Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization

Sahand Negahban* Martin J. Wainwright†,*

Department of Statistics†, and
Department of Electrical Engineering and Computer Sciences*
UC Berkeley, Berkeley, CA 94720

May 5, 2009

Technical Report,
Department of Statistics, UC Berkeley

Abstract

Given a collection of $r \geq 2$ linear regression problems in p dimensions, suppose that the regression coefficients share partially common supports. This set-up suggests the use of ℓ_1/ℓ_∞-regularized regression for joint estimation of the $p \times r$ matrix of regression coefficients. We analyze the high-dimensional scaling of ℓ_1/ℓ_∞-regularized quadratic programming, considering both consistency rates in ℓ_∞-norm, and also how the minimal sample size n required for performing variable selection grows as a function of the model dimension, sparsity, and overlap between the supports. We begin by establishing bounds on the ℓ_∞-error as well sufficient conditions for exact variable selection for fixed design matrices, as well as designs drawn randomly from general Gaussian matrices. Our second set of results applies to $r = 2$ linear regression problems with standard Gaussian designs whose supports overlap in a fraction $\alpha \in [0, 1]$ of their entries: for this problem class, we prove that the ℓ_1/ℓ_∞-regularized method undergoes a phase transition—that is, a sharp change from failure to success—characterized by the rescaled sample size $\theta_{1,\infty}(n, p, s, \alpha) = n/\{(4 - 3\alpha)s \log(p - (2 - \alpha)\, s)\}$. More precisely, given sequences of problems specified by (n, p, s, α), for any $\delta > 0$, the probability of successfully recovering both supports converges to 1 if $\theta_{1,\infty}(n, p, s, \alpha) > 1 + \delta$, and converges to 0 for problem sequences for which $\theta_{1,\infty}(n, p, s, \alpha) < 1 - \delta$. An implication of this threshold is that use of ℓ_1/ℓ_∞-regularization yields improved statistical efficiency if the overlap parameter is large enough ($\alpha > 2/3$), but has *worse* statistical efficiency than a naive Lasso-based approach for moderate to small overlap ($\alpha < 2/3$). Empirical simulations illustrate the close agreement between these theoretical predictions, and the actual behavior in practice. These results indicate that some caution needs to be exercised in the application of ℓ_1/ℓ_∞ block regularization: if the data does not match its structure closely enough, it can impair statistical performance relative to computationally less expensive schemes.[1]

1 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

[1]This work was presented in part at the NIPS 2008 conference in Vancouver, Canada, December 2008. Supported in part by NSF grants DMS-0528488, DMS-0605165, and CCF-0545862.

to or larger than n, an active line of research is based on imposing structural conditions on the data (e.g., sparsity, manifold constraints, or graphical model structure), and studying the high-dimensional consistency (or inconsistency) of various types of estimators.

This paper deals with high-dimensional scaling in the context of solving multiple regression problems, where the regression vectors are assumed to have shared sparse structure. More specifically, suppose that we are given a collection of r different linear regression models in p dimensions, with regression vectors $\bar{\beta}^i \in \mathbb{R}^p$, for $i = 1, \ldots, r$. We let $S(\bar{\beta}^i) = \{j \mid \bar{\beta}^i_j \neq 0\}$ denote the support set of $\bar{\beta}^i$. In many applications—among them sparse approximation, graphical model selection, and image reconstruction—it is natural to impose a sparsity constraint, corresponding to restricting the cardinality $|S(\bar{\beta}^i)|$ of each support set. Moreover, one might expect some amount of overlap between the sets $S(\bar{\beta}^i)$ and $S(\bar{\beta}^j)$ for indices $i \neq j$ since they correspond to the sets of active regression coefficients in each problem. Let us consider some examples to illustrate:

- Consider the problem of image denoising or compression, say using a wavelet transform or some other type of multiresolution basis [17]. It is well known that natural images tend to have sparse representations in such bases [27]. Moreover, similar images—say the same scene taken from multiple cameras—would be expected to share a similar subset of active features in the reconstruction. Consequently, one might expect that using a block-regularizer that enforces such joint sparsity could lead to improved image denoising or compression.

- Consider the problem of identifying the structure of a Markov network or graphical model [10] based on a collection of samples (e.g., such as observations of a social network). For networks with a single parameter per edge (e.g., Gaussian models [19], Ising models [25]), a line of recent work has shown that ℓ_1-based methods can be successful in recovering the network structure. However, many graphical models have multiple parameters per edge (e.g., for discrete models with non-binary state spaces), and it is natural that the subset of parameters associated with a given edge are zero (or non-zero) in a grouped manner. Thus, any method for recovering the graph structure should impose a block-structured regularization that groups together the subset of parameters associated with a single edge.

- Finally, consider a standard problem in genetic analysis: given a set of gene expression arrays, where each array corresponds to a different patient but the same underlying tissue type (e.g., tumor), the goal is to discover the subset of features relevant for tumorous growths. This problem can be expressed as a joint regression problem, again with a shared sparsity constraint coupling together the different patients. In this context, the recent work of Liu et al. [14] shows that imposing additional structural constraints can be beneficial (e.g., they are able to greatly reduce the number of expressed genes while maintaining the same prediction performance).

Given these structural conditions of shared sparsity in these and other applications, it is reasonable to consider how this common structure can be exploited so as to increase the statistical efficiency of estimation procedures.

There is now a substantial and relatively mature body of work on ℓ_1-regularization for estimation of sparse models, dating back to the introduction of the Lasso and basis pursuit [28, 5]. With contributions from various researchers (e.g., [7, 19, 29, 37, 3]), there is now a fairly complete theory of the behavior of the Lasso for high-dimensional sparse estimation. A more

recent line of work (e.g., [31, 35, 22, 30, 36]), motivated by applications in which block or hierarchical structure arises, has proposed the use of block $\ell_{a,b}$ norms for various $a, b \in [1, \infty]$. Of particular relevance to this paper is the block ℓ_1/ℓ_∞ norm, proposed initially by Turlach et al. [31] and Tropp et al. [30]. This form of block regularization is a special case of the more general family of composite or hierarchical penalties, as studied by Zhao et al. [36].

Various authors have empirically demonstrated that block regularization schemes can yield better performance for different data sets [36, 22, 14]. Some recent work by Bach [1] has provided consistency results for ℓ_1/ℓ_2 block-regularization schemes under classical scaling, meaning that $n \to +\infty$ with p fixed. Meier et al. [18] has established high-dimensional consistency for the predictive risk of ℓ_1/ℓ_2 block-regularized logistic regression. The papers [15, 21, 24] have provided high-dimensional consistency results for ℓ_1/ℓ_q block regularization for support recovery using fixed design matrices, but the rates do not provide sharp differences between the case $q = 1$ and $q > 1$.

To date, there has a relatively limited amount of theoretical work characterizing if and when the use of block regularization schemes actually leads to gains in statistical efficiency. As we elaborate below, this question is significant due to the greater computational cost involved in solving block-regularized convex programs. In the case of ℓ_1/ℓ_2 regularization, concurrent work by Obozinski et al [23] (involving a subset of the current authors) has shown that that the ℓ_1/ℓ_2 method can yield statistical gains up to a factor of r, the number of separate regression problems; more recent concurrent work [9, 16] has provided related high-dimensional consistency results for ℓ_1/ℓ_2 regularization, emphasizing the gains when the number of tasks r is much larger than $\log p$.

This paper considers this issue in the context of variable selection using block ℓ_1/ℓ_∞ regularization. Our main contribution is to obtain some precise—and arguably surprising—insights into the benefits and dangers of using block ℓ_1/ℓ_∞ regularization, as compared to simpler ℓ_1-regularization (separate Lasso for each regression problem). We begin by providing a general set of sufficient conditions for consistent support recovery for both fixed design matrices, and random Gaussian design matrices. In addition to these basic consistency results, we then seek to characterize rates, for the particular case of standard Gaussian designs, in a manner precise enough to address the following questions:

(a) First, under what structural assumptions on the data does the use of ℓ_1/ℓ_∞ block-regularization provide a quantifiable reduction in the scaling of the sample size n, as a function of the problem dimension p and other structural parameters, required for consistency?

(b) Second, are there any settings in which ℓ_1/ℓ_∞ block-regularization can be harmful relative to computationally less expensive procedures?

Answers to these questions yield useful insight into the *tradeoff between computational and statistical efficiency* in high-dimensional inference. Indeed, the convex programs that arise from using block-regularization typically require a greater computational cost to solve. Accordingly, it is important to understand under what conditions this increased computational cost guarantees that fewer samples are required for achieving a fixed level of statistical accuracy.

The analysis of this paper gives conditions on the designs and regression matrix $\overline{B}$ for which ℓ_1/ℓ_∞ yields improvements (question (a)), and also shows that if there is sufficient mismatch between the regression matrix $\overline{B}$ and the ℓ_1/ℓ_∞ norm, then use of this regularizer actually impairs statistical efficiency relative to a naive ℓ_1-approach. As a representative instance of our theory, consider the special case of standard Gaussian design matrices and two regression

problems ($r = 2$), with the supports $S(\bar{\beta}^1)$ and $S(\bar{\beta}^2)$ each of size s and overlapping in a fraction $\alpha \in [0, 1]$ of their entries. For this problem, we prove that block ℓ_1/ℓ_∞ regularization undergoes a phase transition—meaning a *sharp threshold between success and recovery*—that is specified by the rescaled sample size

$$\theta_{1,\infty}(n, p, s, \alpha) \quad := \quad \frac{n}{(4 - 3\alpha)s \log(p - (2 - \alpha)s)}. \tag{1}$$

In words, for any $\delta > 0$ and for scalings of the quadruple (n, p, s, α) such that $\theta_{1,\infty} \geq 1 + \delta$, the probability of successfully recovering both $S(\bar{\beta}^1)$ and $S(\bar{\beta}^2)$ converges to one, whereas for scalings such that $\theta_{1,\infty} \leq 1 - \delta$, the probability of success converges to zero.

Figure 1 illustrates how the theoretical threshold (1) agrees with the behavior observed in practice. This figure plots the probability of successful recovery using the block ℓ_1/ℓ_∞ approach versus the rescaled sample size $n/\{2s \log[p - (2 - \alpha)s\}$; the results shown here are for $r = 2$ regression parameters. The plots show twelve curves, corresponding to three different problem sizes $p \in \{128, 256, 512\}$ and four different values of the overlap parameter $\alpha \in \{0.1, 0.3, 0.7, 1\}$. First, let us focus on the set of curves labeled with $\alpha = 1$, corresponding to case of complete overlap between the regression vectors. Notice how the curves for all three problem sizes p, when plotted versus the rescaled sample size, line up with one another; this "stacking effect" shows that the rescaled sample size captures the phase transition behavior. Similarly, for other choices of the overlap, the sets of three curves (over problem size p) exhibit the same stacking behavior. Secondly, note that the results are consistent with the theoretical prediction (1): the stacks of curves shift to the right as the overlap parameter α decreases

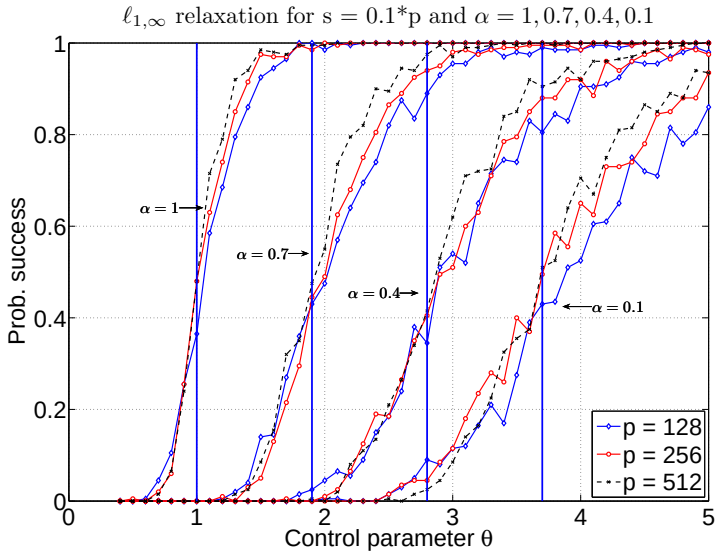


Figure 1. Probability of success in recovering the joint signed supports plotted against the rescaled sample size $\theta_{\text{Las}} := n/[2s \log(p - (2 - \alpha)s))]$ for linear sparsity $s = 0.1p$. Each stack of graphs corresponds to a fixed overlap α, as labeled on the figure. The three curves within each stack correspond to problem sizes $p \in \{128, 256, 512\}$; note how they all align with each other and exhibit step-like behavior, consistent with Theorem 3. The vertical lines correspond to the thresholds $\theta^*_{1,\infty}(\alpha)$ predicted by Theorem 3; note the close agreement between theory and simulation.

from 1 towards 0, showing that problems with less overlap require a larger rescaled sample size. More interesting is the sharpness of agreement in *quantitative terms:* the vertical lines

in the center of each stack show the point at which our theory (1) predicts that the method should transition from failure to success.

By comparison to previous theory on the behavior of the Lasso (ordinary ℓ_1-regularized quadratic programming), the scaling (1) has two interesting implications. For the s-sparse regression problem with standard Gaussian designs, the Lasso has been shown [33] to transition from success to failure as a function of the rescaled sample size

$$\theta_{\mathrm{Las}}(n, p, s) \quad := \quad \frac{n}{2s \log(p - s)}. \tag{2}$$

In particular, under the conditions imposed here, solving two separate Lasso problems, one for each regression problem, would recover both supports for problem sequences (n, p, s) such that $\theta_{\mathrm{Las}} > 1$. Thus, one consequence of our analysis is to characterize the *relative statistical efficiency* of ℓ_1/ℓ_∞ regularization versus ordinary ℓ_1-regularization, as described by the ratio $R := \frac{\theta_{1,\infty}}{\theta_{\mathrm{Las}}}$.

Our theory predicts that (disregarding some $o(1)$ factors) the relative efficiency scales as $R(\alpha) \sim \frac{4-3\alpha}{2}$, which (as we show later) shows excellent agreement with empirical behavior in simulation. Our characterization of $R(\alpha)$ confirms that if the regression matrix $\overline{B}$ is well-aligned with the block ℓ_1/ℓ_∞ regularizer—more specifically for overlaps $\alpha \in [\frac{2}{3}, 1]$—then block-regularization increases statistical efficiency. On the other hand, our analysis also conveys a *cautionary message*: if the overlap is too small—more precisely, if $\alpha < 2/3$—then block $\ell_{1,\infty}$ is actually relative to the naive Lasso-based approach. This fact illustrates that some care is required in the application of block regularization schemes.

In terms of proof techniques, the analysis of this paper is considerably more delicate than the analogous arguments required to show support consistency for the Lasso [19, 33, 37]. The major difference—and one that presents substantial technical challenges—is that the sub-differential[2] of the block ℓ_1/ℓ_∞ is a much more subtle object than the subdifferential of the ordinary ℓ_1-norm. In particular, the ℓ_1-norm has an ordinary derivative whenever the coefficient vector is non-zero. In contrast, even for non-zero rows of the regression matrix, the block ℓ_1/ℓ_∞ norm may be non-differentiable, and these non-differentiable points play a key role in our analysis. (See Section 4.1 for more detail on the sub-differential of this block norm.) As we show, it is the Frobenius norm of the sub-differential on the regression matrix support that controls high-dimensional scaling. For the ordinary ℓ_1-norm, this Frobenius norm is always equal to s, whereas for matrices with $r = 2$ columns and α fraction overlap, this Frobenius norm can be as small as $\frac{(4-3\alpha)\, s}{2}$. As our analysis reveals, it is precisely the differing structures of these sub-differentials that leads to different high-dimensional scaling for ℓ_1 versus $\ell_{1,\infty}$ regularization.

The remainder of this paper is organized as follows. In Section 2, we provide a precise description of the problem. Section 3 is devoted to the statement of our main results, some discussion of their consequences, and illustration by comparison to empirical simulations. In Section 4, we provide an outline of the proof, with the technical details of many intermediate lemmas deferred to the appendices.

Notational conventions: For the convenience of the reader, we summarize here some notation to be used throughout the paper. We reserve the index $i \in \{1, \ldots, r\}$ as a superscript

[2] As we describe in more detail in Section 4.1, the sub-differential is the appropriate generalization of gradient to convex functions that are allowed to have "corners", like the ℓ_1 and ℓ_1/ℓ_∞ norms; the standard books [26, 8] contain more background on sub-differentials and their properties.

in indexing the different regression problems, or equivalently the columns of the matrix $\overline{B} \in \mathbb{R}^{p \times r}$. Given a design matrix $X \in \mathbb{R}^{n \times p}$ and a subset $S \subseteq \{1, \ldots, p\}$, we use X_S to denote the $n \times |S|$ sub-matrix obtained by extracting those columns indexed by S. For a pair of matrices $A \in \mathbb{R}^{m \times \ell}$ and $B \in \mathbb{R}^{m \times n}$, we use the notation $\langle A, B \rangle := A^T B$ for the resulting $\ell \times n$ matrix.

