)\exp\big(-3\log(p-k)\big),$$

showing that the middle term of Equation (27) goes to zero. Finally, using the condition $n \geq (2+\varepsilon)k\log(p-k)$, we obtain

$$\begin{aligned}(p-k)\exp\big(-n(\gamma+\frac{1}{2\sqrt{k}})\frac{3\delta_1^2}{16}\big) &\leq (p-k)\exp\big(-n(\frac{1}{2\sqrt{k}})\frac{3\delta_1^2}{16}\big) \\ &\leq (p-k)\exp\big(-\big[\frac{(2+\varepsilon)\sqrt{k}}{2}\big]\log(p-k)\frac{3\delta_1^2}{16}\big).\end{aligned}$$

This quantity tends to zero, because ε and δ_1 are fixed constants and $\sqrt{k}$ tends to infinity. We conclude that the last term in Equation (27) goes to zero, thereby concluding the proof.

3.4 Proof of Lemma 5

We first observe that conditioned on X_S, each U_i is Gaussian with mean and variance

$$\begin{aligned} m_i &:= \mathbb{E}[U_i \mid X_S] = e_i^T\big(\frac{1}{n}X_S^TX_S\big)^{-1}\big[-\rho_n\vec{1}\big], \quad \text{and} \\ \psi_i &:= \text{var}[U_i \mid X_S] = \frac{1}{\gamma n}e_i^T\big(\frac{1}{n}X_S^TX_S\big)^{-1}e_i, \end{aligned}$$

respectively. Let us define the function

$$T(\gamma, k, p, \theta, t) \quad := \quad \frac{1}{\gamma}\sqrt{\max\Big\{\frac{\log(t)}{\theta k \log(p-k)}, \frac{\log[\theta \log(p-k)]}{\theta \log(p-k)}\Big\}}, \tag{28}$$

as well as the the upper bounds

$$m^* := \rho_n(1 + C\sqrt{k}T(\gamma, k, p, 1, k)), \quad \text{and} \quad \psi^* := \frac{1}{\gamma n}(1 + CT(\gamma, k, p, 1, k)).$$

Now consider the event

$$\mathcal{T}(m^*, \psi^*) \quad := \quad \{\max_{i \in S}|m_i| \leq m^* \text{ and } \max_{i \in S}|\psi_i| \leq \psi^*\}.$$

Conditioning on $\mathcal{T}$ and its complement, we have

$$\begin{aligned}
\mathbb{P}[(\mathcal{E}(U))^c] \quad &= \quad \mathbb{P}[\frac{1}{\beta^*_{\min}} \max_{i \in S} U_i| > 1] \\
&\leq \quad \mathbb{P}[\frac{1}{\beta^*_{\min}} \max_{i \in S} |U_i| > 1 \;|\; \mathcal{T}(m^*, \psi^*)] + \mathbb{P}[(\mathcal{T}(m^*, \psi^*))^c].
\end{aligned} \tag{29}$$

In order to upper bound $\mathbb{P}[(\mathcal{T}(m^*, \psi^*))^c]$, we first upper bound the terms $\mathbb{P}(|m_i| > m^*)$ and $\mathbb{P}(|\psi_i| > \psi^*)$, and then apply the union bound. Beginning with the mean, we have

$$\begin{aligned}
|m_i| \quad &:= \quad \rho_n \big| e_i^T \big(\frac{1}{n}X_S^T X_S\big)^{-1}\vec{1}\big| \\
&= \quad \rho_n \big| e_i^T \big[\big(\frac{1}{n}X_S^T X_S\big)^{-1} - I_{k,k}\big]\vec{1} + e_i^T I_{k,k}\vec{1}\big| \\
&\leq \quad \rho_n \big| e_i^T \big[\big(\frac{1}{n}X_S^T X_S\big)^{-1} - I_{k,k}\Big\}\vec{1}\big| + \rho_n.
\end{aligned}$$

Our next step is to upper bound the operator norm of the matrix within curly braces, which we do by applying Lemma 10 from Section D, with the parameters $\theta = 1$ and $t = k$. We conclude there is an universal constant C such that $|||(\frac{1}{n}X_S^T X_S)^{-1} - I_{k,k}|||_2 \leq CT(\gamma, k, p, 1, k)$ with probability at least $1 - O(k^{-2})$. Consequently, if we define $m^* := \rho_n\big[\sqrt{k}CT(\gamma, k, p, 1, k)\big] + \rho_n$, then we have the bound

$$\mathbb{P}\big[|m_i| \geq m^*\big] \quad = \quad O(1/k^2).$$

A similar argument can be used to bound each term ψ_i, thereby obtaining

$$\mathbb{P}[\big|\psi_i\big| > \psi^*] \quad = \quad O(k^{-2}).$$

Since there are k versions of each of m_i and ψ_i, the union bound implies that

$$\mathbb{P}[(\mathcal{T}(m^*, \psi^*))^c] \leq 2kO(k^{-2}) = O(k^{-1}).$$

We now turn to the first term of Equation (29). Letting $Y_i \sim \mathcal{N}(0, \psi_i)$, and using $\mathcal{T}$ as shorthand for the event $\mathcal{T}(m^*, \psi^*)$, we have

$$\begin{aligned}
\mathbb{P}[\frac{1}{\beta^*_{\min}} \max_{i \in S} |U_i| > 1 \ \mid \ \mathcal{T}] &= \mathbb{E}\big\{\mathbb{P}\big[\max_{i \in S} |U_i| > \beta^*_{\min} \ \mid \ X_S, \mathcal{T}\big]\big\} \\
&\leq \mathbb{E}\big\{\mathbb{P}\big[\max_{i \in S} \big(|m_i| + |Y_i|\big) > \beta^*_{\min} \ \mid \ X_S, \mathcal{T}\big]\big\} \\
&\leq \mathbb{E}\big\{\mathbb{P}\big[m^* + \max_{i \in S} |Y_i| > \beta^*_{\min} \ \mid \ X_S, \mathcal{T}\big]\big\} \\
&= \mathbb{E}\big\{\mathbb{P}\big[\frac{1}{\beta^*_{\min}} \max_{i \in S} |Y_i| > 1 - \frac{m^*}{\beta^*_{\min}} \ \mid \ X_S, \mathcal{T}\big]\big\}.
\end{aligned}$$