We use the following standard asymptotic notation: for functions f, g, the notation $f(n) = \mathcal{O}(g(n))$ means that there exists a fixed constant $0 < C < +\infty$ such that $f(n) \leq Cg(n)$; the notation $f(n) = \Omega(g(n))$ means that $f(n) \geq Cg(n)$, and $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$.

2 Problem set-up

We begin by setting up the problem to be studied in this paper, including multivariate regression and family of block-regularized programs for estimating sparse vectors.

2.1 Multivariate regression and block regularization schemes

In this paper, we consider the following form of multivariate regression. For each $i = 1, \ldots, r$, let $\bar{\beta}^i \in \mathbb{R}^p$ be a regression vector, and consider the r-variate linear regression problem

$$y^i \quad = \quad X^i \bar{\beta}^i + w^i, \qquad i = 1, 2, \ldots, r. \tag{3}$$

Here each $X^i \in \mathbb{R}^{n \times p}$ is a design matrix, possibly different for each vector $\bar{\beta}^i$, and $w^i \in \mathbb{R}^n$ is a noise vector. We assume that the noise vectors w^i and w^j are independent for different regression problems $i \neq j$. In this paper, we assume that each w^i has a multivariate Gaussian $N(0, \sigma^2 I_{n \times n})$ distribution. However, we note that qualitatively similar results will hold for any noise distribution with sub-Gaussian tails (see the book [4] for more background on sub-Gaussian variates).

For compactness in notation, we frequently use $\overline{B}$ to denote the $p \times r$ matrix with $\bar{\beta}^i \in \mathbb{R}^p$ as the i^{th} column. Given a parameter $q \in [1, \infty]$, we define the ℓ_1/ℓ_q block-norm as follows:

$$\|\overline{B}\|_{\ell_1/\ell_q} \quad := \quad \sum_{k=1}^{p} \|(\bar{\beta}_k^1, \bar{\beta}_k^2, \ldots, \bar{\beta}_k^r)\|_q, \tag{4}$$

corresponding to applying the ℓ_q norm to each row of $\overline{B}$, and the ℓ_1-norm across all of these blocks. We note that all of these block norms are special cases of the CAP family of penalties [36].

This family of block-regularizers (4) suggests a natural family of M-estimators for estimating $\overline{B}$, based on solving the block-ℓ_1/ℓ_q-regularized quadratic program

$$\widehat{B} \quad \in \quad \arg\min_{B \in \mathbb{R}^{p \times r}} \Big\{ \frac{1}{2n} \sum_{i=1}^{r} \|y^i - X^i \beta^i\|_2^2 + \lambda_n \|B\|_{\ell_1/\ell_q} \Big\}, \tag{5}$$

where $\lambda_n > 0$ is a user-defined regularization parameter. Note that the data term is separable across the different regression problems $i = 1, \ldots, r$, due to our assumption of independence on the noise vectors. Any coupling between the different regression problems is induced by the block-norm regularization.

In the special case of univariate regression ($r = 1$), the parameter q plays no role, and the block-regularized scheme (6) reduces to the Lasso [28, 5]. If $q = 1$ and $r \geq 2$, the

block-regularization function (like the data term) is separable across the different regression problems $i = 1, \ldots, r$, and so the scheme (6) reduces to solving r separate Lasso problems. For $r \geq 2$ and $q = 2$, the program (6) is frequently referred to as the group Lasso [35, 22]. Another important case [31, 30] and the focus of this paper is the setting $q = \infty$ and $r \geq 2$, which we refer to as block ℓ_1/ℓ_∞ regularization.

The motivation for using block ℓ_1/ℓ_∞ regularization is to encourage *shared sparsity* among the columns of the regression matrix B. Geometrically, like the ℓ_1 norm that underlies the ordinary Lasso, the ℓ_1/ℓ_∞ block norm has a polyhedral unit ball. However, the block norm captures potential interactions between the columns β^i in the matrix B. Intuitively, taking the maximum encourages the elements $(\beta_k^1, \beta_k^2 \ldots, \beta_k^r)$ in any given row $k = 1, \ldots, p$ to be zero simultaneously, or to be non-zero simultaneously. Indeed, if $\beta_k^i \neq 0$ for at least one $i \in \{1, \ldots, r\}$, then there is no additional penalty to have $\beta_k^j \neq 0$ as well, as long as $|\beta_k^j| \leq |\beta_k^i|$.

2.2 Estimation in ℓ_∞ norm and support recovery

For a given $\lambda_n > 0$, suppose that we solve the block ℓ_1/ℓ_∞ program, thereby obtaining an estimate

$$\widehat{B} \quad \in \quad \arg\min_{B \in \mathbb{R}^{p \times r}} \Big\{ \frac{1}{2n} \sum_{i=1}^{r} \|y^i - X^i \beta^i\|_2^2 + \lambda_n \|B\|_{\ell_1/\ell_\infty} \Big\}, \tag{6}$$

We note that under high-dimensional scaling ($p \gg n$), this convex program (6) is not necessarily strictly convex, since the quadratic term is rank deficient and the block ℓ_1/ℓ_∞ norm is polyhedral, which implies that the program is not strictly convex. However, a consequence of our analysis is that under appropriate conditions, the optimal solution $\widehat{B}$ is in fact unique.

In this paper, we study the accuracy of the estimate $\widehat{B}$, as a function of the sample size n, regression dimensions p and r, and the sparsity index $s = \max_{i=1,\ldots,r} |S(\bar{\beta}^i)|$. There are various metrics with which to assess the "closeness" of the estimate $\widehat{B}$ to the truth $\overline{B}$, including predictive risk, various types of norm-based bounds on the difference $\widehat{B} - \overline{B}$, and variable selection consistency. In this paper, we prove results bounding the ℓ_∞/ℓ_∞ difference

$$\|\widehat{B} - \overline{B}\|_{\ell_\infty/\ell_\infty} \quad := \quad \max_{k=1,\ldots,p} \max_{i=1,\ldots,r} |\widehat{B}_k^i - \overline{B}_k^i|.$$

In addition, we prove results on support recovery criteria. Recall that for each vector $\bar{\beta}^i \in \mathbb{R}^p$, we use $S(\bar{\beta}^i) = \{k \mid \bar{\beta}_k^i \neq 0\}$ to denote its support set. The problem of *row support recovery* corresponds to recovering the set

$$U \quad := \quad \bigcup_{i=1}^{r} S(\bar{\beta}^i), \tag{7}$$

corresponding to the subset $U \subseteq \{1, \ldots, p\}$ of indices that are active in at least one regression problem. Note that the cardinality of $|U|$ is upper bounded by rs, but can be substantially smaller (as small as s) if there is overlap among the different supports.

As discussed at more length in Appendix A, given an estimate of the row support of $\overline{B}$, it is possible to either use additional structure of the solution $\widehat{B}$ or perform some additional computation to recover *individual signed supports* of the columns of $\overline{B}$. To be precise, define the sign function

$$\operatorname{sign}(t) \quad = \quad \begin{cases} +1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0. \end{cases} \tag{8}$$

Then the recovery of individual signed supports means estimating the signed vectors with entries $\text{sign}(\bar{\beta}_k^i)$, for each $i = 1, 2, \ldots, r$ and for all $k = 1, 2, \ldots, p$. Interestingly, when using block ℓ_1/ℓ_∞ regularization, there are multiple ways in which the support (or signed support) can be estimated, depending on whether we use primal or dual information from an optimal solution.

The *dual recovery method* involves the following steps. First, solve the block-regularized program (6), thereby obtaining an primal solution $\widehat{B} \in \mathbb{R}^{p \times r}$. For each row $k = 1, \ldots, p$, compute the set $\mathbb{M}_k := \arg\max_{i=1,\ldots,r} |\widehat{\beta}_k^i|$. Estimate the support union via $\widehat{U} = \bigcup_{i=1,\ldots,r} S(\widehat{\beta}^i)$, and estimate the signed support vectors

$$[\mathbb{S}_{\text{dua}}(\widehat{\beta}_k^i)] \quad = \quad \begin{cases} \text{sign}(\widehat{\beta}_k^i) & \text{if } i \in \mathbb{M}_k \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

As our development will clarify, this procedure (9) corresponds to estimating the signed support on the basis of a dual optimal solution associated with the optimal primal solution. We discuss the primal-based recovery method and its differences with the dual-based method at more length in Appendix A.

3 Main results and their consequences

In this section, we provide precise statements of the main results of this paper. Our first main result (Theorem 1) provides sufficient conditions for deterministic design matrices $X^1, \ldots, X^r$, whereas our second main result (Theorem 2) provides sufficient conditions for design matrices drawn randomly from sub-Gaussian ensembles. Both of these results allow for an arbitrary number r of regression problems, and the random design case allows for random Gaussian designs X^k with i.i.d. rows and covariance matrix $\Sigma^k \in \mathbb{R}^{p \times p}, k = 1, \ldots, r$. Not surprisingly, these results show that the high-dimensional scaling of block ℓ_1/ℓ_∞ is *qualitatively similar* to that of ordinary ℓ_1-regularization: for instance, in the case of random Gaussian designs and bounded r, our sufficient conditions ensure that $n = \Omega(s \log p)$ samples are sufficient to recover the union of supports correctly with high probability, which matches known results on the Lasso [33], as well as known information-theoretic results on the problem of support recovery [32].

As discussed in the introduction, we are also interested in the more refined question: can we provide necessary and sufficient conditions that are sharp enough to reveal *quantitative differences* between ordinary ℓ_1-regularization and block regularization? Addressing this question requires analysis that is sufficiently precise to control the constants in front of the rescaled sample size $n/s \log(p - s)$ that controls the performance of both ℓ_1 and block ℓ_1/ℓ_∞ methods. Accordingly, in order to provide precise answers to this question, our final two results concern the special case of $r = 2$ regression problems, both with supports of size s that overlap in a fraction α of their entries, and with design matrices drawn randomly from the standard Gaussian ensemble. In this setting, our final result (Theorem 3) shows that block ℓ_1/ℓ_∞ regularization undergoes a phase transition—that is, a rapid change from failure to success—specified by the rescaled sample size $\theta_{1,\infty}(n, p, s, \alpha)$ previously defined (1). We then discuss some consequences of these results, and illustrate their sharpness with some simulation results.

3.1 Sufficient conditions for general deterministic and random designs

In addition to the sample size n, problem dimensions p and r, sparsity index s and overlap parameter α, our results involve certain quantities associated with the design matrices X^i. To begin, in the deterministic case, we assume that the columns of each design matrix $X^i, i = 1, \ldots, r$ are normalized[3] so that

$$\|X^i_k\|_2^2 \;\leq\; 2n \qquad \text{for all } k = 1, 2, \ldots p. \tag{10}$$

More significantly, we require that the following *incoherence condition* on the design matrix be satisfied:

$$\gamma(\Sigma) \;:=\; 1 - \max_{\ell=1,\ldots,|U^c|} \sum_{i=1}^{r} \|\big\langle X^i_\ell,\, X^i_U(\langle X^i_U,\, X^i_U\rangle)^{-1}\big\rangle\|_1 \;>\; 0. \tag{11}$$

For the case of the ordinary Lasso, conditions of this type are known [19, 37, 33] to be both necessary and sufficient for successful support recovery.[4]

In addition, the statement of our results involve certain quantities associated with the $|U| \times |U|$ matrices $\frac{1}{n}\langle X^i_U,\, X^i_U\rangle$; in particular, we define a lower bound on the minimum eigenvalue

$$C_{\min}(X) \;\leq\; \min_{i=1,\ldots,r} \lambda_{\min}\big(\frac{1}{n}\langle X^i_U,\, X^i_U\rangle\big), \tag{12}$$

as well as an upper bound maximum $\ell_{\infty,\infty}$-operator norm of the inverses

$$D_{\max}(X) \;\geq\; \max_{i=1,\ldots,r} |\!|\!|\big(\frac{1}{n}\langle X^i_U,\, X^i_U\rangle\big)^{-1}|\!|\!|_\infty. \tag{13}$$

Remembering that our analysis applies to to sequences $\{X_{n,p}\}$ of design matrices, in the simplest scenario, both of the bounding quantities $C_{\min}$ and $D_{\max}$ do not scale with (n, p, s). To keep notation compact, we write $C_{\min}$ and $D_{\max}$ in the analysis to follow.

We also define the *support minimum value*

$$\overline{B}_{\min} \;=\; \min_{k \in U} \max_{i=1,\ldots,r} |\bar{\beta}^i_k|, \tag{14}$$

corresponding to the minimum value of the ℓ_∞ norm of any row $k \in U$.

Theorem 1 (Sufficient conditions for deterministic designs)**.** *Consider the observation model* (3) *with design matrices X^i satisfying the column bound* (10) *and incoherence condition* (11)*. Suppose that we solve the block-regularized ℓ_1/ℓ_∞ convex program* (6) *with regularization parameter $\lambda_n^2 \geq \frac{4\xi\sigma^2}{\gamma^2}\frac{r^2 + r\log(p)}{n}$ for some $\xi > 1$. Then with probability greater than*

$$\phi_1(\xi, p, s) \;:=\; 1 - 2\exp(-(\xi - 1)[r + \log p]) - 2\exp(-(\xi^2 - 1)\log(rs)), \tag{15}$$

we are guaranteed that

(a) The block-regularized program has a unique solution $\widehat{B}$ such that $\bigcup_{i=1}^r S(\widehat{\beta}^i) \subseteq U$.

[3]The choice of the factor 2 in this bound is for later technical convenience.

[4]Some work [20] has shown that multi-stage methods can allow some relaxation of this incoherence condition; however, as our main interest is in understanding the sample complexity of ordinary ℓ_1 versus ℓ_1/ℓ_∞ relaxations, we do not pursue such extensions here.

(b) Moreover, the solution satisfies the elementwise ℓ_∞-bound

$$\|\widehat{B} - \overline{B}\|_{\ell_\infty/\ell_\infty} \;\leq\; \underbrace{\xi\sqrt{\frac{4\sigma^2}{C_{\min}}\,\frac{\log|U|}{n}} + D_{\max}\,\lambda_n}_{b_1(\xi,\lambda_n,n,s)}. \tag{16}$$

Consequently, as long as $\overline{B}_{\min} \geq b_1(\xi,\lambda_n,n,s)$, then $\bigcup_{i=1}^r S(\widehat{\beta}^i) = U$, so that the solution $\widehat{B}$ correctly specifies the union of supports U.

We now state an analogous result for random design matrices; in particular, consider the observation model (3) with design matrices X^i chosen with i.i.d. rows from covariance matrices Σ^i. In analogy to definitions (12) and (13) in the deterministic case, we define the lower bound

$$C_{\min}(\Sigma) \;\leq\; \min_{i=1,\ldots,r} \lambda_{\min}\big(\Sigma^i_{UU}\big), \tag{17}$$

as well as an analogous upper bound on ℓ_∞-operator norm of the inverses

$$D_{\max}(\Sigma) \;\geq\; \max_{i=1,\ldots,r} |||(\Sigma^i_{UU})^{-1}|||_\infty \leq D_{\max}. \tag{18}$$

Note that unlike the case of deterministic designs, these quantities are *not* functions of the design matrix X, which is now a random variable. Finally, our results involve an analogous incoherence parameter of the covariance matrices $\Sigma = \{\Sigma^i, i = 1,\ldots,r\}$, defined as

$$\gamma(\Sigma) \;:=\; 1 - \max_{k=1,\ldots,|U^c|} \sum_{i=1}^r \big\|\Sigma^i_{kU}\,(\Sigma^i_{UU})^{-1}\big\|_1 \;>\; 0. \tag{19}$$

With this notation, the following result provides an analog of Theorem 1 for random design matrices:

Theorem 2 (Sufficient conditions for random Gaussian designs)**.** *Suppose that we are given n i.i.d. observations from the model* (3) *with*

$$n \;>\; \frac{8\,\kappa\,r}{C_{\min}\gamma^2}\,s\big(r + \log p\big) \tag{20}$$

for some $\kappa > 1$. If we solve the convex program (6) *with regularization parameter satisfying $\lambda_n \geq \frac{4\xi\sigma^2}{\gamma^2}\big[\frac{r^2 + r\log(p)}{n}\big]$ for some $\xi > 1$, then with probability greater than*

$$\phi_2(\kappa,\xi,n,p,s) := 1 - 2\exp\big\{-2(\xi^2-1)\log(rs)\big\} - 2\exp\big\{-\kappa(r+\log p)\big\} \to 1, \tag{21}$$

we are guaranteed that

(a) The block-regularized program (6) *has a unique solution $\widehat{B}$ such that $\bigcup_{i=1}^r S(\widehat{\beta}^i) \subseteq U$.*

(b) The solution satisfies the elementwise ℓ_∞ bound

$$\|\widehat{B} - \overline{B}\|_{\ell_\infty/\ell_\infty} \;\leq\; \underbrace{\xi\,\sqrt{\frac{100\sigma^2}{C_{\min}}\,\frac{\log|U|}{n}} + \lambda_n\big[\frac{4s}{\sqrt{n}} + D_{\max}\big]}_{b_2(\xi,\lambda_n,n,s)}, \tag{22}$$

Consequently, if $B^*_{\min} \geq b_2(\xi,\lambda_n,n,s)$, then $\bigcup_{i=1}^r S(\widehat{\beta}^i) = U$, so that the solution $\widehat{B}$ correctly specifies the union of supports U.