For sufficiently large p and k, we have

$$\begin{aligned}
T(\gamma, k, p, 1, k) &= \frac{1}{\gamma}\sqrt{\max\big\{\frac{\log(k)}{k \log(p-k)}, \frac{\log[\log(p-k)]}{\log(p-k)}\big\}} \\
&\leq \frac{1}{\gamma}\sqrt{\big\{\frac{\log[\log(p-k)]}{\log(p-k)}\big\}},
\end{aligned}$$

using the facts that $\frac{\log(k)}{k} \to 0$ and $\log[\log(p-k)] \to \infty$, so that the maximum is dominated by the second term. As a result, applying condition (9) yields that $\frac{m^*}{\beta^*_{\min}} \to 0$. Letting $Y^* \sim \mathcal{N}(0, \psi^*)$, we have

$$\begin{aligned}
\mathbb{E}\big\{\mathbb{P}\big[\frac{1}{\beta^*_{\min}} \max_{i \in S} |Y_i| > \frac{1}{2} \ \mid \ X_S, \mathcal{T}\big]\big\} &\leq \mathbb{E}\big\{k\, \mathbb{P}[|Y^*| \geq \frac{\beta^*_{\min}}{2} \ \mid \ X_S, \mathcal{T}]\big\} \\
&\leq 2k \, \exp\big(-\frac{[\beta^*_{\min}]^2}{8\psi^*}\big),
\end{aligned}$$

where the last inequality follows from Gaussian tail bounds (see Appendix A). It remains to verify that this final term converges to zero. Taking logarithms and ignoring constant terms, we have

$$\log(k)(1 - \frac{[\beta^*_{\min}]^2}{\log(k)\ 8\psi^*}) = \log(k)\big(1 - \frac{[\beta^*_{\min}]^2 \gamma n}{8 \log(k)\ (1 + CT(\gamma, k, p, 1, k))}\big).$$

Our goal is to show that that this quantity diverges to $-\infty$. Condition (10) implies that

$$T(\gamma, k, p, 1, k) = \frac{1}{\gamma}\sqrt{\max\big\{\frac{\log(k)}{k \log(p-k)}, \frac{\log\log(p-k)}{\log(p-k)}\big\}} \to 0.$$

Hence, it suffices to show that $\log k \ \big(1 - \frac{[\beta^*_{\min}]^2 \gamma n}{16 \log k}\big)$ diverges to $-\infty$. We have

$$\begin{aligned}
\log(k)\ \big(1 - \frac{[\beta^*_{\min}]^2 \gamma n}{16 \log(k)}\big) &= \log(k)\ (1 - \frac{[\beta^*_{\min}]^2}{\rho_n^2} \frac{\gamma n \rho_n^2}{16 \log(k)}) \\
&= \log(k)\ (1 - \frac{[\beta^*_{\min}]^2}{\rho_n^2} \frac{\gamma n \rho_n^2}{16 \log(p-k)} \frac{\log(p-k)}{\log(k)}).
\end{aligned}$$

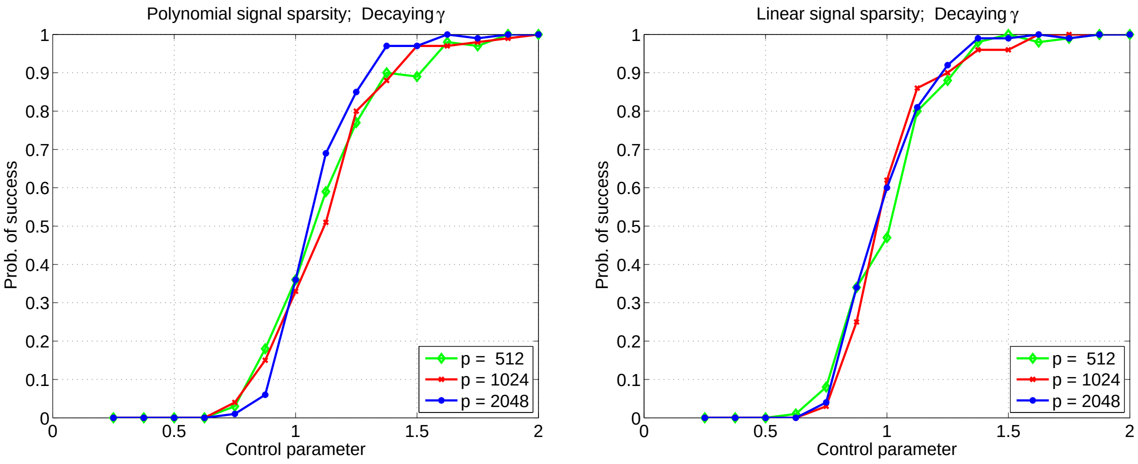
Condition (9) implies that $\frac{[\beta^*_{\min}]^2}{\rho_n^2} \to \infty$, whereas condition (8) ensures that $\frac{\gamma n \rho_n^2}{\log(p-k)} \to \infty$. The signal class $\mathcal{C}(p, k, \beta^*_{\min})$, as previously defined (1), ensures that $k \leq \frac{p}{2}$, so that the third term is greater than one. Putting together the pieces, we conclude that $\mathbb{P}[\mathcal{E}(U)^c]$ tends to zero.

4. Experimental Results

In this section, we provide some experimental results to illustrate the claims of Theorem 1. We consider two different sparsity regimes, namely linear sparsity ($k = \alpha p$) and polynomial sparsity ($k = \sqrt{p}$), and we show simulations in which the fraction γ of non-zero entries in each row of the measurement matrix X converges to zero. For all experiments, the additive noise standard deviation is set to $\sigma = 0.25$ and we fix the vector β^* by setting the first k entries are set to one, and the remaining entries to zero. There is no loss of generality in fixing the support in this way, since the ensemble of models is invariant under permutations.

Although it is possible to solve the Lasso using a variety of methods, our theory (in particular, Lemma 2) shows that it suffices to simulate the random variables $\{V_j^a, V_j^b, j \in S^c\}$ and $\{U_i, i \in S\}$, and then check the equivalent conditions (21) and (22). (These necessary and sufficient conditions give the same result for support recovery as solving the Lasso; however, they are much faster to simulate.) In all cases, we plot the success probability $\mathbb{P}[S(\widehat{\beta}) = S(\beta^*)]$ versus the *control parameter* $\theta(n,p,k) = \frac{n}{2k\log(p-k)}$. Note that Theorem 1 predicts that the Lasso should transition from failure to success at $\theta \approx 1$.

In Figure 1, the empirical success rate of the Lasso is plotted against the control parameter $\theta(n,p,k) = \frac{n}{2k\log(p-k)}$. Each panel shows three curves, corresponding to the problem sizes $p \in \{512, 1024, 2048\}$, and each point on the curve represents the average of 100 trials. For all trials shown, we set $\gamma = 0.5\frac{\log(p-k)}{\sqrt{p-k}}$, which converges to zero at a rate slightly faster than that guaranteed by Theorem 1. Nonetheless, we still observe the ”stacking” behavior around the predicted threshold $\theta^* = 1$.



(a) (b)

Figure 1: Plots of the success probability $\mathbb{P}[\widehat{S} = S]$ versus the control parameter $\theta(n,p,k) = \frac{n}{k\log(p-k)}$ for γ-sparsified ensembles, with decaying measurement sparsity $\gamma = \frac{.5\log(p-k)}{\sqrt{p-k}}$. (a) Polynomial signal sparsity $k = O(\sqrt{p})$. (b) Linear signal sparsity $k = \Theta(p)$.

5. Discussion

In this paper, we have studied the problem of recovery the support set of a sparse vector β^* based on noisy observations. The main result is to show that it is possible to "sparsify" standard dense measurement matrices, so that they have a vanishing fraction of non-zeroes per row, while retaining the same sample complexity (number of observations n) required for exact recovery. We also showed that under the support recovery metric and in the presence of noise, no method can succeed without the number of non-zeroes per column tending to infinity, so that our results cannot be improved substantially. Thus, our results show that it is possible to use sparse measurement matrices while retaining the same guarantees regarding the recovery of the support. Note that our sparsification scheme is the simplest one, and requires no additional overhead to implement. Although this paper focused on sparsified Gaussian measurement matrices, it is possible to obtain qualitatively similar results for sparsified sub-Gaussian ensembles (for instance, the discrete uniform distribution on $\{-1,1\}$).

The approach taken in this paper is to find rates which γ (as a function of n, p, k) can safely tend towards zero while maintaining the same statistical efficiency as dense random matrices. In various practical settings (Wakin et al., 2006), it may be preferable to make the measurement ensembles even sparser at the cost of taking more measurements n, thereby decreasing statistical efficiency relative to dense random matrices. A natural question is the sample complexity $n(\gamma,p,k)$ in this regime as well. Finally, this work has focused only on a randomly sparsified matrices, as opposed to particular sparse designs (e.g., based on LDPC or expander-type constructions Feldman et al., 2007; Sarvotham et al., 2006; Xu and Hassibi, 2007). Although our results imply that exact support recovery with noisy observations is impossible with bounded degree designs, it would be interesting to examine the trade-off between other loss functions (e.g., ℓ_2 reconstruction error) and sparse measurement matrices.

Acknowledgments

This work was partially supported by NSF grants CAREER-CCF-0545862 and DMS-0605165, a Vodafone-US Foundation Fellowship (DO), and a Sloan Foundation Fellowship (MJW).

Appendix A. Standard Concentration Results

In this appendix, we collect some tail bounds used repeatedly throughout this paper.

Lemma 7 (Hoeffding bound—Hoeffding, 1963) *Given a binomial variate $Z \sim \mathrm{Bin}(n,\gamma)$, we have for any $\delta > 0$*

$$\mathbb{P}[|Z-\gamma n| \geq \delta n] \quad \leq \quad 2\exp\big(-2n\delta^2\big).$$

Lemma 8 (χ^2-concentration—Johnstone, 2001) *Let $X \sim \chi^2_m$ be a chi-squared variate with m degrees of freedom. Then for all $\frac{1}{2} > \delta \geq 0$, we have*

$$\mathbb{P}[X-m \geq \delta m] \quad \leq \quad \exp\big(-\frac{3}{16}m\delta^2\big).$$

We will also find the following standard Gaussian tail bound (e.g., Ledoux and Talagrand, 1991) useful:

Lemma 9 (Gaussian tail behavior) *Let $V \sim \mathcal{N}(0,\sigma^2)$ be a zero-mean Gaussian with variance σ^2. Then for all $\delta > 0$, we have*

$$\mathbb{P}[|V| > \delta] \quad \leq \quad 2\, \exp\big(-\frac{\delta^2}{2\sigma^2}\big).$$

Appendix B. Convex Optimality Conditions

In this section we discuss the optimality conditions that the Lasso must satisfy and some implications that follow.

B.1 Proof of Lemma 2

Let $f(\beta) := \frac{1}{2n}\|Y - X\beta\|_2^2 + \rho_n\|\beta\|_1$ denote the objective function of the Lasso (4). By standard convex optimality conditions (Rockafellar, 1970), a vector $\widehat{\beta} \in \mathbb{R}^p$ is a solution to the Lasso if and only if $0 \in \mathbb{R}^p$ is an element of the subdifferential of $f(\beta)$ at $\widehat{\beta}$. These conditions lead to

$$\frac{1}{n}X^T(X\widehat{\beta} - Y) + \rho_n \widehat{z} \quad = \quad 0,$$

where the dual vector $\widehat{z} \in \mathbb{R}^p$ is an element of the subdifferential of the ℓ_1-norm, given by

$$\partial\|\widehat{\beta}\|_1 \quad = \quad \big\{z \in \mathbb{R}^p \;\mid\; z_i = \text{sign}(\widehat{\beta}_i) \text{ if } \widehat{\beta}_i \neq 0, \qquad z_i \in [-1,1] \text{ otherwise}\big\}.$$

Now suppose that we are given a pair $(\widehat{\beta}, \widehat{z}) \in \mathbb{R}^p \times \mathbb{R}^p$ that satisfy the assumptions of Lemma 2. Condition (13) is equivalent to $(\widehat{\beta}, \widehat{z})$ satisfying the zero subgradient condition. Conditions (14), (16) and (17) ensure that $\widehat{z}$ is an element of the subdifferential of the ℓ_1-norm at $\widehat{\beta}$. Finally, conditions (15) and (17) ensure that $\widehat{\beta}$ correctly specifies the signed support.