To clarify the interpretation of Theorems 1 and Theorem 2, part (a) of each claim guarantees that the estimator has *no false inclusions*, in that the row support of the estimate $\widehat{B}$ is contained within the row support of the true matrix $\overline{B}$. One consequence of part (b) is that as long as the minimum signal parameter $B^*_{\min}$ decays slowly enough, then the estimators have *no false exclusions*, so that the true row support is correctly recovered.

In terms of consistency rates in block ℓ_∞/ℓ_∞ norm, assuming that the design-related quantities $C_{\min}$, $D_{\max}$ and γ do not scale with p, Theorem 1(a) guarantees consistency in elementwise ℓ_∞-norm at the rate

$$\|\widehat{B} - \overline{B}\|_{\ell_\infty/\ell_\infty} \quad = \quad \mathcal{O}\big(\sigma^2 \sqrt{\frac{r^2 + r\log p}{n}}\big).$$

Here we have used the fact that $\log|U| \leq \log(rs) = o(r\log p)$. Similarly, Theorem 2(b) guarantees consistency in elementwise ℓ_∞-norm at the rate

$$\|\widehat{B} - \overline{B}\|_{\ell_\infty/\ell_\infty} \quad = \quad \mathcal{O}\big(\sigma^2 \sqrt{\max\{1, \frac{s}{r\log p}\}\ \frac{r^2 + r\log p}{n}}\big).$$

In this expression, the extra term $\max\{1, s/(r\log p)\}$ arises in the analysis due to the need to control the norms of the random design matrices. For sufficiently sparse problems (e.g., $s = \mathcal{O}(\log p)$), this factor is constant.

At a high level, our results thus far show that for a fixed number r of regression problems, the ℓ_1/ℓ_∞ method guarantees exact support recovery with $n = \Omega(s\log p)$ samples, and guarantees consistency in an elementwise sense at rate $\mathcal{O}(\sqrt{\frac{\log p}{n}})$. In qualitative terms, these results match the known scaling [33] for the Lasso (ℓ_1-regularized QP), which is obtained as the special case for univariate regression ($r = 1$). It should be noted that this scaling is known to be optimal in an information-theoretic sense: no algorithm can recover support correctly if the rescaled sample size $\theta_{\text{Las}} = \frac{n}{2s\log(p-s)}$ is below a critical threshold [32, 34].

3.2 A phase transition for standard Gaussian ensembles

In order to provide keener insight into the advantages and/or disadvantages associated with using ℓ_1/ℓ_∞ block regularization, we need to obtain even sharper results, ones that are capable of distinguishing constants in front of the rescaled sample size θ_{Las}. With this aim in mind, the following results are specialized to the case of $r = 2$ regression problems, where the corresponding design matrices $X^i, i = 1, 2$ are sampled from the standard Gaussian ensemble—i.e., with i.i.d. rows $N(0, I_{p\times p})$. By studying this simpler class of problems, we can make *quantitative comparisons* to the sample complexity of the Lasso, which provide insight into the benefits and dangers of block ℓ_1/ℓ_∞ regularization.

The main result of this section asserts that there is a phase transition in the performance of ℓ_1/ℓ_∞ quadratic programming for suppport recovery—by which we mean a sharp transition from failure to success—and provide the exact location of this transition point as a function of (n, p, s) and the overlap parameter $\alpha \in (0, 1)$. The phase transition involves the *support gap*

$$\overline{B}_{\text{gap}} \quad = \quad \min_{i \in S(\bar{\beta}^1) \cap S(\bar{\beta}^2)} \big|\, |\bar{\beta}^1_i| - |\bar{\beta}^2_i| \,\big|. \tag{23}$$

This quantity measures how close the two regression vectors are in absolute value on their *shared support*. Our main theorem treats the case in which this gap vanishes (i.e., $\overline{B}_{\text{gap}} = o(1)$); note that block ℓ_1/ℓ_∞ regularization is best-suited to this type of structure. A subsequent corollary provides more general but technical conditions for the cases of non-vanishing support gaps. Our main result specifies a phase transition in terms of the *rescaled sample size*

$$\theta_{1,\infty}(n,p,s,\alpha) \quad := \quad \frac{n}{(4-3\alpha)s\log(p-(2-\alpha)s)}, \tag{24}$$

as stated in the theorem below.

Theorem 3 (Phase transition)**.** *Consider sequences of problems, indexed by (n,p,s,α) drawn from the observation model (3) with random design X drawn with i.i.d. standard Gaussian entries and with $C_{\min} = 1 = D_{\max}$.*

(a) Success: Suppose that the problem sequence (n,p,s,α) satisfies

$$\theta_{1,\infty}(n,p,s,\alpha) \quad > \quad 1+\delta \qquad \text{for some } \delta > 0. \tag{25}$$

If we solve the block-regularized program (6) with $\lambda_n \geq \sqrt{\frac{\xi\sigma^2\log p}{n}}$ for some $\xi > 2$ and $\overline{B}_{\text{gap}} = o(\lambda_n)$, then with probability greater than $1 - c_1\exp(-c_2\log(p-(2-\alpha)s))$, the block $\ell_{1,\infty}$-program (6) has a unique solution $\widehat{B}$ such that $S(\widehat{B}) \subseteq U$, and moreover it satisfies the elementwise bound (22) with $C_{\min} = 1 = D_{\max}$. In addition, if $B^*_{\min} > b_2(\xi,\lambda_n,n,s)$, then the unique solution recovers the correct signed support.

(b) Failure: For problem sequences (n,p,s,α) such that

$$\theta_{1,\infty}(n,p,s,\alpha) \quad < \quad 1-\delta \qquad \text{for some } \delta > 0 \tag{26}$$

and for any non-increasing regularization sequence $\lambda_n > 0$, no solution $\widehat{B} = (\widehat{\beta}^1, \widehat{\beta}^2)$ to the block-regularized program (6) has the correct signed support.

In a nutshell, Theorem 3 states that block ℓ_1/ℓ_∞ regularization recovers the correct support with high probability for sequences (n,p,s,α) such that $\theta_{1,\infty}(n,p,s,\alpha) > 1$, and otherwise fails with high probability.

We now consider the case in which the support gap does not vanish, and show that it only further degrades the performance of block ℓ_1/ℓ_∞ regularization. To make the degree of this degradation precise, we define the λ_n-truncated gap vector $T_{\lambda_n}(\widehat{B}) \in \mathbb{R}^p$, with elements

$$[T_{\lambda_n}(\overline{B})]_i \quad := \quad \begin{cases} \min\left\{\lambda_n, \ \big|\,|\bar{\beta}^1_i| - |\bar{\beta}^2_i|\,\big|\right\} & \text{if } i \in S(\bar{\beta}^1) \cap S(\bar{\beta}^2) \\ 0 & \text{otherwise} \end{cases}$$

Recall that support overlap $S(\bar{\beta}^1) \cap S(\bar{\beta}^2)$ has cardinality αs by assumption. Therefore, $T_{\lambda_n}(\overline{B})$ has at most αs non-zero entries, and moreover $\|T_{\lambda_n}(\overline{B})\|_2^2 \leq \lambda_n^2 \alpha s$. We then define the rescaled gap limit

$$\Delta(\overline{B},\lambda_n) \quad := \quad \limsup_{(n,p,s)} \frac{\|T_{\lambda_n}(\overline{B})\|_2^2}{\lambda_n^2 s}. \tag{27}$$

Note that $\Delta(\overline{B},\lambda_n) \in [0,\alpha]$ by construction. With these definitions, we have the following:

Corollary 1 (Poorer performance with non-vanishing gap)**.** *If for any $\delta > 0$, the sample size n is upper bounded as*

$$n \;<\; (1-\delta)\,\big[(4-3\alpha) + \Delta(\overline{B}, \lambda_n)\big] s \log[p - (2-\alpha)s], \tag{28}$$

then the dual recovery method (9) *fails to recover the individual signed supports.*

To understand the implications of this result, suppose that all αs of the gaps $\big|\,|\bar{\beta}_i^1| - |\bar{\beta}_i^2|\,\big|$ were above the regularization level λ_n. Then by definition, we have $\Delta(\overline{B}, \lambda_n) = \alpha$, so that condition (28) implies that the method fails for all $n < (1-\delta)\,[4-2\alpha] s \log[p-(2-\alpha)s]$. Since the factor $(4-2\alpha)$ is strictly greater than 2 for all $\alpha < 1$, this scaling is *always worse*[5] than the Lasso scaling given by $n \asymp 2s\log(p-s)$ (see equation (2)), unless there is perfect overlap ($\alpha = 1$), in which case it yields no improvements. Consequently, Corollary 1 shows that the performance ℓ_1/ℓ_∞ regularization is also very sensitive to the numerical amplitudes of the signal vectors.

3.3 Illustrative simulations and some consequences

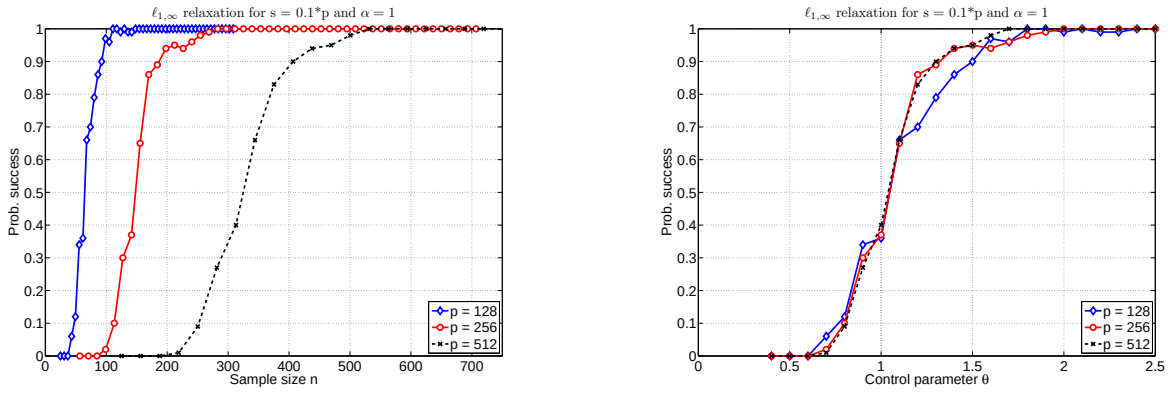
In this section, we provide some simulation results to illustrate the phase transition predicted by Theorem 3. Interestingly, these results show that the theory provides an accurate description of practice even for relatively small problem sizes (e.g., $p = 128$). As specified in Theorem 3, we simulate multivariate regression problems with $r = 2$ columns, with the design matrices X^i drawn from the standard Gaussian ensemble. In all cases, we initially solved the ℓ_1/ℓ_∞ program using MATLAB, and then verified that the behavior of the solution agreed with the primal-dual optimality conditions specified by our theory. In subsequent simulations, we solved directly for the dual variables, and then checked whether or not the dual feasibility conditions are met.

We first illustrate the difference between unscaled and rescaled plots of the empirical performance, which demonstrate that the rescaled sample size $n/[s\log(p-s)]$ specifies the high-dimensional scaling of block ℓ_1/ℓ_∞ regularization. Figure 2(a) shows the empirical behavior of the block ℓ_1/ℓ_∞ method for joint support recovery. For these simulations, we applied the method to $r = 2$ regression problems with overlap $\alpha = 1$, and to three different problem sizes $p \in \{128, 256, 512\}$, in all cases with the sparsity index $s = \lfloor 0.1p \rfloor$. Each curve in panel (a) shows the probability of correct support recovery $\mathbb{P}[\widehat{U} = U]$ versus the raw sample size n. As would be expected, all the curves initially start at $\mathbb{P}[\widehat{U} = U] = 0$, but then transition to 1 as n increases, with the transition taking place at larger and larger sampler sizes as p is increased. The purpose of the rescaling is to determine exactly how this transition point depends on the problem size p and other structural parameters (s and α). Figure 2(b) shows the same simulation results, now plotted versus the rescaled sample size $\theta := n/[2s\log(p-s)]$, which is the appropriate rescaling predicted by our theory. Notice how all three curves now lie on top of another, and moreover transition from failure to success at $\theta \approx 1$, consistent with our theoretical predictions.

We now seek to explore the dependence of the sample size on the overlap fraction $\alpha \in [0, 1]$ of the two regression vectors. For this purpose, we plot the probability of successful recovery versus the rescaled sample size

$$\theta_{1,\infty}(n, p, s, \alpha) = \frac{n}{(4-3\alpha)s\log(p-(2-\alpha)s)}.$$

[5]Here we are assuming that $s/p = o(1)$, so that $\log(p-s) \asymp \log[p-(2-\alpha)s]$.



(a) (b)

Figure 2. (a) Plots of the probability $\mathbb{P}[\widehat{U} = U]$ of successful joint support recovery versus the sample size n. Each curve corresponds to a different problem size p; notice how the curves shift to the right as p increases, reflecting the difficulty of solving larger problems. (b) Plots of the same data versus the rescaled sample size $n/[2s\log(p-s)]$; note how all three curves now align with one another, showing that this order parameter is the correct scaling for assessing the method.

As shown by Figure 2(b), when plotted with this rescaling, there is any longer size p. Moreover, if we choose the sparsity index s to grow in a fixed way with p (i.e., $s = f(p)$ for some fixed function f), then the only remaining free variable is the overlap parameter α. Note that the theory predicts that the required sample size should decrease as α increases towards 1.

As shown earlier in Section 1, Figure 1 plots the probability of successful recovery of the joint supports versus the rescaled samples size $\theta_{1,\infty}(n, p, s, \alpha)$. Notice that the plot shows four sets of 'stacked" curves, where each stack corresponds to a different choice of the overlap parameter, ranging from $\alpha = 1$ (left-most stack), to $\alpha = 0.1$ (right-most stack). Each stack contains three curves, corresponding to the problem sizes $p \in \{128, 256, 512\}$. In all cases, we fixed the support size $s = 0.1p$. As with Figure 2(b), the "stacking" behavior of these curves demonstrates that Theorem 3 isolates the correct dependence on p. Moreover, their step-like behavior is consistent with the theoretical prediction of a phase transition. Notice how the curves shift towards the left as the overlap parameter α parameter increases towards one, reflecting that the problems become easier as the amount of shared sparsity increases. To assess this shift in a qualitative manner for each choice of overlap $\alpha \in \{0.1, 0.3.0.7.1\}$, we plot a vertical line within each group, which is obtained as the threshold value of $\theta_{1,\infty}$ predicted by our theory. Observe how the theoretical value shows excellent agreement with the empirical behavior.

As noted previously in Section 1, Theorem 3 has some interesting consequences, particularly in comparison to the behavior of the "naive" Lasso-based individual decoding of signed supports—that is, the method that simply applies the Lasso (ordinary ℓ_1-regularization) to each column $i = 1, 2$ separately. By known results [33] on the Lasso, the performance of this naive approach is governed by the order parameter $\theta_{\text{Las}}(n, p, s) = \frac{n}{2s\log(p-s)}$, meaning that for any $\delta > 0$, it succeeds for sequences such that $\theta_{\text{Las}} > 1 + \delta$, and conversely fails for sequences such that $\theta_{\text{Las}} < 1 - \delta$. To compare the two methods, we define the relative efficiency coefficient $R(\theta_{1,\infty}, \theta_{\text{Las}}) := \theta_{\text{Las}}(n, p, s)/\theta_{1,\infty}(n, p, s, \alpha)$. A value of $R < 1$ implies that the block method is more efficient, while $R > 1$ implies that the naive method is more

efficient. With this notation, we have the following:

Corollary 2. *The relative efficiency of the block $\ell_{1,\infty}$ program* (6) *compared to the Lasso is given by* $R(\theta_{1,\infty}, \theta_{\text{Las}}) = \frac{4-3\alpha}{2} \frac{\log(p-(2-\alpha)s)}{\log(p-s)}$. *Thus, for sublinear sparsity $s/p \to 0$, the block scheme has greater statistical efficiency for all overlaps $\alpha \in (2/3, 1]$, but lower statistical efficiency for overlaps $\alpha \in [0, 2/3)$.*

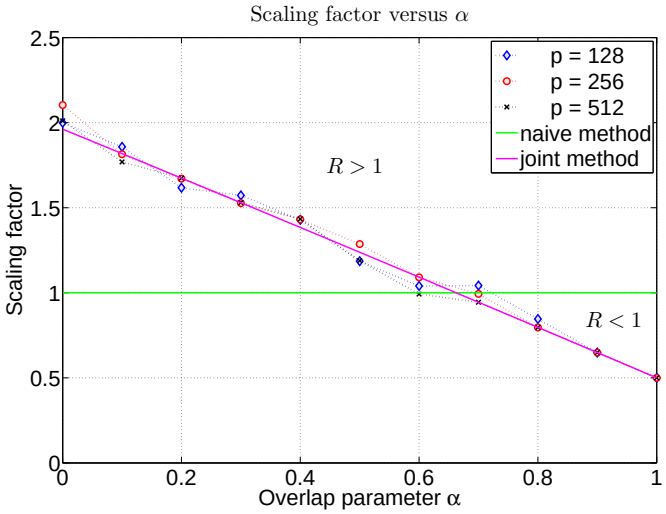


Figure 3. Plots of the relative statistical efficiency $R(\alpha)$ of a method based on block-ℓ_1/ℓ_∞ regularization versus the Lasso (ordinary ℓ_1-regularization). For each value of the parameter $\alpha \in [0, 1]$ that measures overlap between the regression problems, the quantity $R(\alpha)$ is the ratio of sample size required by an estimator based on block ℓ_1/ℓ_∞-regularization relative to the sample size required by the Lasso (ordinary ℓ_1-regularization). The error criterion here is recovery of the correct subset of active variables in the regression. Over a range of overlaps, the empirical thresholds of the ℓ_1/ℓ_∞ block regularization method closely align with the theoretical prediction of $(4 - 3\alpha)/2$. The block-based method begins to give benefits versus the "naive" Lasso-based method at the critical overlap $\alpha^* \approx 2/3$, at which point the relative efficiency $R(\alpha)$ first drops below 1. For overlaps $\alpha \in [0, 2/3)$, the joint method actually requires more samples than the naive method.