It remains to verify that $\widehat{\beta}$ is the *unique* optimal solution. By Lagrangian duality, the Lasso problem (4) (given in penalized form) can be written as an equivalent constrained optimization problem over the ball $\|\beta\|_1 \leq C(\rho_n)$, for some constant $C(\rho_n) < +\infty$. Equivalently, we can express this single ℓ_1-constraint as a set of 2^p linear constraints $\vec{v}^T\beta \leq C$, one for each sign vector $\vec{v} \in \{-1,+1\}^p$. The vector $\widehat{z}$ can be written as a convex combination $\widehat{z} = \sum_{\vec{v}} \alpha^*_{\vec{v}}\vec{v}$, where the weights $\alpha^*_{\vec{v}}$ are non-negative and sum to one. By construction of $\widehat{\beta}$ and $\widehat{z}$, the weights α^* form an optimal Lagrange multiplier vector for the problem. Consequently, any other optimal solution—say $\widetilde{\beta}$—must also minimize the associated Lagrangian

$$L(\beta;\alpha^*) \quad = \quad f(\beta) + \sum_{\vec{v}} \alpha^*_{\vec{v}}\big[\vec{v}^T\beta - C\big],$$

and satisfy the complementary slackness conditions $\alpha^*_{\vec{v}}\big(\vec{v}^T\widetilde{\beta} - C\big) = 0$ for every $\vec{v}$.

Note that these complementary slackness conditions imply that $\widehat{z}^T\widetilde{\beta} = C$. But this can only happen if $\widetilde{\beta}_j = 0$ for all indices where $|\widehat{z}_j| < 1$. Therefore, any optimal solution $\widetilde{\beta}$ satisfies $\widetilde{\beta}_{S^c} = 0$. Finally, given that all optimal solutions satisfy $\beta_{S^c} = 0$, we may consider the restricted optimization problem subject to this set of constraints. If the Hessian submatrix $\widehat{\Sigma}_{SS}$ is strictly positive definite, then this sub-problem is strictly convex, so that $\widehat{\beta}$ must be the unique optimal solution, as claimed.

B.2 Derivation of $\{V_j^a, V_j^b, U_i\}$

In this appendix, we derive the form of the $\{V_j^a, V_j^b\}$ and $\{U_i\}$ variables defined in Equations (18) through (20). We begin by writing the zero sub-gradient condition in a block-form, and substituting the relations specified in conditions (14) and (15):

$$\begin{bmatrix} \widehat{\Sigma}_{SS} & \widehat{\Sigma}_{SS^c} \\ \widehat{\Sigma}_{S^cS} & \widehat{\Sigma}_{S^cS^c} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_S - \beta^*_S \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} X_S^T W \\ \frac{1}{n} X_{S^c}^T W \end{bmatrix} + \rho_n \begin{bmatrix} \operatorname{sign}(\beta^*_S) \\ \widehat{z}_{S^c} \end{bmatrix} \quad = \quad 0.$$

By solving the top block, we obtain

$$U := \widehat{\beta}_S - \beta^*_S \quad = \quad \widehat{\Sigma}_{SS}^{-1} \big\{ \frac{1}{n} X_S^T W - \rho_n \operatorname{sign}(\beta^*_S) \big\}.$$

By back-substituting this relation into the lower block, we can solve explicitly for $\widehat{z}_{S^c}$; doing so yields that $\rho_n \widehat{z}_{S^c} = V^a + V^b$, where the $(p-k)$-vectors are defined in Equations (18) and (19).

Appendix C. Proof of Lemma 6

Let $Z \in \mathbb{R}^{n \times n}$ denote a $n \times n$ matrix, for which the off-diagonal elements $Z_{ij} = 0$ for all $i \neq j$, and the diagonal elements $Z_{ii} \sim \operatorname{Ber}(\gamma)$ are i.i.d. With this notation, we can write $H \overset{d}{=} Zh$. Using the definition (23) of h, we have

$$\begin{aligned} \|H\|_2^2 \quad &= \quad \|Zh\|_2^2 \\ &= \quad \|Z \frac{X_S}{n} (\widehat{\Sigma}_{SS})^{-1} \vec{1}\|_2^2 \\ &= \quad \vec{1}^T (\widehat{\Sigma}_{SS})^{-1} (Z \frac{X_S}{n})^T (Z \frac{X_S}{n}) (\widehat{\Sigma}_{SS})^{-1} \vec{1} \\ &= \quad \frac{\gamma}{n} \vec{1}^T (\widehat{\Sigma}_{SS})^{-1} \underbrace{\big\{ \frac{1}{\gamma n} \sum_{i=1}^n \mathbb{I}[Z_{ii} = 1] \, x_i x_i^T \big\}}_{\Gamma(Z)} (\widehat{\Sigma}_{SS})^{-1} \vec{1}, \end{aligned}$$

where x_i is the i^{th} row of the matrix X_S. We can apply Lemma 10 from Appendix D with parameters $\theta = 1$ and $t = (p-k)$ yielding

$$\mathbb{P}\big[\|\!|\widehat{\Sigma}_{SS}^{-1}\|\!|_2 \geq f_1(p,k,\gamma) \big] \quad \leq \quad O\big(\frac{1}{(p-k)^2} \big), \tag{30}$$

where $f_1(p,k,\gamma) := 1 + T(\gamma, k, p, 1, p-k)$, and the function $T(\gamma, k, p, \theta, t)$ was defined in Equation (28).