Figure 3 provides an alternative perspective on the data, where we have plotted how the sample size required by block regression changes as a function of the overlap parameter $\alpha \in [0, 1]$. Each set of data points plots a scaled form of the sample size required to hit 50% success, for a range of overlaps, and the straight line $(4-3\alpha)/2$ that is predicted by Theorem 3 Note the excellent agreement between the experimental results, for all three problem sizes for $p \in \{128, 256, 512\}$, and the full range of overlaps. The line $(4 - 3\alpha)/2$ also characterizes the relative efficiency R of block regularization versus the naive Lasso-based method, as described in Corollary 2. For overlaps $\alpha > 2/3$, this parameter R drops below 1. On the other hand, for overlaps $\alpha < 1$, we have $R > 1$, so that applying the joint optimization problem actually decreases statistical efficiency. Intuitively, although there is still some fraction of overlap, the regularization is misleading, in that it tries to enforce a higher degree of shared sparsity than is actually present in the data.

4 Proofs

This section contains the proofs of our three theorems. Our proofs are constructive in nature, based on a procedure that constructs pair of matrices $\widetilde{B} = (\widetilde{\beta}^1, \ldots, \widetilde{\beta}^r) \in \mathbb{R}^{p \times r}$ and $\widetilde{Z} = (\widetilde{z}^1, \ldots, \widetilde{z}^r) \in \mathbb{R}^{p \times r}$. The goal of the construction is to show that matrix $\widetilde{B}$ is an optimal primal solution to the convex program (6), and that the matrix $\widetilde{Z}$ is a corresponding dual-optimal solution, meaning that it belongs to the sub-differential of the $\ell_{1,\infty}$-norm (see Lemma 1), evaluated at $\widetilde{B}$. If the construction succeeds, then the pair $(\widetilde{B}, \widetilde{Z})$ acts as a witness for the success of the convex program (6) in recovering the correct signed support—in particular, success of the primal-dual witness procedure implies that $\widetilde{B}$ is the unique optimal solution of the convex program (6), with its row support contained with U. To be clear, the procedure for constructing this candidate primal-dual solution is *not* a practical algorithm (as it exploits knowledge of the true support sets), but rather a proof technique for certifying the correctness of the block-regularized program.

We begin by providing some background on the sub-differential of the ℓ_1/ℓ_∞ norm; we refer the reader to the books [26, 8] for more background on convex analysis.

4.1 Structure of ℓ_1/ℓ_∞-norm sub-differential

The sub-differential of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is the set of all vectors $y \in \mathbb{R}^d$ such that $f(x') \geq f(x) + \langle y, x' - x \rangle$ for all $x' \in \mathbb{R}^d$. See the standard references [26, 8] for background on subdifferentials and their properties.

We state for future reference a characterization of the sub-differential of the ℓ_1/ℓ_∞ block norm:

Lemma 1. *The matrix $\widetilde{Z} \in \mathbb{R}^{p \times r}$ belongs to the sub-differential $\partial \|\widetilde{B}\|_{\ell_1/\ell_\infty}$ if and only if the following conditions hold for each $k = 1, \ldots, p$.*

(i) If $\widetilde{\beta}^i_k \neq 0$ for at least one index $i \in \{1, \ldots, r\}$, then

$$\widetilde{z}^i_k \;=\; \begin{cases} t_i \operatorname{sign}(\widetilde{\beta}^i_k) & \text{if } i \in \mathbb{M}_k \\ 0 & \text{otherwise.} \end{cases},$$

where $\mathbb{M}_k := \arg\max_{i=1,\ldots,r} |\widetilde{\beta}^i_k|$, for a set of non-negative scalars $\{t_i,\ i \in \mathbb{M}_k\}$ such that $\sum_{i \in \mathbb{M}_k} t_i = 1$.

(ii) If $\widetilde{\beta}^i_k = 0$ for all $i = 1, \ldots, r$, then we require $\sum_{i=1}^r |\widetilde{z}^i_k| \leq 1$.

4.2 Primal-dual construction

We now describe our method for constructing the matrix pair $(\widetilde{B}, \widetilde{Z})$. Recalling that $U = \bigcup_{i=1}^r S(\bar{\beta}^i)$ denotes the union of supports of the true regression vectors, let U^c denote the complement of $\{1, \ldots, p\} \backslash U$. With this notation, Figure 4 provides the four steps of the primal-dual witness construction.

The following lemma summarizes the utility of the primal-dual witness method:

Lemma 2. *Suppose that for each $i = 1, \ldots, r$, the $|U| \times |U|$ sub-matrix $\langle X^i_U,\ X^i_U \rangle$ is invertible. Then for any $\lambda_n > 0$, we have the following correspondences:*

Primal-dual witness construction:

(A) First, we solve the restricted program

$$\widetilde{B} = \arg \min_{B \in \mathbb{R}^{p \times r}, B_{U^c}=0} \{ \frac{1}{2n} \sum_{i=1}^{r} \|y^i - X^i \beta^i\|_2^2 + \lambda_n \|B\|_{\ell_1/\ell_\infty} \}. \quad (29)$$

Given our assumption that the $|U| \times |U|$ sub-matrices $\langle X_U^i, X_U^i \rangle$ are invertible, the solution to this convex program is unique. Moreover, note that $\widetilde{B}_{U^c} = 0$ by construction.

(B) We choose $\widetilde{Z}_U \in \mathbb{R}^{|U| \times r}$ as an element of the subdifferential $\partial \|\widetilde{B}_U\|_{\ell_1/\ell_\infty}$.

(C) Using the optimality conditions associated with the original convex program (6), we then solve for the matrix $\widetilde{Z}_{U^c}$, and verify that its rows satisfy the *strict dual feasibility* condition

$$\sum_{i=1}^{r} |\widetilde{z}_k^i| \quad < \quad 1 \qquad \text{for all } k \in U^c. \quad (30)$$

(D) A final (optional) step is to verify that $\widetilde{B}_U$ satisfies the *sign consistency* conditions $\text{sign}(\widetilde{B}_U) = \text{sign}(\overline{B}_U)$.

Figure 4. Steps in the primal-dual witness construction. Steps (A) and (B) are straightforward; the main difficulties lie in verifying the strict dual feasibility and sign consistency conditions stated in step (C) and (D).

(i) If steps (A) through (C) of the primal-dual construction succeed, then $(\widetilde{B}_U, 0) \in \mathbb{R}^{p \times r}$ is the unique optimal solution of the original convex program (6).

(ii) Conversely, suppose that there is a solution $\widehat{B} \in \mathbb{R}^{p \times r}$ to the convex program (6) *with support contained within U. Then steps (A) through (C) of the primal-dual witness construction succeed.*

We provide the proof of Lemma 2 in Appendix D.2. It is convex-analytic in nature, based on exploiting the subgradient optimality conditions associated with both the restricted convex program (29) and the original program (6), and performing some algebra to characterize when the convex program recovers the correct signed support. Lemma 2 lies at the heart of all three of our theorems. In particular, the positive results of Theorem 1, Theorem 2 and Theorem 3(a) are based on claims (i) and (iii), which show that it is sufficient to verify that the primal-dual witness construction succeeds with high probability. The negative result of Theorem 3(b), in contrast, is based on part (ii), which can be restated as asserting that if the primal-dual witness construction fails, then no solution has support contained with U.

Before proceeding to the proofs themselves, we introduce some additional notation and develop some auxiliary results concerning the primal-dual witness procedure, to be used in subsequent development. With reference to steps (A) and (B), we show in Appendix D.2 that unique solution $\widetilde{B}_U$ has the form

$$\widetilde{B}_U \; = \; \overline{B}_U + \Delta_U, \quad (31)$$

where the matrix $\Delta_U \in \mathbb{R}^{|U| \times r}$ has columns

$$\Delta^i \;:=\; \big(\frac{1}{n}\langle X_U^i,\, X_U^i\rangle\big)^{-1}\Big[\frac{1}{n}\langle X_U^i,\, w^i\rangle - \lambda_n \widetilde{z}_U^i\Big], \qquad \text{for } i = 1, \ldots, r, \tag{32}$$

and $\widetilde{z}_U^i$ is the i^{th} column of the sub-gradient matrix $\widetilde{Z}_U$.

With reference to step (C), we obtain the candidate dual solution $\widetilde{Z}_{U^c} \in \mathbb{R}^{|U^c| \times r}$ as follows. For each $i = 1, \ldots, r$, let $\Pi_{X_U^i}$ denote the orthogonal projection onto the range of X_U^i. Using the sub-matrix $\widetilde{Z}_U \in \mathbb{R}^{|U| \times r}$ obtained from step (B), we define column i of the matrix $\widetilde{Z}_{U^c}$ as follows:

$$\widetilde{z}_{U^c}^i \;=\; \frac{1}{\lambda_n n}\big\langle X_{U^c}^i,\, (I - \Pi_{X_U^i})w^i\big\rangle + \frac{1}{n}\big\langle X_{U^c}^i,\, X_U^i(\frac{1}{n}\langle X_U^i,\, X_U^i\rangle)^{-1}\widetilde{z}_U^i\big\rangle \quad \text{for } i = 1, \ldots, r. \tag{33}$$

See the end of Appendix D.2 for derivation of this condition.

Finally, in order to further simplify notation in our proofs, for each $k \in U^c$, we define the random variable

$$V_k \;:=\; \sum_{i=1}^{r} |\widetilde{z}_k^i| \tag{34}$$

With this notation, the strict dual feasibility condition (30) is equivalent to the event $\{\max_{k \in U^c} V_k < 1\}$.

5 Proof of Theorem 1

We begin by establishing a set of sufficient conditions for deterministic design matrices, as stated in Theorem 1.

5.1 Establishing strict dual feasibility

We begin by obtaining control on the probability of the event $\mathcal{E}(V)$, so as to show that step (C) of the primal-dual witness construction succeeds. Recall that $\Pi_{X_U^i}$ denotes the orthogonal projection onto the range space of X_U^i, and the definition (11) of the incoherence parameter $\gamma \in (0, 1]$. By the mutual incoherence condition (11), we have

$$\max_{k \in U^c}\Big\{\sum_{i=1}^{r} |\frac{1}{n}\big\langle X_k^i,\, X_U^i\,\big(\langle\frac{1}{n}X_U^i,\, X_U^i\rangle\big)^{-1}\widetilde{z}_U^i\big\rangle\Big\} \;\leq\; 1 - \gamma, \tag{35}$$

where we have used the fact that $\sum_{i=1}^r |\widetilde{z}_j^i| = 1$ for each $j \in U$. Recalling that $V_k = \sum_{i=1}^r |\widetilde{z}_k^i|$ and using the definition (33), we have by triangle inequality

$$\mathbb{P}[\max_{k \in U^c} V_k > 1] \;\leq\; \mathbb{P}[\mathbb{A}(\gamma)],$$

where we have defined the event

$$\mathbb{A}(\gamma) \;:=\; \big\{\max_{k \in U^c}\sum_{i=1}^{r} \big|\frac{1}{\lambda_n n}\langle X_k^i,\, (I - \Pi_{X_U^i})w^i\rangle\big| \geq \gamma\big\}. \tag{36}$$

To analyze this remaining probability, for each index $i = 1, \ldots, r$ and $k \in U^c$, define the random variable

$$W_k^i \;:=\; \frac{1}{\lambda_n n}\big\langle X_k^i,\, (I - \Pi_{X_U^i})\big\rangle w^i. \tag{37}$$

Since the elements of the n-vector w^i follow a $N(0, \sigma^2)$ distribution, the variable W_k^i is zero-mean Gaussian with variance $\frac{\sigma^2}{\lambda_n^2 n^2}\langle X_k^i, (I - \Pi_{X_U^i})X_k^i\rangle$. Since $\|X_k^i\|_2^2 \leq 2n$ by assumption and $(I - \Pi_{X_U^i})$ is an orthogonal projection matrix, the variance of each W_k^i is upper bounded by $\frac{2\sigma^2}{\lambda_n^2 n}$. Consequently, for any choice of sign vector $b \in \{-1, +1\}^r$, the variance of the zero-mean Gaussian $\sum_{i=1}^r b_i W_k^i$ is upper bounded by $\frac{2r\sigma^2}{\lambda_n^2 n}$.

Consequently, by taking the union bound over all sign vectors and over indices $k \in U^c$, we have

$$\mathbb{P}[\mathbb{A}(\gamma)] \;=\; \mathbb{P}\big[\max_{k \in U^c} \max_{b \in \{-1,+1\}^r} \sum_{i=1}^r b_i W_k^i > \gamma\big] \;\;\leq\;\; 2\,\exp\big(-\frac{\lambda_n^2 n \gamma^2}{4r\sigma^2} + r + \log p\big).$$

With the choice $\lambda_n^2 \geq \frac{4\xi\sigma^2}{\gamma^2}\frac{r^2 + r\log(p)}{n}$ for some $\xi > 1$, we conclude that

$$\mathbb{P}[\mathcal{E}(V)] \;\;\geq\;\; 1 - 2\exp(-(\xi - 1)[r + \log p]) \;\rightarrow\; 1.$$

By Lemma 2(i), this event implies the uniqueness of the solution $\widehat{B}$, and moreover the inclusion of the supports $S(\widehat{B}) \subseteq S(\overline{B})$, as claimed.

5.2 Establishing ℓ_∞ bounds

We now turn to establishing the claimed ℓ_∞-bound (16) on the difference $\widehat{B} - \overline{B}$. We have already shown that this difference is exactly zero for rows in U^c; it remains to analyze the difference $\Delta_U = \widehat{B}_U - \overline{B}_U$. It suffices to prove the ℓ_∞ bound for the columns Δ_U^i separately, for each $i = 1, \ldots, r$.

We split the analysis of the random variable $\max_{k \in U} |\Delta_k^i|$ into two terms, based on the form of Δ from equation (32), one involving the dual variables $\widetilde{z}_U^i$, and the other involving the observation noise w^i, as follows:

$$\max_{k \in U} |\Delta_k^i| \;\;\leq\;\; \underbrace{\big\|\big(\frac{1}{n}\langle X_U^i, X_U^i\rangle\big)^{-1}\frac{1}{n}\langle X_U^i, w^i\rangle\big\|_\infty}_{T_a^i} + \underbrace{\big\|\big(\langle\frac{1}{n}X_U^i, X_U^i\rangle\big)^{-1}\lambda_n \widetilde{z}_U^i\big\|_\infty}_{T_b^i}.$$

The second term is easy to control: from the characterization of the subdifferential (Lemma 1), we have $\|\widetilde{z}_U^i\|_\infty \leq 1$, so that $T_b^i \leq \lambda_n \|(\langle\frac{1}{n}X_U^i, X_U^i\rangle)^{-1}\|_\infty \;\leq\; D_{\max}\lambda_n$.

Turning to the first term T_a^i, we note that since X_U^i is fixed, the $|U|$-dimensional random vector $Y := \big(\langle\frac{1}{n}X_U^i, X_U^i\rangle\big)^{-1}\frac{1}{n}\langle X_U^i, w^i\rangle$ is zero-mean Gaussian, with covariance $\frac{1}{n}\big(\langle\frac{1}{n}X_U^i, X_U^i\rangle\big)^{-1}$. Therefore, we have $\mathrm{var}(Y_k) \leq \frac{1}{C_{\min} n}$, and can use this in standard Gaussian tail bounds. By applying the union bound twice, first over $k \in U$, and then over $i \in \{1, 2, \ldots, r\}$, we obtain

$$\mathbb{P}[\max_{i=1,\ldots,r} T_a^i \geq t] \;\;\leq\;\; 2\exp(-t^2 n C_{\min}/(2) + \log(rs) + \log r),$$

where we have used the fact that $|U| \leq rs$. Setting $t = \xi\sqrt{\frac{4\log(rs)}{C_{\min} n}}$ yields that

$$\max_{i=1,\ldots,r} \;\; \max_{k \in U} |\Delta_k^i| \;\;\leq\;\; \xi\sqrt{\frac{4}{C_{\min}}\frac{\log rs}{n}} + D_{\max}\lambda_n \;=: b_1(\xi, \lambda_n, n, s),$$

with probability greater than $1 - 2\exp(-(\xi^2 - 1)\log(rs))$, as claimed.