Next we control the spectral norm of the random matrix $\Gamma(Z)$, conditioned on the total number $\sum_{i=1}^n Z_{ii}$ of non-zero entries. In particular, applying Lemma 10 with $t = p-k$, and $\theta = \frac{z}{n}$, we have

$$\mathbb{P}\big[\|\Gamma(Z)\|_2 \geq \frac{z}{n\gamma} \big[1 + T(\gamma, k, p, \frac{z}{n}, p-k) \big] \;\big|\; \sum_{i=1}^n Z_{ii} = z \big] \quad \leq \quad \frac{1}{(p-k)^2}, \tag{31}$$

as long as $k \frac{z}{n} \to \infty$.

The next step is to deal with the conditioning. Define the event

$$\mathcal{T}(k,\gamma) \quad := \quad \big\{ Z \;\mid\; \gamma - \frac{1}{2\sqrt{k}} \leq \frac{1}{n}\sum_{i=1}^{n} Z_{ii} \leq \gamma + \frac{1}{2\sqrt{k}} \big\}.$$

We need to find an upper bound on $\| \Gamma(Z) \|_2$ that will hold with high probability for all Z that satisfy the above property. One function that suffices is

$$f_2(p,k,\gamma) \quad := \quad \left(1+\frac{1}{2\sqrt{k}\gamma}\right) \left[1 + C\Big(\frac{1}{\gamma}\sqrt{\max\left\{\frac{1}{k(\gamma-\frac{1}{2\sqrt{k}})}, \frac{\log\left[(\gamma+\frac{1}{2\sqrt{k}})\log(p-k)\right]}{(\gamma-\frac{1}{2\sqrt{k}})\log(p-k)}\right\}}\Big)\right].$$

We then have that

$$\begin{aligned} \mathbb{P}[|\!|\!|\Gamma(Z)|\!|\!|_2 \geq f_2(p,k,\gamma)] \quad &\leq \quad \mathbb{P}[|\!|\!|\Gamma(Z)|\!|\!|_2 \geq f_2(p,k,\gamma) \;\mid\; \mathcal{T}(k,\gamma)] + \mathbb{P}[(\mathcal{T}(k,\gamma))^c] \\ &\leq \quad \exp(-2\log(p-k)) + 2\exp(-\frac{n}{2k}) \\ &\leq \quad 3\exp(-\min\{2\log(p-k), \frac{n}{2k}\}), \end{aligned} \qquad (32)$$

where we have used the bound (31), and the Hoeffding bound (see Lemma 7).

Combining the bounds (30) and (32), we conclude that as long as $\gamma k \to \infty$, then:

$$\mathbb{P}\big[|\!|\!|\widehat{\Sigma}^{-1}\Gamma(Z)\widehat{\Sigma}^{-1}|\!|\!|_2 \geq f_1^2 f_2\big] \quad \leq \quad 4\exp(-\min\{2\log(p-k), \frac{n}{2k}\}).$$

Since $\|\vec{1}\|_2 = \sqrt{k}$, we have

$$\mathbb{P}[\|H\|_2^2 \geq \frac{\gamma k}{n} f_1^2 f_2] \quad \leq \quad 4\exp(-\min\{2\log(p-k), \frac{n}{2k}\}).$$

To conclude the proof, we must show that both $f_1(p,k,\gamma)$ and $f_2(p,k,\gamma)$ converge to 1 as (p,k,γ) scale. The term $f_1(p,k,\gamma) = 1 + T(\gamma,k,p,1,p-k)$ converges to one, since the quantity

$$T(\gamma,k,p,1,p-k) \quad = \quad \frac{1}{\gamma}\sqrt{\max\{\frac{1}{k}, \frac{\log[\log(p-k)]}{\log(p-k)}\}}$$

converges[1] to zero under assumption (10). Next, we need to demonstrate that $f_2(p,k,\gamma)$ converge to 1 as (p,k,γ) scale. Since assumption (10) ensures that $\gamma\sqrt{k} \to \infty$, it suffices to study the simpler function

$$f_3(p,k,\gamma) \quad := \quad 1 + C\Big(\frac{1}{\gamma}\sqrt{\max\left\{\frac{1}{k\gamma}, \frac{\log\left[\gamma\log(p-k)\right]}{\gamma\log(p-k)}\right\}}\Big),$$

which has the same asymptotic behavior as $f_2(p,k,\gamma)$. Observe that $f_3(p,k,\gamma)$ satisfies the sandwich relation

$$1 \;\leq\; f_3(p,k,\gamma) \quad \leq \quad 1 + C\Big(\sqrt{\max\{\frac{1}{k\gamma^3}, \frac{\log\left[\log(p-k)\right]}{\gamma^3\log(p-k)}\}}\Big),$$

By assumption (10), this upper bound converges to one, showing that f_3 and hence f_2 converge to one, as desired. In particular, for any fixed $\delta > 0$, we have $f_1^2 f_2 < (1+\delta)$ for p,k sufficiently large, thereby completing the proof of Lemma 6.

[1] In particular, the left-hand side of the expression (10) satisfies $\frac{\gamma}{[T(\gamma,k,p,1,p-k)]^2} \leq \frac{1}{[T(\gamma,k,p,1,p-k)]^2}$.

Appendix D. Singular Values of Sparsified Matrices

Let $\theta(p,k) \in (0,1]$ and $t(p,k) \in \{1,2,3,\ldots\}$ be functions. Let X be an $\theta n \times k$ random matrix with i.i.d. entries X_{ij} distributed according to the γ-sparsified ensemble (6). Recall the definition of the function $T(\gamma,k,p,\theta,t)$ defined in Equation (28), and let $t > 0$ be arbitrary.