Finally, to establish support recovery, recall that we proved above that Δ^i is bounded by $b_1(\xi, \lambda_n, n, s)$. Hence, as long as $B_{\min}^* > b_1(\xi, \lambda_n, n, s)$, then we are guaranteed that if $\overline{B}_k^i \neq 0$, then $\widehat{B}_k^i \neq 0$.

6 Proof of Theorem 2

We now turn to the proof of Theorem 2, providing sufficient conditions for general Gaussian ensembles. Recall that for $i = 1, 2, \ldots, r$, each $X^i \in \mathbb{R}^{n \times p}$ is a random design matrix, with rows drawn i.i.d. from a zero-mean Gaussian with $p \times p$ covariance matrix Σ^i.

6.1 Establishing strict dual feasibility

Recalling that $V_k = \sum_{i=1}^r |\widetilde{z}_k^i|$ and using the definition (33), we have the decomposition

$$\max_{k \in U^c} |V_k| \quad \leq \quad \underbrace{\max_{k \in U^c} \sum_{i=1}^r |\frac{1}{\lambda_n n} \langle X_k^i, \, (I - \Pi_{X_U^i}) w^i \rangle|}_{M_1} \quad + \quad \underbrace{\max_{k \in U^c} \sum_{i=1}^r \big| \frac{1}{n} \langle X_k^i, \, X_U^i (\frac{1}{n} \langle X_U^i, \, X_U^i \rangle)^{-1} \widetilde{z}_U^i \rangle \big|}_{M_2}.$$

In order to show that $\max_{k \in U^c} |V_k| < 1$ with high probability, we deal with each of these two terms in turn, showing that $M_1 < \gamma/2$, and $M_2 < 1 - \gamma/2$, both with high probability.

In order to bound M_1, we require the following condition on the columns of the design matrices:

Lemma 3. *Let $\sigma_{\max} = \max_i \Sigma$. For $n > 2 \log(rp)$, each column of the design matrices $X^i, i = 1, \ldots, r$ has controlled ℓ_2-norm:*

$$\mathbb{P}\big[\max_{i=1,\ldots,r} \max_{k=1,\ldots,p} \|X_k^i\|_2^2 \leq 2 \sigma_{\max} n \big] \quad \leq \quad 2 \exp\big(- \frac{n}{2} + \log(pr) \big) \; \to 0. \tag{38}$$

This claim follows immediately by union bound and concentration results for χ^2-variates; in particular, the bound (66a) in Appendix E.

Under the condition of Lemma 3, each variable $W_k^i := \frac{1}{\lambda_n n} \langle X_k^i, \, (I - \Pi_{X_U^i}) w^i \rangle$ is zero-Gaussian, with variance at most $\frac{2\sigma^2}{\lambda_n^2 n}$. Consequently, for any choice of signs $b \in \{-1, +1\}^r$, the vector $\sum_{i=1}^r b_i W_k^i$ is zero-mean Gaussian, with variance at most $\frac{2\sigma^2 r}{\lambda_n^2 n}$. Therefore, for any $t > 0$, we have

$$\begin{aligned} \mathbb{P}[\max_{k \in U^c} \sum_{i=1}^r |W_k^i| \geq t] \quad &= \quad \mathbb{P}[\max_{k \in U^c} \max_{b \in \{-1,+1\}^r} \sum_{i=1}^r b_i W_k^i \geq t] \\ &\leq \quad 2 \; \exp\big(- \frac{\lambda_n^2 n}{4 \sigma^2 r} t^2 + r + \log p \big) \end{aligned}$$

Setting $t = \gamma/2$ yields that

$$\mathbb{P}[M_1 \geq \gamma/2] \quad \leq \quad 2 \; \exp\big(- \frac{\lambda_n^2 n}{16 \sigma^2 r} \gamma^2 + r + \log p \big).$$

Lemma 4. *Suppose that the design covariance matrices $\Sigma^i, i = 1, \ldots, r$ satisfy the mutual incoherence condition* (11)*. Then we have*

$$M_2 \quad \leq \quad (1 - \gamma) + \underbrace{\max_{k \in U^c} \sum_{i=1}^r \big| \frac{1}{n} \langle Y_k^i, \, X_U^i (\langle \frac{1}{n} X_U^i, \, X_U^i \rangle)^{-1} \widetilde{z}_U^i \rangle \big|}_{M_2'}, \tag{39}$$

where each random vector $Y_k^i \in \mathbb{R}^{n \times 1}$ has i.i.d. $N(0,1)$ entries, and is independent of w^i and X_U^i.

See Appendix B for the proof of this claim.

It remains to show that the random variable M_2' defined in equation (39) is upper bounded by $\gamma/2$ with high probability. Conditioning on X_U^i and w^i, the scalar random variable $\frac{1}{n}\langle Y_k^i,\, X_U^i(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle)^{-1}\widetilde{z}_U^i\rangle$ is zero-mean Gaussian, with variance upper bounded as

$$\frac{1}{n}\langle \widetilde{z}_U^i,\, (\langle \frac{1}{n}X_U^i,\, X_U^i\rangle)^{-1}\widetilde{z}_U^i\rangle \;\;\leq\;\; \frac{\|\widetilde{z}_U^i\|_2^2}{C_{\min}\, n}.$$

Recalling that $\sum_{i=1}^r |\widetilde{z}_j^i| = 1$, for any choice of signs $b \in \{-1,+1\}^r$, the variable

$$\sum_{i=1}^r b_i \frac{1}{n}\langle Y_k^i,\, X_U^i(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle)^{-1}\widetilde{z}_U^i\rangle$$

is zero-mean Gaussian, with variance at most $\frac{rs}{C_{\min}\, n}$. Therefore, we have

$$\begin{aligned}\mathbb{P}[M_2' \geq \gamma/2] &\leq \mathbb{P}\big[\max_{k\in U^c}\max_{b\in\{-1,+1\}^r} |\sum_{i=1}^r b_i \frac{1}{n}\langle Y_k^i,\, X_U^i(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle)^{-1}\widetilde{z}_U^i\rangle| \geq \gamma/2\big] \\ &\leq 2\exp(-\frac{C_{\min} n}{8rs}\gamma^2 + r + \log p).\end{aligned}$$

This probability vanishes faster than $2\exp\big\{-\kappa(r+\log p)\big\} \to 0$, as long as

$$n \;\;>\;\; \frac{8\,\kappa\, r}{C_{\min}\gamma^2}\, s\big(r+\log p\big)$$

6.2 Establishing ℓ_∞ bounds

We now turn to establishing the claimed ℓ_∞-bound (16) on the difference $\widehat{B} - \overline{B}$. As in the analogous portion of the proof of Theorem 1, we use the decomposition

$$\max_{k\in U}|\Delta_k^i| \;\;\leq\;\; \underbrace{\big\|\big(\frac{1}{n}\langle X_U^i,\, X_U^i\rangle\big)^{-1}\frac{1}{n}\langle X_U^i,\, w^i\rangle\big\|_\infty}_{T_a^i} + \underbrace{\big\|\big(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle\big)^{-1}\lambda_n\widetilde{z}_U^i\big\|_\infty}_{T_b^i}.$$

In the setting of random design matrices, a bit more work is required to control these terms.

Beginning with the second term, by triangle inequality, we have

$$\begin{aligned}T_b^i &\leq \big\|\big[\big(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle\big)^{-1} - (\Sigma_{UU}^i)^{-1}\big)\lambda_n\widetilde{z}_U^i\big]\|_\infty + \big\|(\Sigma_{UU}^i)^{-1}\lambda_n\widetilde{z}_U^i\big\|_\infty \\ &\leq \|\big[\big(\langle \frac{1}{n}X_U^i,\, X_U^i\rangle\big)^{-1} - (\Sigma_{UU}^i)^{-1}\big)\|_2\lambda_n\sqrt{s} + D_{\max}\lambda_n\end{aligned}$$

where we have used the facts that $\|\widetilde{z}_U^i\|_2 \leq \sqrt{s}$, since $\widetilde{z}_U^i$ belongs to the sub-differential of the block ℓ_1/ℓ_∞ norm (see Lemma 1) so that $\sum_{i=1}^r |\widetilde{z}_j^i| \leq 1$ for all $j \in U$. By, concentration bounds for eigenvalues of Gaussian random matrices (see equation (69b) in Appendix E), we conclude that

$$T_b^i \;\;\leq\;\; 4\lambda_n\sqrt{s}\,\sqrt{\frac{s}{n}} + D_{\max}\lambda_n \;=\; \lambda_n\,\big[\frac{4s}{\sqrt{n}} + D_{\max}\big].$$

Now consider the first term T_a^i: if we condition on X_U^i, then the $|U|$-dimensional random vector $Y := \big(\langle \frac{1}{n} X_U^i, X_U^i \rangle\big)^{-1} \frac{1}{n} \langle X_U^i, w^i \rangle$ is zero-mean Gaussian, with covariance $\frac{1}{n}\big(\langle \frac{1}{n} X_U^i, X_U^i \rangle\big)^{-1}$. By concentration bounds for eigenvalues of Gaussian random matrices (see equation (69b) in Appendix E), we have

$$\begin{aligned}
\frac{1}{n} |\!|\!| \big(\langle \frac{1}{n} X_U^i, X_U^i \rangle\big)^{-1} |\!|\!|_2 &\leq \frac{1}{n} \Big\{ |\!|\!| \big(\langle \frac{1}{n} X_U^i, X_U^i \rangle\big)^{-1} - (\Sigma_{UU}^i)^{-1} |\!|\!|_2 + |\!|\!| (\Sigma_{UU}^i)^{-1} |\!|\!|_2 \Big\} \\
&\leq \frac{4}{C_{\min} n} \sqrt{\frac{rs}{n}} + \frac{1}{C_{\min} n} \\
&\leq \frac{5}{C_{\min} n},
\end{aligned}$$

since $rs/n \leq 1$. Therefore, we have shown that the variance of each element of Y is upper bounded by $5/(C_{\min} n)$, so that we can apply standard Gaussian tail bounds. By applying the union bound twice, first over $k \in U$, and then over $i \in \{1, 2, \ldots, r\}$, we obtain

$$\mathbb{P}[\max_{i=1,\ldots,r} T_a^i \geq t] \leq 2 \exp(-t^2 n C_{\min}/(50) + \log |U| + \log r).$$

Setting $t = \xi \sqrt{\frac{100 \log(rs)}{C_{\min} n}}$ yields that

$$\begin{aligned}
\mathbb{P}[\max_{i=1,\ldots,r} T_a^i \geq t] &\leq 2 \exp\big\{ -2\xi^2 \log(rs) + \log(rs) + \log r \big\} \\
&\leq 2 \exp\big\{ -2(\xi^2 - 1) \log(rs) \big\},
\end{aligned}$$

where we have used the fact that $|U| \leq rs$. Combining the pieces, we conclude that

$$\max_{k \in U} |\Delta_k^i| \leq \xi \sqrt{\frac{100 \log(rs)}{C_{\min} n}} + \lambda_n \big[\frac{4s}{\sqrt{n}} + D_{\max} \big],$$

with probability greater than

$$1 - 2 \exp\big\{ -2(\xi^2 - 1) \log(rs) \big\} - c_1 \exp(-c_2 n),$$

as claimed.

7 Proof of Theorem 3

We now turn to the proof of the phase transition predicted by Theorem 3, which applies to random design matrices X^1 and X^2 drawn from the standard Gaussian ensemble. This proof requires significantly more technical work than the preceding two proofs, since we need to control all the constants exactly, and to establish both necessary and sufficient conditions on the sample size.

7.1 Proof of Theorem 3(a)

We begin with the achievability result. Our proof parallels that of Theorems 1 and 2, in that we first establish strict dual feasibility, and then turn to proving ℓ_∞ bounds and exact support recovery.

7.1.1 Establishing strict dual feasibility

Recalling that $V_k = \sum_{i=1}^2 |\widetilde{z}_k^i|$, we have

$$\max_{k \in U^c} |V_k| \ \leq \ M_1 \ + \ M_2,$$

where the random variables M_1 and M_2 were defined at the start of Section 6.1. In order to prove that $\max_{k \in U^c} |V_k| < 1$ with high probability for the values of n, s, and p, we will first establish that $M_1 < \epsilon/2$ and $M_2 < 1 - \epsilon$ for an appropriately chosen value of ϵ.

By the results from the previous section, we have $M_1 < \epsilon/2$ with probability

$$\mathbb{P}[\max_{k \in U^c} \sum_{i=1}^2 |W_k^i| \geq \epsilon/2] \quad \leq \quad 2 \ \exp\big(- \frac{\lambda_n^2 n \epsilon^2}{32 \sigma^2} + 2 + \log p \big)$$

Recall that

$$M_2 = \max_{k \in U^c} \sum_{i=1}^2 \big| \frac{1}{n} \big\langle X_k^i, \, X_U^i (\frac{1}{n} \langle X_U^i, \, X_U^i \rangle)^{-1} \widetilde{z}_U^i \big\rangle \big|,$$

and that $X_{U^c}^i$ is independent of X_U^i and w^i. We will show that $M_2 < 1 - \epsilon$ with high probability by using results on Gaussian extrema. Conditioning on $(X_U, w, \widetilde{z}_U^1)$, the random variable $Y_k^i = \frac{1}{n} \big\langle X_k^i, \, X_U^i (\frac{1}{n} \langle X_U^i, \, X_U^i \rangle)^{-1} \widetilde{z}_U^i \big\rangle$ is zero-mean with variance upper-bounded as

$$\frac{1}{n} \big\langle \widetilde{z}_U^i, \, (\langle \frac{1}{n} X_U^i, \, X_U^i \rangle)^{-1} \widetilde{z}_U^i \big\rangle \quad \leq \quad |\!|\!| (\langle \frac{1}{n} X_U^i, \, X_U^i \rangle)^{-1} |\!|\!|_2^2 \frac{\| \widetilde{z}_U^i \|_2^2}{n},$$

Under the given conditioning, the random variables Y_k^1 and Y_k^2 are independent and for any sign vector $b \in \{-1, +1\}^2$, the random variable $\sum_{i=1}^2 b_i Y_k^i$ is Gaussian, zero-mean with variance upper bounded as

$$\sum_{i=1}^2 |\!|\!| (\langle \frac{1}{n} X_U^i, \, X_U^i \rangle)^{-1} |\!|\!|_2^2 \frac{\| \widetilde{z}_U^i \|_2^2}{n}$$

By Lemma 13, $|\!|\!| (\langle \frac{1}{n} X_U^i, \, X_U^i \rangle)^{-1} |\!|\!|_2^2 \leq (1 + \delta)$ with probability at least $1 - c_1 \exp - c_2 n$ for sufficiently large s and n under the given scaling for each i. Hence, $\sum_{i=1}^2 b_i Y_k^i$ is normal, zero-mean, with variance upper bounded as

$$\frac{(1 + \delta)}{n} \sum_{i=1}^2 \| \widetilde{z}_U^i \|_2^2$$

Recall that $\widetilde{Z}_U$ was obtained from Step (B) of the Prima-dual witness construction. The next lemma provides control over $\sum_{i=1}^2 \| \widetilde{z}_U^i \|_2^2$.

Lemma 5. *Under the assumptions of Theorem 3 and Corollary 1, if $\lambda_n^2 n \to +\infty$ and $s/n \to 0$, then $\| \widetilde{z}_U^1 \|_2^2$ is concentrated: for all $\delta > 0$, we have that for sufficiently large s and n*

$$\mathbb{P}\big[\| \widetilde{z}_U^1 \|_2^2 + \| \widetilde{z}_U^2 \|_2^2 \leq (1 - \delta) \frac{s}{2} \big\{ (4 - 3\alpha) + \frac{1}{\lambda_n^2 s} \| \overline{B}_{\mathrm{diff}} \|_2^2 \big\} \big] \quad \to \quad 0, \qquad \textit{and} \tag{40a}$$

$$\mathbb{P}\big[\| \widetilde{z}_U^1 \|_2^2 + \| \widetilde{z}_U^2 \|_2^2 \geq (1 + \delta) \frac{s}{2} \big\{ (4 - 3\alpha) + \frac{1}{\lambda_n^2 s} \| \overline{B}_{\mathrm{diff}} \|_2^2 \big\} \big] \quad \leq \quad c_1 \exp(-c_2 n), \tag{40b}$$

See Appendix C for the proof of this claim.