Lemma 10 *Suppose that $n \geq (2+\nu)k\log(p-k)$ for some $\nu > 0$. If as k and $p-k \to \infty$, we have $T(\gamma,k,p,\theta,t) \longrightarrow 0$, then there are universal positive constants C_i such that*

$$\mathbb{P}\big[\sup_{\|u\|_2=1}\big|\frac{1}{\sqrt{\theta n}}\|Xu\|_2 - 1\big| \geq C_1\,T(\gamma,k,p,\theta,t)\big] \;=\; O(\frac{1}{t^2}), \quad \textit{and} \tag{33}$$

$$\mathbb{P}\big[|\|(\frac{1}{\theta n}X^TX)^{-1} - I_{k\times k}\|_2 \geq C_2\,T(\gamma,k,p,\theta,t)\big] \;=\; O(\frac{1}{t^2}). \tag{34}$$

Remark 3 *(a) Note that Equation* (33) *implies that the eigenvalues of the matrix $\frac{1}{\theta n}X^TX$ are contained in the interval $(1-C_1T(\gamma,k,p,\theta,t), 1+C_1T(\gamma,k,p,\theta,t))$. Since $T(\gamma,k,p,\theta,t) = o(1)$ by assumption, we can always find a constant C_2 such that the eigenvalues of the inverted matrix $(\frac{1}{\theta n}X^TX)^{-1}$ are contained in the interval $(1-C_2T(\gamma,k,p,\theta,t), 1+C_2T(\gamma,k,p,\theta,t))$. Consequently, Equation* (34) *is a consequence of the assumptions of Lemma 10 and Equation* (33)*.*

(b) In addition, observe that Lemma 10 with $\theta = 1$ and $t = p-k$ implies that $\widehat{\Sigma} = \frac{1}{n}X_S^TX_S$ is invertible with probability greater than $1-O(\frac{1}{(p-k)^2})$, there establishing Lemma 1. Other settings in which this lemma is applied are $(\theta,t) = (\gamma,p-k)$ and $(\theta,t) = (1,k)$.

The remainder of this section is devoted to the proof of Lemma 10.

D.1 Bounds on Expected Values

Let $X \in \mathbb{R}^{\theta n \times k}$ be a random matrix with i.i.d. entries from the γ-sparsified ensemble

$$X_{ij} \quad \sim \quad (1-\gamma)\delta_X(0) + \gamma\mathcal{N}(0,\frac{1}{\gamma}).$$

Note that $\mathbb{E}[X_{ij}] = 0$ and $\text{var}(X_{ij}) = 1$ by construction.

We follow the proof technique outlined in the lecture notes (Vershynin, 2006). We first note the tail bound:

Lemma 11 *Let $Q_1,\ldots,Q_d$ be i.i.d. samples of the γ-sparsified Gaussian ensemble. Given any vector $a \in \mathbb{R}^d$ and $t > 0$, we have $\mathbb{P}[\sum_{i=1}^d a_iQ_i > t] \leq \exp\big(-\frac{\gamma t^2}{2\|a\|_2^2}\big)$.*

To establish this bound, note that each Y_i is dominated (stochastically) by the random variable $Z \sim \mathcal{N}(0,\frac{1}{\gamma})$. In particular, for any $\lambda > 0$ we have

$$\mathbb{M}_{Q_i}(\lambda) \quad = \quad \mathbb{E}[\exp(\lambda Q_i)] \;=\; (1-\gamma) + \gamma\mathbb{E}[\exp(\lambda Z)] \leq \exp(\lambda^2/2\gamma),$$

from which the claim follows by optimizing the Chernoff bound.

Now let us bound the maximum singular value $s_{\max}(X)$ of the random matrix X. Letting S^{d-1} denote the ℓ_2 unit ball in d dimensions, we begin with the variational representation

$$\begin{aligned} s_{\max}(X) &= \max_{u \in S^{k-1}} \|Xu\| \\ &= \max_{v \in S^{\theta n-1}} \max_{u \in S^{k-1}} v^T X u. \end{aligned}$$

For an arbitrary $\varepsilon \in (0,1)$, we can find ε-covers (in ℓ_2 norm) of $S^{\theta n-1}$ and S^{k-1} with $M_{\theta n}(\varepsilon) = (3/\varepsilon)^{\theta n}$ and $M_k(\varepsilon) = (3/\varepsilon)^k$ points respectively (Matousek, 2002). Denote these covers by $C_{\theta n}(\varepsilon)$ and $C_k(\varepsilon)$ respectively. A standard argument shows that for all $\varepsilon \in (0,1)$, we have

$$\|X\|_2 \;\leq\; \frac{1}{(1-\varepsilon)^2} \max_{u_\alpha \in C_k(\varepsilon)} \max_{v_\beta \in C_{\theta n}(\varepsilon)} v_\beta^T X u_\alpha.$$

Let us analyze the maximum on the RHS: for a fixed pair (u,v) in our covers, we have

$$u^T X v \;=\; \sum_{i=1}^{k} \sum_{j=1}^{\theta n} X_{ij} u_i v_j.$$

Let us apply Lemma 11 with $d = \theta nk$, and weights $a_{ij} = u_i v_j$. Note that we have

$$\|a\|_2^2 \;=\; \sum_{i,j} a_{ij}^2 \;=\; \sum_i u_i^2 (\sum_j v_j^2) \;=\; 1,$$

since each u and v are unit norm. Consequently, for any fixed u,v in the covers, we have

$$\mathbb{P}[u^T X v > t] \;\leq\; \exp\big(-\frac{\gamma t^2}{2}\big).$$

By the union bound, we have

$$\begin{aligned} \mathbb{P}\big[\max_{u_\alpha \in C_k(\varepsilon)} \max_{v_\beta \in C_{\theta n}(\varepsilon)} v_\beta^T X u_\alpha > t\big] &\leq M_k(\varepsilon) M_{\theta n}(\varepsilon) \exp\big(-\frac{\gamma t^2}{2}\big) \\ &\leq \exp\big((k+\theta n)\log(3/\varepsilon) - \frac{\gamma t^2}{2}\big). \end{aligned}$$

By choosing $\varepsilon = \frac{1}{2}$ and $t = \sqrt{\frac{4}{\gamma}(k+\theta n)\log 6}$, we can conclude that

$$s_{\max}(X)/\sqrt{\theta n} \;=\; \|X\|_2/\sqrt{\theta n} \;\leq\; C\sqrt{\frac{1}{\gamma}}\sqrt{1+\frac{k}{\theta n}},$$

with probability at least $1-\exp(-(k+\theta n)\log 6)$. Note that

$$\frac{k}{\theta n} \;=\; O\big(\frac{1}{(2+\nu)\theta\log(p-k)}\big) \to 0,$$

since $\frac{\log[\theta\log(p-k)]}{\theta\log(p-k)} \to 0$, which implies that $\theta\log(p-k) \to \infty$.

Consequently, we can conclude that

$$\|X\|_2/\sqrt{\theta n} \;\leq\; O(1/\sqrt{\gamma}),$$

w.p. one as $\theta n, k \to \infty$. Although this bound is essentially correct for a $\mathcal{N}(0,\frac{1}{\gamma})$ ensemble with γ *fixed*, it is very crude for the sparsified case with $\gamma \to 0$, but will useful in obtaining tighter control on $s_{\max}(X)$ and $s_{\min}(X) := \min_{u \in S^{k-1}} \|Xu\|$ in the sequel.

D.2 Tightening the Bound

For a given $u \in S^{k-1}$, consider the random variable $\|Xu\|_2^2 := \sum_{i=1}^{\theta n}(Xu)_i^2$. We first claim that each variate $Z_i = (Xu)_i^2$ is subexponential, or more precisely:

Lemma 12 *For any $s > 0$, we have $\mathbb{P}[Z_i > s] \leq 2\exp\left(-\frac{\gamma s}{2}\right)$.*

Proof We can write $(Xu)_i = \sum_{j=1}^{k} X_{ij}u_j$ where $\|u\|_2 = 1$. Consequently, Lemma 11 implies that $\mathbb{P}[\sum_{j=1}^{k} X_{ij}u_j > \delta] \leq \exp(-\frac{\gamma\delta^2}{2})$. By symmetry, we have

$$\mathbb{P}[Z_i > s] = \mathbb{P}[|\sum_{j=1}^{k} X_{ij}u_j| > \sqrt{s}] \;\leq\; 2\exp(-\frac{\gamma s}{2}),$$

from which the claim follows. ■

Now consider the event

$$\left\{\left|\frac{\|Xu\|_2^2}{\theta n} - 1\right| > \delta\right\} \quad = \quad \left\{\left|\sum_{i=1}^{\theta n} Z_i - \mathbb{E}[\sum_{i=1}^{\theta n} Z_i]\right| > \delta\theta n\right\}.$$

Let us apply Theorem 1.4 from Vershynin (2000) with $a = 2$, $b = 8\theta n/\gamma^2$ and $d = 2/\gamma$. With these choices, we have $4b/d = 16\theta n/\gamma$, which grows at least linearly in θn. Consequently, for any $\delta > 0$ less than $16/\gamma$ (we will in fact take $\delta \to 0$), we have

$$\mathbb{P}\left[\left|\frac{\|Xu\|_2^2}{\theta n} - 1\right| > \delta\right] \quad \leq \quad 2\exp\left(-\frac{\delta^2(\theta n)^2}{256\theta n/\gamma^2}\right) \;=\; 2\exp\left(-\frac{\gamma^2\,\delta^2\theta n}{256}\right).$$

Now take an ε-cover of the k-dimensional ℓ_2 ball, say with $N(\varepsilon) \;=\; (3/\varepsilon)^k$ elements. By the union bound, we have

$$\mathbb{P}\left[\inf_{i=1,\ldots,N(\varepsilon)} \frac{\|Xu_i\|_2^2}{\theta n} < 1-\delta\right] \quad \leq \quad \exp\left(-\frac{\gamma^2\,\delta^2\theta n}{256} + k\log(3/\varepsilon)\right).$$

Now set

$$\delta = \frac{\sqrt{2}}{\gamma}\sqrt{\frac{256 f(k,p)k\log(3/\varepsilon)}{\theta n}},$$

where $f(k,p) \geq 1$ is a function to be specified. Doing so yields that the infimum is bounded by $1+\delta$ with probability $1-\exp(-kf(k,p)\log(3/\varepsilon))$. (Note that the choice of $f(k,p)$ influences the rate of convergence, hence its utility.)

For any element $u \in S^{k-1}$, we have some u_i in the cover, and moreover

$$\begin{aligned}
\left|\|Xu\|^2 - \|Xu_i\|^2\right| \quad &= \quad \left|\left\{\|Xu\| - \|Xu_i\|\right\}\left\{\|Xu\| + \|Xu_i\|\right\}\right| \\
&\leq \quad \left|\left\{\|Xu\| - \|Xu_i\|\right\}\right|(2\|X\|) \\
&\leq \quad (\|X\|\,\|u-u_i\|)\;(2\|X\|) \;\leq\; 2\|X\|^2\varepsilon.
\end{aligned}$$

From our earlier result, we know that $\|X\|^2 = O(\theta n/\gamma)$ with probability $1-\exp(\log 6(k+\theta n))$. Putting together the pieces, we have there is a universal constant $C_2 > 0$, independent of $((\theta n), k, \gamma)$, such that the bound

$$\frac{1}{\theta n}\inf_{u\in S^{k-1}}\|Xu\|^2 \quad \geq \quad 1-\delta-C_3\varepsilon/\gamma \;=\; 1-\frac{2}{\gamma}\sqrt{\frac{32 f(k,p)k\log(3/\varepsilon)}{\theta n}}-\frac{C_2}{\gamma}\varepsilon,$$

holds with probability at least

$$1-\exp(-kf(k,p)\log(3/\varepsilon))-\exp(-\log 6(k+\theta n)). \tag{35}$$

Setting $\varepsilon = 3k/\theta n$ yields the bound

$$\frac{1}{\theta n}\inf_{u\in S^{k-1}}\|Xu\|^2 \quad \geq \quad 1-\frac{C_3}{\gamma}\sqrt{f(k,p)\frac{k}{\theta n}\log(\frac{\theta n}{k})},$$

where we have used the fact that $\frac{k}{\theta n} = o\Big(\sqrt{f(k,p)\frac{k}{\theta n}\log(\frac{\theta n}{k})}\Big)$. To understand how to choose the function $f(k,p)$, let us consider the rate of convergence (35). To establish the claim (33), we need rates fast enough to dominate a $2\log(t)$ term in the exponent, which guides our choice of $f(k,p)$. Recall that we are seeking to prove a scaling of the form $n = \Theta(k\log(p-k))$, so that this requirement (with $\varepsilon = 3k/\theta n = \frac{3}{\theta\log(p-k)}$) is equivalent to the quantity

$$kf(k,p)\log(3/\varepsilon)-2\log(t) \quad = \quad kf(k,p)\log[\theta\log(p-k)]-2\log(t) \tag{36}$$

tending to infinity.