Now, by applying the union bound and using Gaussian tail bounds, we obtain that the probability $\mathbb{P}[M_2 \geq 1-\epsilon]$ is upper bounded by

$$c_1 \exp(-c_2 n) \; + \; 4\exp\big(-(1-\epsilon)^2 n/[(1+\delta)s\big\{(4-3\alpha) + \frac{1}{\lambda_n^2 s}\|\overline{B}_{\mathrm{diff}}\|_2^2\big\}] + \log(p-(2-\alpha)s)\big),$$

which goes to 0 as $n \to \infty$ under the condition

$$n \;\; > \;\; [(1+\delta)s\big\{(4-3\alpha) + \frac{1}{\lambda_n^2 s}\|\overline{B}_{\mathrm{diff}}\|_2^2\big\}]/(1-\epsilon)^2 \log(p-(2-\alpha)s).$$

7.2 Proof of Theorem 3(b)

We now turn to the proof of the converse claim in Theorem 3. We establish the claim by contradiction. We show that if a solution $\widehat{B}$ exists such that $\widehat{B}_{U^c} = 0$, then under the stated upper bound on the sample size n, there exists some $\epsilon > 0$ such that $\mathbb{P}[\max_{k \in U^c}(|\widetilde{z}_k^1| + |\widetilde{z}_k^2|) > 1+\epsilon]$ converges to one. From the definition (33), we see that conditioned on $(X_U, w, \widetilde{z}_U^1)$, the variables $\{\widetilde{z}_k^1, k \in U^c\}$ are i.i.d. zero-mean Gaussians, with variance given by

$$\mathrm{var}(\widetilde{z}_k^1) := \|\frac{1}{\lambda_n n}\Pi_{U^\perp} w - \frac{1}{n} X_U (\frac{1}{n} X_U^T X_U)^{-1} \widetilde{z}_U^1\|_2^2.$$

By orthogonality, we have $\mathrm{var}(\widetilde{z}_k^1) = \|\frac{1}{\lambda_n n}\Pi_{U^\perp} w\|_2^2 + \|\frac{1}{n} X_U (\frac{1}{n} X_U^T X_U)^{-1} \widetilde{z}_U^1\|_2^2$, so that (using the idempotency of projection operators), we have

$$\mathrm{var}(\widetilde{z}_k^1) \;\; \geq \;\; \sigma^2 \;\; := \;\; \max\left\{ \frac{1}{\lambda_n^2 n} \, \frac{\|\Pi_{U^\perp} w\|_2^2}{n}, \; \lambda_{\min}((\frac{1}{n} X_U^T X_U)^{-1}) \, \frac{\|\widetilde{z}_U^1\|_2^2}{n} \right\}. \tag{41}$$

Note that $\sigma^2 = \sigma^2(X_U, w, \widetilde{z}_U^1)$ is a scalar random variable, but fixed under the conditioning. Turning to the variables $\{\widetilde{z}_k^2, k \in U^c\}$, a similar argument shows that have $\mathrm{var}(\widetilde{z}_k^2) \geq \widetilde{\sigma}^2$, where $\widetilde{\sigma}^2 = \widetilde{\sigma}^2(\overline{X}_U, \overline{w}, \widetilde{z}_U^2)$ is the analogous random variable.

For $k \in U^c$, let $\widetilde{z}_k^1 \sim N(0, \sigma^2)$ and $\widetilde{z}_k^2 \sim N(0, \widetilde{\sigma}^2)$. We then have

$$\begin{aligned}
\mathbb{P}[\max_{k \in U^c}(|\widetilde{z}_k^1| + |\widetilde{z}_k^2|) > (1+\epsilon)] \;\; &\stackrel{(a)}{\geq} \;\; \mathbb{P}[\max_{k \in U^c} |\widetilde{z}_k^1| + |\widetilde{z}_k^2| > 1+\epsilon] \\
&\geq \;\; \mathbb{P}[\max_{k \in U^c}(\widetilde{z}_k^1 + \widetilde{z}_k^2) > 1+\epsilon] \\
&\stackrel{(b)}{=} \;\; \mathbb{P}[\max_{k \in U^c} Z_k > 1+\epsilon],
\end{aligned}$$

where $Z_j \sim N(0, \sigma^2 + \widetilde{\sigma}^2)$. Here inequality (a) follows because σ^2 and $\widetilde{\sigma}^2$ are lower bounds on the variances of $\{\widetilde{z}_k^1, k \in U^c\}$ and $\{\widetilde{z}_k^2, k \in U^c\}$ respectively, and equality (b) follows since $\widetilde{z}_j^1$ and $\widetilde{z}_j^2$ are independent zero-mean Gaussians with variances σ^2 and $\widetilde{\sigma}^2$, respectively.

To simplify notation, let $N = |U^c| = p - (2-\alpha)s$. By standard results for Gaussian maxima [13], for any $\delta > 0$, there exists an integer $N(\delta)$ such that for all $N \geq N(\delta)$,

$$\mathbb{E}[\max_{j \in U^c} Z_j] \;\; \geq \;\; (1-\delta)\sqrt{2(\sigma^2 + \widetilde{\sigma}^2)\log N}.$$

Moreover, the maximum function is Lipschitz, so that by Gaussian concentration for Lipschitz functions [13, 12], for any $\eta > 0$, we have

$$\mathbb{P}\big[\max_{j \in U^c} Z_j \leq \mathbb{E}[\max_{j \in U^c} Z_j] - \eta\big] \;\; \leq \;\; \exp\big(-\frac{\eta^2}{2(\sigma^2 + \widetilde{\sigma}^2)}\big).$$

Combining these two statements yields that for all $N \geq N(\delta)$, we have

$$\mathbb{P}\big[\max_{j \in U^c} Z_j \leq (1-\delta)\sqrt{2(\sigma^2 + \widetilde{\sigma}^2)\log N} - \eta\big] \quad \leq \quad \exp\big(-\frac{\eta^2}{2(\sigma^2+\widetilde{\sigma}^2)}\big). \tag{42}$$

It remains to show that there exists some $\epsilon > 0$ such that $\mathbb{P}[\max_{k \in U^c} Z_k \leq 1+\epsilon]$ converges to zero.

Case 1: First suppose that $\lambda_n^2 n = \mathcal{O}(1)$. In this case, we have $\sigma^2 = \Omega\big(\frac{\|\Pi_{U^\perp} w\|_2^2}{n}\big)$. With probability greater than $1 - c_1\exp(-c_2 n)$, this quantity is lower bounded by a constant, using concentration for χ^2-variates. In this case, $\sqrt{2(\sigma^2+\widetilde{\sigma}^2)\log N} - \eta \to +\infty$ w.h.p., so that the result follows trivially.

Case 2: Otherwise, we must have $\lambda_n^2 n \to +\infty$. Under this condition, we now establish a lower bound on σ^2 that holds with high probability; it will be seen that a similar lower bound holds for $\widetilde{\sigma}^2$. We begin by noting the lower bound $\sigma^2 \geq \frac{\|\widetilde{z}_U^1\|_2^2}{n}\lambda_{\min}((\frac{1}{n}X_U^T X_U)^{-1})$. To control the minimum eigenvalue, define the event

$$\mathcal{T}(X_U) := \big\{X_U \;\mid\; \lambda_{\min}((\frac{1}{n}X_U^T X_U)^{-1}) \geq (1+\sqrt{s/n})^{-2}\big\}. \tag{43}$$

By standard random matrix concentration arguments (see Appendix E), for some fixed $c > 0$, we are guaranteed that $\mathbb{P}[\mathcal{T}^c(X_U)] \leq 2\exp(-cn)$. Consequently, conditioned on $\mathcal{T}(X_U)$, we have

$$\sigma^2 + \widetilde{\sigma}^2 \;\geq\; \frac{\|\widetilde{z}_U^1\|_2^2 + \|\widetilde{z}_U^2\|_2^2}{n}\,(1+\sqrt{s/n})^{-2}. \tag{44}$$

From Lemma 5, we note that if $s/n = o(1)$, then for any $\delta > 0$, we have the lower bound

$$\sigma^2 + \widetilde{\sigma}^2 \;\geq\; (1-\delta)\frac{s}{2n}\big\{(4-3\alpha) + (\Delta(\overline{B}, \lambda_n))^2\big\}(1-o(1)). \tag{45}$$

The following result is the final step in the proof of Theorem 3(b).

Lemma 6. *Suppose that $\lambda_n^2 n \to +\infty$. Under this condition:*

(a) If $\frac{s}{n} = \Omega(1)$, then $\mathbb{P}[\max_{k \in U^c} Z_k \leq 2] \to 0$.

(b) If $\frac{s}{n} \to 0$, then there exists some $\epsilon > 0$ such that $\mathbb{P}[\max_{k \in U^c} Z_k \leq 1+\epsilon] \to 0$.

Proof. (a) If $\frac{s}{n}$ is bounded below by some constant $c > 0$, then we have

$$\sigma^2 \;\geq\; (1-\alpha)\frac{s}{n} \;\geq\; (1-\alpha)c,$$

which implies that $(\sigma^2 + \widetilde{\sigma}^2)\log N \to +\infty$. Thus, setting $\delta = 1/4$ and $\eta = \frac{1}{2}\sqrt{2(\sigma^2+\widetilde{\sigma}^2)\log N}$ in equation (42) yields that (for N sufficiently large):

$$\begin{aligned}\mathbb{P}\big[\max_{k \in U^c} Z_k \leq (\frac{1}{2}-\delta)\sqrt{2(\sigma^2+\widetilde{\sigma}^2)\log N}\big] \quad &= \quad \mathbb{P}\big[\max_{k \in U^c} Z_k \leq \frac{1}{4}\,\sqrt{2(\sigma^2+\widetilde{\sigma}^2)\log N}\big] \\ &\leq \quad \exp\big(-\frac{\log N}{4}\big) \to 0.\end{aligned}$$

Since $\frac{1}{4}\sqrt{2(\sigma^2+\widetilde{\sigma}^2)\log N} \geq 2$ for N large enough, the claim follows.

(b) In this case, we may apply the lower bound (45), so that, for any $\delta > 0$, we have

$$\sigma^2 + \widetilde{\sigma}^2 \ \geq \ (1-\delta)\frac{s}{2n}\big\{(4-3\alpha) + (\Delta(\overline{B}, \lambda_n))\big\}(1 - o(1))$$

with high probability. Since $n < (1-\nu)[(4-3\alpha) + (\Delta(\overline{B}, \lambda_n))]s \log N$ by assumption, we have

$$\begin{aligned}\sqrt{2(\sigma^2 + \widetilde{\sigma}^2)\log N} &\geq (1 - o(1)) \sqrt{(1-\delta)\frac{s}{n}\big\{(4-3\alpha) + (\Delta(\overline{B}, \lambda_n))\big\}\log N} \\ &\geq (1 - o(1)) \sqrt{\frac{1-\delta}{1-\nu}}.\end{aligned}$$

Consequently, from equation (42), for any $\eta > 0$ and $\delta > 0$, we have for all $N \geq N(\delta)$,

$$\mathbb{P}\big[\max_{k \in U^c} Z_k \leq (1-\delta)\,(1 - o(1))\,\frac{1}{\sqrt{1-\nu}} - \eta\big] \ \leq \ \exp\big(-\frac{\eta^2}{2(\sigma^2 + \widetilde{\sigma}^2)}\big). \tag{46}$$

Since $\nu > 0$, we may choose $\eta, \delta > 0$ sufficiently small so that for sufficiently large choices of (s, n), we have

$$(1-\delta)\,(1 - o(1))\frac{1}{\sqrt{1-\nu}} - \eta \ \geq \ 1 + \epsilon$$

for some $\epsilon > 0$. Since from Lemma 5, the condition $s/n = o(1)$ implies that $\sigma^2 + \widetilde{\sigma}^2 = o(1)$ w.h.p, we thus conclude that, using these choices of η and δ, we have

$$\mathbb{P}[\max_{k \in U^c} Z_k \leq 1 + \epsilon] \ \leq \ o(1) + \exp\big(-\frac{\eta^2}{2(\sigma^2 + \widetilde{\sigma}^2)}\big) \ \to \ 0,$$

as claimed. □

8 Discussion

In this paper, we provided a number of theoretical results that provide a sharp characterization of when, and if so by how much the use of block ℓ_1/ℓ_∞ regularization actually leads improvements in statistical efficiency in the problem of multivariate regression. As suggested in a body of past work, the use of block ℓ_1/ℓ_∞ regularization is well-motivated in many application contexts. However, since it involves greater computational cost than more naive approaches, the question of whether this greater computational price yields statistical gains is an important one.

This paper assessed statistical efficiency in terms of the number of samples required to recover the support exactly; however, one could imagine studying the same issue for related loss functions (e.g., ℓ_2-loss or prediction loss), and it would be interesting to see if the results were qualitatively similar or not. Our results demonstrate that some care needs to be exercised in the application of ℓ_1/ℓ_∞ regularization. Indeed, it can yield improved statistical efficiency when the regression matrix exhibits structured sparsity, with high overlaps among the sets of active coefficients within each column. However, our analysis shows that these improvements are quite sensitive to the exact structure of the regression matrix, and how well it aligns with the regularizing norm. When this alignment is not high enough, then the use of ℓ_1/ℓ_∞ can actually impair performance relative to more naive (and less computationally intensive) schemes based on ℓ_1-regularization, such as the Lasso. Moreover, whether or not the ℓ_1/ℓ_∞ yields statistical improvements is very sensitive to the actual magnitudes of the different

regression problems. In comparison to related results obtained by Obozinski et al. [23] on block ℓ_1/ℓ_2 regularization, the block ℓ_1/ℓ_∞ exhibits some fragility, in that the conditions under which it actually improves statistical efficiency are delicate and easily violated. An interesting open direction is study whether or not it is possible to develop computationally efficient methods that are *fully adaptive* to the sparsity overlap–namely, methods that behave like ordinary ℓ_1-regularization when there is no or little shared sparsity, and behave like block regularization schemes in the presence of shared sparsity.

A Recovering individual signed supports

In this appendix, we discuss some issues associated with recovering individual signed supports. We begin by observing that once the support union U has been recovered, one can restrict the regression problem to this subset U, and then apply Lasso to each problem separately (with substantially lower cost, since each problem is now low-dimensional) in order to recover the individual signed supports. If one is not willing to perform some extra computation in this way, then the the interpretation of Theorems 1 and 2—in terms of recovering the individual signed supports—requires a more delicate treatment, which we discuss in this appendix.

Interestingly, the structure of the block ℓ_1/ℓ_∞ norm permits two ways in which to recover the individual signed supports.

ℓ_1/ℓ_∞ primal recovery: Solve the block-regularized program (6), thereby obtaining a (primal) optimal solution $\widehat{B} \in \mathbb{R}^{p \times r}$. Estimate the support union via $\widehat{U} := \bigcup_{i=1,\ldots,r} S(\widehat{\beta}^i)$, and and estimate the signed support vectors via

$$[\mathbb{S}_{\mathrm{pri}}(\widehat{\beta}^i)]_k \quad := \quad \mathrm{sign}(\widehat{\beta}^i_k). \tag{47}$$

ℓ_1/ℓ_∞ dual recovery: Solve the block-regularized program (6), thereby obtaining an primal solution $\widehat{B} \in \mathbb{R}^{p \times r}$. For each row $k = 1, \ldots, p$, compute the set $\mathbb{M}_k := \arg\max_{i=1,\ldots,r} |\widehat{\beta}^i_k|$. Estimate the support union via $\widehat{U} = \bigcup_{i=1,\ldots,r} S(\widehat{\beta}^i)$, and estimate the signed support vectors

$$[\mathbb{S}_{\mathrm{dua}}(\widehat{\beta}^i_k)] \quad = \quad \begin{cases} \mathrm{sign}(\widehat{\beta}^i_k) & \text{if } i \in \mathbb{M}_k \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

The procedure (48) corresponds to estimating the signed support on the basis of a dual optimal solution associated with the optimal primal solution.

The dual signed support recovery method (48) is more conservative in estimating the individual support sets. In particular, for any given $i \in \{1, \ldots, r\}$, it only allows an index k to enter the signed support estimate $\mathbb{S}_{\mathrm{dua}}(\widehat{\beta}^i)$ when $|\widehat{\beta}^i_k|$ achieves the maximum magnitude (possibly non-unique) across all indices $i = 1, \ldots, r$. Consequently, unlike the primal estimator (48), a corollary of Theorem 1 guarantees that the dual signed support method (48) never suffers from false inclusions in the signed support set. On the other hand, unlike the primal estimator, it may incorrectly exclude indices of some supports—that is, it may exhibit false exclusions.

To provide a concrete illustration of this distinction, suppose that $p = 4$ and $r = 3$, and that the true matrix $\overline{B}$ and estimate take the following form:

$$\overline{B} = \begin{bmatrix} 2 & 0 & -3 \\ 2 & 4 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \widehat{B} = \begin{bmatrix} 1.9 & 0.1 & -2.9 \\ 1.7 & 3.9 & -0.1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Consistent with the claims of Theorem 1, the estimate $\widehat{B}$ correctly recovers the support union—viz. $S(\widehat{B}) = \widehat{U} = \{1, 2\} = S(\overline{B})$. The primal (47) and dual (48) methods return the following estimates of the individual signed supports:

$$\mathbb{S}_{\text{pri}}(\widehat{B}) = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{and} \quad \mathbb{S}_{\text{dua}}(\widehat{B}) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Consequently, the primal estimate includes false non-zeros in positions $(1, 2)$ and $(2, 3)$, whereas the dual estimate includes false zeros in positions $(1, 1)$ and $(2, 1)$.

We note that it is possible to ensure that under some conditions that the dual support method (48) will correctly recover each of the individual signed supports, without any incorrect exclusions. However, as illustrated by Theorem 3 and Corollary 1, doing so requires additional assumptions on the size of the gap $|\bar{\beta}^i_k| - |\bar{\beta}^j_k|$ for indices $k \in B := S(\bar{\beta}^i) \cap S(\bar{\beta}^j)$.

B Proof of Lemma 4

Note that conditioned X_U, the rows of the random matrix $X^i_{U^c}$ are i.i.d. Gaussian random vectors with mean $\langle \Sigma^i_{U^cU}(\Sigma^i_{UU})^{-1}, X^i_U \rangle$ and covariance

$$\Sigma^i_{U^c|U} \quad = \quad \Sigma^i_{U^cU^c} - \Sigma^i_{U^cU}(\Sigma^i_{UU})^{-1}\Sigma^i_{UU^c}.$$

$$\frac{1}{n}\langle X^i_{U^c}, X^i_U(\langle \frac{1}{n}X^i_U, X^i_U\rangle)^{-1}\rangle \quad \stackrel{d}{=} \quad \Sigma^i_{U^cU}(\Sigma^i_{UU})^{-1}\widetilde{z}^i_U + \frac{1}{n}\langle Y^i_{U^c}, X_U(\langle \frac{1}{n}X^i_U, X^i_U\rangle)^{-1}\rangle$$

where $Y^i_{U^c} \sim N(0, \Sigma^i_{U^c|U})$.

Using these expressions and triangle inequality, we obtain that M is upper bounded by

$$\max_{k \in U^c}\{\sum_{i=1}^r \|e_k^T\Sigma^i_{U^cU}(\Sigma^i_{UU})^{-1}\|_1\} + \max_{k\in U^c}\sum_{i=1}^r \big|\frac{1}{n}\langle Y^i_k, X^i_U(\langle \frac{1}{n}X^i_U, X^i_U\rangle)^{-1}\widetilde{z}^i_U\rangle\big|.$$

Applying the mutual incoherence assumption (19), we obtain

$$M \quad \leq \quad (1-\gamma) + \max_{k\in U^c}\sum_{i=1}^r \big|\frac{1}{n}\langle Y^i_k, X^i_U(\langle \frac{1}{n}X^i_U, X^i_U\rangle)^{-1}\widetilde{z}^i_U\rangle\big|,$$

as claimed.

C Proof of Lemma 5

Recall that $\widetilde{z}^1_U = (\widetilde{z}^1_{B^c}, \widetilde{z}^1_B)$, $\|\widetilde{z}^1_{B^c}\|_2^2 = (1-\alpha)s$, and that B is the set where $|\widehat{\beta}^1_B| = |\widehat{\beta}^2_B|$. Thus, the claim is equivalent to showing that $\|\widetilde{z}^1_B\|_2^2$ is concentrated. If $\alpha = 0$, then the claim is trivial, so that we may assume that $\alpha > 0$.

Recall that

$$S^1\widetilde{z}^1_B = \frac{1}{\lambda_n}\big\{M^2\big[M^1+M^2\big]^{-1}f^1 - M^1\big[M^2+M^1\big]^{-1}f^2\big\} + M^1\big[M^1+M^2\big]^{-1}\vec{1} - \frac{1}{\lambda_n}[(M^1)^{-1}+(M^2)^{-1}]^{-1}\overline{B}_{\text{diff}}. \quad (49)$$

Using $\|\!|\cdot|\!\|_2$ to denote the spectral norm, we first claim that as long as $s/n \to 0$, then the following events hold with probability greater than $1 - c_1\exp(-c_2 n)$:

$$\|\!|M^1 - I|\!\|_2 = \mathcal{O}(\sqrt{s/n}), \quad (50a)$$

$$\|\!|[(M^1)^{-1}+(M^2)^{-1}]^{-1} - I/2|\!\|_2 = \mathcal{O}(\sqrt{s/n}), \quad \text{and} \quad (50b)$$

$$\|\!|M^1\big[M^1+M^2\big]^{-1} - I/2|\!\|_2 = \mathcal{O}(\sqrt{s/n}), \quad (50c)$$

as well as the analogous events with M^1 and M^2 interchanged.

To verify the bound (50a), we first diagonalize the projection matrix. All of its eigenvalues are 0 or 1, and it has rank $(n-s)$ w.p. one, so that we may write $\Pi_{B^{c\perp}} = U^T D U$ for some orthogonal matrix U, and the diagonal matrix $D = \text{diag}\{1_{n-s}, 0_s\}$,

$$M = n^{-1}X_B^T U^T D U X_B.$$

But the projection $\Pi_{B^{c\perp}}$ is independent of X_B, which implies that the random rotation matrix U is independent of X_B, and hence $X_B \stackrel{d}{=} UX_B$. Since D is diagonal with $(n-s)$ ones and s zeros, $M \stackrel{d}{=} n^{-1}W^TW$, where $W \in \mathbb{R}^{(n-s)\times|B|}$ is a standard Gaussian random matrix. Consequently, we have

$$\begin{aligned}
\|\!|M - I|\!\|_2 &\stackrel{d}{=} \|\!|n^{-1}W^TW - I|\!\|_2 \\
&\leq \big|\frac{n-s}{n} - 1\big|\,\|\!|\frac{1}{n-s}W^TW|\!\|_2 + \|\!|\frac{1}{n-s}W^TW - I|\!\|_2 \\
&= \mathcal{O}(\sqrt{s/n}),
\end{aligned}$$

since $\|\!|W^TW/(n-s)|\!\|_2 = \mathcal{O}(1)$, and

$$\|\!|\frac{1}{n-s}W^TW - I|\!\|_2 = \mathcal{O}(\sqrt{\frac{s}{n-s}}) = \mathcal{O}(\sqrt{s/n}),$$

using concentration arguments for random matrices (see Lemma 13 in Appendix E).

For (50b) we may use the triangle inequality and the submultiplicativity of the norm so that

$$\begin{aligned}
\|\!|[M^{-1}+\widetilde{M}^{-1}]^{-1} - I/2|\!\|_2 &= \|\!|[M^{-1}+\widetilde{M}^{-1}]^{-1}(I - [M^{-1}+\widetilde{M}^{-1}]/2)|\!\|_2 \\
&\leq \|\!|[M^{-1}+\widetilde{M}^{-1}]^{-1}|\!\|_2\,\|\!|I - [M^{-1}+\widetilde{M}^{-1}]/2|\!\|_2 \\
&\leq \frac{1}{2}\big\{\|\!|I/2 - M^{-1}/2|\!\|_2 + \|\!|I/2 - \widetilde{M}^{-1}/2|\!\|_2\big\}\|\!|[M^{-1}+\widetilde{M}^{-1}]^{-1}|\!\|_2 \\
&= \|\!|[M^{-1}+\widetilde{M}^{-1}]^{-1}|\!\|_2\,\mathcal{O}(\sqrt{s/n}),
\end{aligned}$$

Finally, since $\|\|[M^{-1} + \widetilde{M}^{-1}]^{-1}\|\|_2 = \mathcal{O}(1)$, equation (50b) is valid.

In order to establish the bound (50c), we have

$$\begin{aligned} \|\|M[M + \widetilde{M}]^{-1} - I/2\|\|_2 &= \|\|(M/2 - \widetilde{M}/2)[M + \widetilde{M}]^{-1}\|\|_2 \\ &\leq \frac{1}{2}\big\{\|\|M - I\|\|_2 + \|\|\widetilde{M} - I\|\|_2\big\} \, \|\|[M + \widetilde{M}]^{-1}\|\|_2 \\ &= \|\|[M + \widetilde{M}]^{-1}\|\|_2 \, \mathcal{O}(\sqrt{s/n}). \end{aligned}$$

Since $\|\|[M + \widetilde{M}] - 2I\|\|_2 = \mathcal{O}(\sqrt{s/n}) \to 0$, we have $\|\|[M + \widetilde{M}]^{-1}\|\|_2 = \mathcal{O}(1)$, which establishes the claim (50c).

We are now ready to establish the claims of the lemma. From the representation (49), we apply triangle inequality and our bounds on spectral norms, thereby obtaining

$$\begin{aligned} \sqrt{\|\widetilde{z}_B^1\|_2^2 + \|\widetilde{z}_B^2\|_2^2} &\leq \sqrt{\|\frac{\vec{1}}{2} - \frac{1}{2\lambda_n}(|\bar{\beta}_B^2| - |\bar{\beta}_B^1|)\|_2^2 + \|\frac{\vec{1}}{2} + \frac{1}{2\lambda_n}(|\bar{\beta}_B^2| - |\bar{\beta}_B^1|)\|_2^2} + 2\|r\| \\ &\leq \sqrt{s}\Big\{\sqrt{\frac{\alpha}{2} + \frac{1}{2s\lambda_n^2}\||\bar{\beta}_B^1| - |\bar{\beta}_B^2|\|_2^2} + \frac{2}{\sqrt{s}}\|r\|_2\Big\} \end{aligned}$$

with probability greater than $1 - c_1 \exp(-c_2 n)$, where $r = \widetilde{z}_B^1 - \frac{1}{2}(\vec{1} - \frac{1}{\lambda_n}(|\bar{\beta}_B^2| - |\bar{\beta}_B^1|))$. By the decomposition of $\widetilde{z}_B^1$ in equation (49) and applying bounds (50)

$$\|r\|_2 \leq \sqrt{s}\Big\{\mathcal{O}(\sqrt{s/n}) + \frac{1}{\lambda_n\sqrt{s}}\,\mathcal{O}(\sqrt{s/n})\||\bar{\beta}_B^1| - |\bar{\beta}_B^2|\|_2 + \frac{1}{2\sqrt{s}\lambda_n}(1 + \mathcal{O}(\sqrt{s/n}))\,\big[\|f\|_2 + \|\widetilde{f}\|_2\big]\Big\}$$

Since $s/n = o(1)$, in order to establish the upper bound (40b) it suffices to show that $\|f\|_2 + \|\widetilde{f}\|_2 = o(\sqrt{s}\lambda_n)$ w.h.p. Similarly, in the other direction, we have

$$\sqrt{\|\widetilde{z}_B^1\|_2^2 + \|\widetilde{z}_B^2\|_2^2} \geq \sqrt{s}\Big\{\sqrt{\frac{\alpha}{2} + \frac{1}{2s\lambda_n^2}\||\bar{\beta}_B^1| - |\bar{\beta}_B^2|\|_2^2} - \frac{2}{\sqrt{s}}\|r\|_2\Big\}$$

Following the same line of reasoning, in order to prove the lower bound (40a), it suffices to show that $\|f\|_2 + \|\widetilde{f}\|_2 = o(\sqrt{s}\lambda_n)$ w.h.p.

Since $\|f\|_2$ and $\|\widetilde{f}\|_2$ behave similarly, it suffices to show that $\|f\|_2 = o(\lambda_n\sqrt{s})$. From the definition (55a), we see that conditioned on $(X_{B^c}, w, \widetilde{z}_{B^c}^1)$, the random vector f is zero-mean Gaussian, with i.i.d. elements with variance

$$\sigma^2 := \frac{\lambda_n^2}{n}(\widetilde{z}_{B^c}^1)^T(X_{B^c}^T X_{B^c}/n)^{-1}\widetilde{z}_{B^c}^1 + \frac{1}{n}w^T\Pi_{B^{c\perp}}w.$$

Recalling that $\|\widetilde{z}_{B^c}^1\|_2^2 = (1-\alpha)s$, we have

$$\sigma^2 \leq \frac{\lambda_n^2(1-\alpha)s}{n}\,\lambda_{\max}((X_{B^c}^T X_{B^c}/n)^{-1}) + \frac{1}{n}\frac{\|\Pi_{B^{c\perp}}(w)\|_2^2}{n}$$

By random matrix concentration (see the discussion following Lemma 13 in Appendix E), we have $\lambda_{\max}((X_{B^c}^T X_{B^c}/n)^{-1}) \leq 1 + \mathcal{O}(\sqrt{s/n})$ w.h.p., and by χ^2 tail bounds (see Lemma 12 in Appendix E), we have $\frac{\|\Pi_{B^{c\perp}}(w)\|_2^2}{n} = \mathcal{O}(1)$ w.h.p. Consequently, with high probability, we have $\sigma^2 = \mathcal{O}(\frac{\lambda_n^2 s}{n} + \frac{1}{n})$. Since the Gaussian random vector f has length $|B| = \Theta(s)$,

again by concentration for χ^2 random variables, we have (with probability greater than $1 - c_1 \exp(-c_2 s)$), $\|f\|_2^2 = \mathcal{O}(\sigma^2 s)$. Combining the pieces, we conclude that w.h.p.

$$\begin{aligned} \|f\|_2^2 &= \mathcal{O}\big(\lambda_n^2 s \frac{s}{n} + \frac{s}{n}\big) \\ &= \mathcal{O}\big(\lambda_n^2 s \, [\frac{s}{n} + \frac{1}{\lambda_n^2 n}]\big) \; = \; o(\lambda_n^2 s), \end{aligned}$$

where the final equality follows since $s/n = o(1)$ and $1/(\lambda_n^2 n) = o(1)$.

D Convex-analytic characterization of optimal solutions

This section is devoted to the development of various properties of the optimal solution(s) of the block ℓ_1/ℓ_∞-regularized problem (6).

D.1 Basic optimality conditions

By standard conditions for optimality in convex programs [26], the zero-vector must belong to the subdifferential of the objective function in the convex program (6), or equivalently, we must have for each $p = 1, 2, \ldots, r$

$$\frac{1}{n}\big\langle X^i, \, X^i \big\rangle \widehat{\beta}^i - \frac{1}{n}(X^i)^T y^i + \lambda_n \widetilde{z}^i \;\; = \;\; 0, \tag{51}$$

where $\widetilde{Z} \in \mathbb{R}^{p \times r}$ must be an element of the subdifferential $\partial \|\widehat{B}\|_{\infty,1}$. Substituting the relation $y^i = X^i \bar{\beta}^i + w^i$, we obtain

$$\frac{1}{n}\big\langle X^i, \, X^i \big\rangle (\widehat{\beta}^i - \bar{\beta}^i) - \frac{1}{n}(X^i)^T w^i + \lambda_n \widetilde{z}^i = 0. \tag{52}$$

D.2 Proof of Lemma 2

We begin with the proof of part (i): suppose that steps (A) through (C) of the primal-witness construction succeed. By definition, it outputs a primal pair, of the form $(\widetilde{B}_U \; , \; 0)$, along with a candidate dual optimal solution $\{(\widetilde{Z}_U \; , \; \widetilde{Z}_{U^c})\}$. Note that the conditions defining the ℓ_1/ℓ_∞ subdifferential apply in an elementwise manner, to each index $i = 1, \ldots, p$. Since the sub-vector $\widetilde{Z}_U$ was chosen from the subdifferential of the restricted optimal solution, it is dual feasible. Moreover, since the strict dual feasibility condition (30) holds, the matrix $\widetilde{Z}_{U^c}$ constructed in step (C) is dual feasible for the zero-solution in the sub-block U^c. Therefore, we conclude that $(\widetilde{B}_U \; , \; 0)$ is a primal optimal solution for the full block-regularized program (6).

It remains to establish uniqueness of this solution. Define the ball

$$\mathbb{K} \;\; = \;\; \{\widetilde{Z} \in \mathbb{R}^{p \times r} \mid \sum_{i=1}^{r} |\widetilde{z}_k^i| \leq 1 \quad \forall k = 1, \ldots, p\},$$

and observe that we have the variational representation

$$\|B\|_{1,\infty} \;\; = \;\; \sup_{\widetilde{Z} \in \mathbb{K}} \langle \widetilde{Z}, \, B \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. With this notation, the block-regularized program (6) is equivalent to the saddle-point problem

$$\inf_{B \in \mathbb{R}^{p \times r}} \sup_{\widetilde{Z} \in \mathbb{K}} \left\{ \frac{1}{2n} \sum_{i=1}^{r} \|y^i - X^i \beta^i\|_2^2 + \lambda_n \langle \widetilde{Z}, B \rangle \right\}.$$

Since this saddle-point problem is strictly feasible and convex-concave, it has a value. Moreover, given any dual optimal solution—in particular, $\widetilde{Z}$ from the primal-dual construction—any optimal primal solution $\widehat{B}$ must satisfy the saddle point condition

$$\|\widehat{B}\|_{1,\infty} \;=\; \sup_{\widetilde{Z} \in \mathbb{K}} \langle \widetilde{Z}, \widehat{B} \rangle$$

But this condition can only hold if $\forall i \in \{1, 2, \ldots, r\}$, $\beta^i_k = 0$ for any index $k \in \{1, \ldots, p\}$ such that $\sum_{i=1}^r |\widetilde{z}^i_k| < 1$. Therefore, any optimal primal solution must satisfy $\widehat{B}_{U^c} = 0$, so that solving the original program (6) is equivalent to solving the restricted program (29). Lastly, if the matrices $\langle X^i_U, X^i_U \rangle$ are invertible for each $i \in \{1, 2, \ldots, r\}$, then the restricted problem (29) is strictly convex, and so has a unique solution, thereby completing the proof of Lemma 2(i).

We now prove part (ii) of Lemma 2. Suppose that we are given an estimate $\widehat{B}$ of the true parameters $\overline{B}$ by solving the convex program (6) such that $\widehat{B}_{U^c} = 0$.