D.2.1 CASE 1

If $k > \frac{\log(t)}{\log[\theta\log(p-k)]}$, then we may set $f(k,p) = 4$. With this choice, the condition (36) is satisfied, and we have

$$f(k,p)\frac{k}{\theta n}\log(\frac{\theta n}{k}) \quad = \quad 4\frac{\log[\theta\log(p-k)]}{\theta\log(p-k)} \;\to\; 0,$$

where we have used the assumption that $T(\gamma,k,p,\theta,t) = o(1)$.

D.2.2 CASE 2

Otherwise, if $k \leq \frac{\log(t)}{\log\theta\log(p-k)}$, then we may set

$$f(k,p) = 4\frac{\log(t)}{k\log\theta\log(p-k)} \geq 4,$$

so that condition (36) is satisfied, and we have

$$f(k,p)\frac{k}{\theta n}\log(\frac{\theta n}{k}) \quad \leq \quad 4\frac{\log(t)}{k\log\theta\log(p-k)}\frac{1}{\theta\log(p-k)}\log\theta\log(p-k) \;=\; \frac{4}{k}\frac{\log t}{\theta\log(p-k)} \to 0,$$

where we have again used the assumption $T(\gamma,k,p,\theta,t) = o(1)$ from Lemma 10.

Recalling the definition of $T(\gamma,k,p,\theta,t)$ from Equation (28), we can summarize both cases cleanly as

$$\mathbb{P}\left[\frac{1}{\theta n}\inf_{u\in S^{k-1}}\|Xu\|^2 \leq 1 - CT(\gamma,k,p,\theta,t)\right] \quad = \quad O(1/t^2).$$

For (p,k) sufficiently large, we have $CT(\gamma,k,p,\theta,t) < 1$, so that we can take square roots. Using the expansion $\sqrt{1+x} = 1 + \frac{x}{2} + o(x)$ for x small, we conclude that

$$\frac{1}{\sqrt{\theta n}}\inf_{u\in S^{k-1}}\|Xu\| \quad \geq \quad 1 - \frac{C}{2}T(\gamma,k,p,\theta,t) - o(T(\gamma,k,p,\theta,t)),$$

with probability greater than $1 - O(1/t^2)$. For (k,p) sufficiently large, the second term is smaller than $\frac{C}{4}T(\gamma,k,p,\theta,t)$, so that we conclude that

$$\mathbb{P}\left[\frac{1}{\sqrt{\theta n}}\inf_{u\in S^{k-1}}\|Xu\| \geq 1 - \frac{3C}{4}T(\gamma,k,p,\theta,t)\right] \quad \geq \quad 1 - O(1/t^2)$$

for (k,p) sufficiently large.

This same process can be repeated to bound the maximum singular value, yielding the bound

$$\mathbb{P}\left[\frac{1}{\sqrt{\theta n}}\sup_{u\in S^{k-1}}\|Xu\| \leq 1 + \frac{3C}{4}T(\gamma,k,p,\theta,t)\right] \quad \geq \quad 1 - O(1/t^2)$$

for (k,p) sufficiently large. Combining these two bounds yields the claim of Lemma 10.

References

D. Achlioptas. Database-friendly random projections. In *Proc. ACM Symp. Princ. Database Systems (PODS)*, pages 274–281, New York, USA, 2001. ACM.

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *STOC '96: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, New York, NY, USA, 1996. ACM Press.

R. Baraniuk, M. Davenport, R. Devore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx*, 2007.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

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

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

G. Cormode and S. Muthukrishnan. Towards an algorithmic theory of compressed sensing. Technical report, Rutgers University, July 2005.

S. Dasgupta. Learning mixtures of Gaussians. In *IEEE Symp. Foundations of Computer Science (FOCS)*, September 1999.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

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.

J. Feldman, T. Malkin, R. A. Servedio, C. Stein, and M. J. Wainwright. LP decoding corrects a constant fraction of errors. *IEEE Trans. Information Theory*, 53(1):82–89, January 2007.

A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin. Algorithmic linear dimension reduction in the ℓ_1-norm for sparse vectors. In *Proc. Allerton Conference on Communication, Control and Computing*, Allerton, IL, September 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, May 2006.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

I. M. 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.

I. M. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

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

P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, New York, NY, USA, 2006. ACM.

P. Li, T. J. Hastie, and K. W. Church. Nonlinear estimators and tail bounds for dimension reduction in l1 using cauchy random projections. *Journal of Machine Learning Research*, 8:2497–2532, 2007.

J. Matousek. *Lectures on Discrete Geometry*. Springer-Verlag, New York, 2002.

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

D. Paul. Asymptotics of sample eigenstructure for a large-dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

S. Sarvotham, D. Baron, and R. G. Baraniuk. Sudocodes: Fast measurement and reconstruction of sparse signals. In *Int. Symposium on Information Theory*, Seattle, WA, July 2006.

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

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

R. Vershynin. On large random almost euclidean bases. *Acta. Math. Univ. Comenianae*, LXIX: 137–144, 2000.

R. Vershynin. Random matrix theory. Technical report, UC Davis, 2006. Lecture Notes.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theor.*, 55(5):2183–2202, 2009.

M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk. An architecture for compressive imaging. *IEEE Int. Conf. Image Proc.*, 2006.

W. Wang, M. Garofalakis, and K. Ramchandran. Distributed sparse random projections for refinable approximation. In *Proc. International Conference on Information Processing in Sensor Networks*, Nashville, TN, April 2007.

W. Wang, M.J. Wainwright, and K. Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theor.*, 56(6):2967–2979, 2010.

W. Xu and B. Hassibi. Efficient compressive sensing with deterministic guarantees using expander graphs. *Information Theory Workshop, 2007. ITW '07. IEEE*, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

S. Zhou, J. Lafferty, and L. Wasserman. Compressed regression. In *Neural Information Processing Systems*, December 2007.