Since $\widehat{B}$ is an optimal solution to the convex program (6), the the optimality conditions of equation (52), must be satified. We may rewrite those conditions as

$$\begin{aligned} \frac{1}{n} \langle X^i_U, X^i \rangle (\Delta^i) - \frac{1}{n} (X^i_U)^T w^i + \lambda_n \widetilde{z}^i_U = 0 \\ \frac{1}{n} \langle X^i_{U^c}, X^i \rangle (\Delta^i) - \frac{1}{n} (X^i_{U^c})^T w^i + \lambda_n \widetilde{z}^i_{U^c} = 0, \end{aligned}$$

where $\Delta^i = \widehat{\beta}^i - \bar{\beta}^i$. Recalling that $\widehat{B}_{U^c} = \overline{B}_{U^c} = 0$, we obtain

$$\frac{1}{n} \langle X^i_U, X^i_U \rangle (\Delta^i_U) - \frac{1}{n} (X^i_U)^T w^i + \lambda_n \widetilde{z}^i_U = 0, \qquad \text{and} \tag{53a}$$

$$\frac{1}{n} \langle X^i_{U^c}, X^i_U \rangle (\Delta^i_U) - \frac{1}{n} (X^i_{U^c})^T w^i + \lambda_n \widetilde{z}^i_{U^c} = 0. \tag{53b}$$

Again, by standard conditions for optimality in convex programs [2, 8], the first of these two equations is exactly the condition that must be satisfied by an optimal solution of the restricted program (29). However, we have already shown that the candidate solution $\widehat{B}_U$ satisfies this condition, so that it must also be an optimal solution of the convex program (29). Additionally, the value of $\widetilde{Z}_U$ that satisfies equation (53a) for each $i \in \{1, 2, \ldots, r\}$ is an element of $\partial \|\widehat{B}\|_{\infty,1}$. We have thus shown that steps (B) and (C) of the primal-witness construction succeed. It remains to establish uniqueness in part (A). However, we note that $\langle X^i_U, X^i_U \rangle$ is invertible for each i. Hence, for any solution $\widehat{B}$ such that $\widehat{B}_{U^c} = 0$,

$$\Delta^i_U = (\frac{1}{n} \langle X^i_U, X^i_U \rangle)^{-1} \left[\frac{1}{n} (X^i_U)^T w^i - \lambda_n \widetilde{z}^i_U \right]$$

is well-defined and unique, noting that $\Delta^i_{U^c} = 0$. Thus, we have established the equality (32) and that $\widehat{B}_U$ is unique. Therefore, $\widehat{B}$ gives solutions to steps (A) and (B) when solving the restricted convex program over the set U.

Finally, we derive the form of the dual solution $\widetilde{z}^i_{U^c}$, as a function of $\langle X^i_U, X^i_U\rangle$, $\widetilde{Z}_U$, and $\widehat{B} - \overline{B}$. Recall that $\langle X^i_U, X^i_U\rangle$ is invertible, $\widetilde{Z}_U$ is an element of the subdifferential of $\partial\|\widetilde{B}_U\|_{\ell_1/\ell_\infty}$, and $\widehat{B}_{U^c} = \overline{B}_{U^c} = 0$. From equation (32), we have

$$\widetilde{z}^i_{U^c} \;=\; \frac{1}{\lambda_n n}\langle X^i_{U^c},\, (I - \Pi_{X^i_U})w^i\rangle + \frac{1}{n}\langle X^i_{U^c},\, X^i_U(\frac{1}{n}\langle X^i_U,\, X^i_U\rangle)^{-1}\widetilde{z}^i_U\rangle \quad \text{for } i = 1, \ldots, r. \quad (54)$$

The claimed form of the dual solution follows by substituting equation (32) into equation (53b).

D.3 Subgradients on the support

In this section, we focus on the specific form of the dual variables $\widetilde{z}^i_U$. Our approach is to construct a candidate set of dual variables, and then show that they are valid. We begin by defining the sets $B = S(\bar{\beta}^i) \cap S(\bar{\beta}^j)$, corresponding to the intersection of the supports, and the set $B^c = U \setminus B$ corresponding to elements in one (but not both) of the supports. For $i = 1, 2$, we let $S^i \in \mathbb{R}^{\alpha s \times \alpha s}$ is a diagonal matrix whose diagonal entries correspond to $\operatorname{sign}(\bar{\beta}^i_B)$. In addition, we define the vectors $f^i \in \mathbb{R}^{\alpha s}$ and matrices $M^i \in \mathbb{R}^{\alpha s \times \alpha s}$ via

$$f^i \;:=\; S^i\left[\frac{1}{n}\langle X^i_B,\, X^i_{B^c}(\frac{1}{n}\langle X^i_{B^c},\, X^i_{B^c}\rangle)^{-1}\rangle\lambda_n\widetilde{z}^i_{B^c} - \frac{1}{n}\langle X^i_B,\, I - \Pi_{X^i_{B^c}}\rangle w^i\right] \quad (55a)$$

$$M^i \;:=\; \frac{1}{n}S^i\langle X^i_B,\, (I - \Pi_{(X^i_{B^c})})X^i_B\rangle S^i. \quad (55b)$$

Given these definitions, we have the following lemma:

Lemma 7. *Assume that $r = 2$, and that $|\widehat{\beta}^1_B| = |\widehat{\beta}^2_B|$. If $\widehat{B}_{U^c} = \overline{B}_{U^c} = 0$, then the dual variable $\widetilde{z}^1$ satisfies the relation*

$$S^1\widetilde{z}^1_B = \frac{1}{\lambda_n}\big\{M^2\big[M^1 + M^2\big]^{-1}f^1 - M^1\big[M^2 + M^1\big]^{-1}f^2\big\} + M^1\big[M^1 + M^2\big]^{-1}\vec{1} - \frac{1}{\lambda_n}[(M^1)^{-1} + (M^2)^{-1}]^{-1}\overline{B}_{\text{diff}} \quad (56)$$

and $\widetilde{z}^2_{B^c} = \mathbb{S}_{\pm}(\bar{\beta}^2_{B^c})$, with analogous results holding for $\widetilde{z}^2$.

Given these forms for $S^1\widetilde{z}^1_B$ and $S^2\widetilde{z}^2_B$, it remains to show that the relation $S^1\widetilde{z}^1_B + S^2\widetilde{z}^2_B = 1$ holds under the conditions of Theorem 3(a). Intuitively, this condition should hold since under the conditions of theorem 3(a), the matrix M^i is approximately the identity, and the vector f^i is approaching 0. Finally, we expect that $\overline{B}_{\text{diff}} := |\bar{\beta}^2_B| - |\bar{\beta}^1_B|$ is very small, hence the final term is also very small. Therefore, on the set B, both $S^1\widetilde{z}^1_B$ and $S^2\widetilde{z}^2_B$ are approximately equal to $\frac{1}{2}$. We formalize this rough intuition in the following lemma:

Lemma 8. *Under the assumptions of Theorem 3(a) each of the following conditions hold for sufficiently large n, s, and p with probability greater than $1 - c_1\exp(-c_2 n)$:*

$$\|\frac{1}{\lambda_n}[(M^1)^{-1} + (M^2)^{-1}]^{-1}(\overline{B}_{\text{diff}})\|_\infty \;\leq\; \epsilon \quad (57a)$$

$$\|\frac{1}{\lambda_n}\big\{M^2\big[M^1 + M^2\big]^{-1}f^1 - M^1\big[M^2 + M^1\big]^{-1}f^2\big\}\|_\infty \;\leq\; \epsilon \quad (57b)$$

$$\|M^1\big[M^1 + M^2\big]^{-1}\vec{1} - \frac{1}{2}\|_\infty \;\leq\; \frac{1}{2} - 3\epsilon. \quad (57c)$$

Given Lemmas 7 and 8, we can conclude that the definition for the dual variables on the support is valid. The remaining subsections in this appendix are dedicated to verifying the above results: in particular, we prove Lemma 7 in Appendix D.4 and Lemma 8 in Appendix D.5.

D.4 Proof of Lemma 7

We now proceed to establish the validity of the closed form expressions for $\widetilde{z}_U^1$ and $\widetilde{z}_U^2$. From equation (53a) we have that

$$\Delta_{B^c}^1 = -(\frac{1}{n}\langle X_{B^c}^i,\, X_{B^c}^i\rangle)^{-1}\left[\frac{1}{n}\langle X_{B^c}^1,\, X_B^1\rangle\Delta_B^1 + \lambda_n\widetilde{z}_{B^c}^1\right] + (\frac{1}{n}\langle X_{B^c}^i,\, X_{B^c}^i\rangle)^{-1}(X_{B^c}^1)^T w^1$$

substituting back into (53a)

$$\frac{1}{n}\langle X_B^1,\, X_B^1\rangle\Delta_B^1 + \frac{1}{n}\langle X_B^1,\, X_{B^c}^1\rangle\Delta_{B^c}^1 - \frac{1}{n}(X_B^1)^T w^1 + \lambda_n\widetilde{z}_B^1 \,=\, 0,$$

so that we obtain

$$\begin{aligned} M^1\Delta_B^1 &= f^1 - \lambda_n\widetilde{z}_B^1 \qquad \text{and similarly,} && \text{(58a)}\\ M^2\Delta_B^2 &= f^2 - \lambda_n\widetilde{z}_B^2 && \text{(58b)}\end{aligned}$$

Recall that by assumption that $S^1\widehat{\beta}_B^1 = |\widehat{\beta}_B^1| = |\widehat{\beta}_B^2| = S^2\widehat{\beta}_B^2$, and $S\widetilde{z}_B^1 + \widetilde{S}\widetilde{z}_B^2 = 1$.

Subtracting $M^1S^1\bar{\beta}_B^1$ and $M^2S^2\bar{\beta}_B^2$ from equations (58a) and (58b)

$$\begin{aligned} M^1S^1(\Delta_B^1 - \bar{\beta}_B^1) &= f^1 - \lambda_n S^1\widetilde{z}_B^1 - M^1S^1\bar{\beta}_B^1 && \text{(59a)}\\ M^2S^2(\Delta_B^2 - \bar{\beta}_B^2) &= f^2 - \lambda_n S^2\widetilde{z}_B^1 - M^2S^2\bar{\beta}_B^2 && \text{(59b)}\end{aligned}$$

Applying the fact that $S^1(\Delta_B^1 - \bar{\beta}_B^1) = S^2(\Delta_B^2 - \bar{\beta}_B^2)$.

$$(M^1 + M^2)S^1(\Delta_B^1 - \bar{\beta}_B^1) = (f^1 + f^2) - \lambda_n\vec{1} - M^1S^1\bar{\beta}_B^1 - M^2S^2\bar{\beta}_B^2,$$

where $\vec{1} \in \mathbb{R}^{\alpha s}$. Then solving for $S^1(\Delta_B^1 - \bar{\beta}_B^1)$ letting $S^1\bar{\beta}_B^1 - S^2\bar{\beta}_B^2 = \overline{B}_{\text{diff}}$ and substituting back into equation (59a)

$$\begin{aligned}\lambda_n S^1\widetilde{z}_B^1 = M^1\big[M^1 + M^2\big]^{-1}\lambda_n\vec{1} - [(M^1)^{-1} + (M^2)^{-1}]^{-1}(\overline{B}_{\text{diff}})\\ + M^2\big[M^1 + M^2\big]^{-1}f^1 - M^1\big[M^1 + M^2\big]^{-1}f^2. \quad \text{(60)}\end{aligned}$$

D.5 Proof of Lemma 8

The first term $\frac{1}{\lambda_n}[(M^1)^{-1} + (M^2)^{-1}]^{-1}\,\overline{B}_{\text{diff}}$ can be decomposed as

$$\frac{1}{\lambda_n}[(M^1)^{-1} + (M^2)^{-1}]^{-1}\overline{B}_{\text{diff}} \;=\; \underbrace{\frac{1}{\lambda_n}([(M^1)^{-1} + (M^2)^{-1}]^{-1} - I/2)\overline{B}_{\text{diff}}}_{T_1} + \underbrace{\frac{\overline{B}_{\text{diff}}}{2\lambda_n}}_{T_2}$$

Under the assumptions of Theorem 3(a), we have $\big|\frac{\overline{B}_{\text{diff}}}{2\,\lambda_n}\big| \to 0$, hence, for s large enough, $T_2 \le \epsilon/4$.

In order to bound T_1, we note that with probability greater than $1 - c_1 \exp(-c_2 n)$, the spectral norm of $([(M^1)^{-1} + (M^2)^{-1}]^{-1} - I/2)$ is $\mathcal{O}(\sqrt{s/n})$ (see the bound (50b) from Appendix C). Consequently, we may decompose $([(M^1)^{-1} + (M^2)^{-1}]^{-1} - I/2)$ as QDQ^T where Q and D are independent and Q is distributed uniformly over all orthogonal matrices, and $|||D|||_2 = \mathcal{O}(\sqrt{s/n})$. Using this decomposition, the following lemma, proved in Appendix D.6, allows us to obtain the necessary control on the quantity $\|T_1\|_\infty$:

Lemma 9. *Let $Q \in \mathbb{R}^{s \times s}$ be a matrix chosen uniformly at random from the space of all orthogonal matrices. Consider a second random matrix A, independent of Q. If $s/n = o(1)$, then for any fixed vector $x \in \mathbb{R}^s$ and fixed $\epsilon > 0$, we have:*

(a) If $|||A|||_2 \leq \sqrt{\frac{s}{n}}$, then

$$\mathbb{P}[\|Q^T A Q x\|_\infty \geq \frac{\epsilon}{2}] \quad \leq \quad c_1 \exp\big(-c_2 \epsilon^2 \frac{n}{s\|x\|_\infty^2} + \log(s)\big).$$

(b) If $|||A|||_2 \leq \frac{s}{n}$, then

$$\mathbb{P}[\|Q^T A Q x\|_\infty \geq \frac{\epsilon}{2}] \quad \leq \quad c_1 \exp\big(-c_2 \epsilon^2 \frac{n^2}{s^2\|x\|_\infty^2} + \log(s)\big).$$

With reference to the problem of bounding $\|T_1\|_\infty$, we may apply part (a) of this lemma with $A = D$ and $x = \frac{\bar{B}_{\text{diff}}}{2\lambda_n}$ to conclude that $\|T_1\|_\infty \leq \epsilon/2$ with high probability, thereby establishing the bound (57a).

We now turn the proving the bound (57b). We begin by decomposing the terms involved in this equation as

$$\begin{aligned}
\frac{1}{\lambda_n} M^2\big[M^1 + M^2\big]^{-1} f^1 &= \frac{1}{\lambda_n}\Big[M^2\big[M^1 + M^2\big]^{-1} - \frac{I}{2}\Big] f^1 + \frac{f^1}{2\lambda_n} \\
\frac{1}{\lambda_n} M^1\big[M^1 + M^2\big]^{-1} f^2 &= \frac{1}{\lambda_n}\Big[M^1\big[M^1 + M^2\big]^{-1} - \frac{I}{2}\Big] f^2 + \frac{f^2}{2\lambda_n}
\end{aligned}$$

Recalling the form of $f^i \in \mathbb{R}^{\alpha s}$, conditioned on $X^i_{B^c}$ and w^i, we have

$$f^i/(2\lambda_n) \sim N\left(0, \frac{1}{4}\langle \widetilde{z}^i_{B^c}, \frac{1}{n}(\frac{1}{n}\langle X^i_{B^c}, X^i_{B^c}\rangle)^{-1}\widetilde{z}^i_{B^c}\rangle I_{\alpha s} + \|w^i\|_2^2/(n^2\lambda_n^2) I_{\alpha s}\right).$$

However, by Lemmas 12 and 13 (see Appendix E), as well as the fact that $\|\widetilde{z}^i_s\|_2^2 = (1-\alpha)s$, for n and s large enough, the variance term is bounded by

$$\frac{1}{4}\langle \widetilde{z}^i_{B^c}, \frac{1}{n}(\frac{1}{n}\langle X^i_{B^c}, X^i_{B^c}\rangle)^{-1}\widetilde{z}^i_{B^c}\rangle + \|w^i\|_2^2/(n^2\lambda_n^2) \leq \frac{1}{4}(1-\alpha)\frac{s}{n}(1+\delta) + \frac{1}{2}\frac{1}{n\lambda_n^2} \tag{61}$$

with probability greater than $1 - c_1\exp(-c_2 n)$. Hence, by standard Gaussian tail bounds, the inequalities $\|f^1/(2\lambda_n)\|_\infty < \epsilon/4$ and $\|f^2/(2\lambda_n)\|_\infty < \epsilon/4$ both hold with probability greater than $1 - c_1\exp(-\delta' \log(p-2s))$.

Now to bound the first term in the decomposition we begin by diagonalizing $M^2 = Q^T D Q$. Note that Q is independent of X^1 and D and by symmetry $X^1_B \stackrel{d}{=} Q X^1_B$. Following some algebra, we find that

$$\frac{1}{\lambda_n}\Big[M^2\big[M^1 + M^2\big]^{-1} - \frac{I}{2}\Big] f^1 = \frac{1}{\lambda_n} Q^T \Big[D\big[QM^1Q^T + D\big]^{-1} - \frac{I}{2}\Big] Q f^1$$

The random vector f^1 is independent of Q and Qf^1 is independent of Q by symmetry. Hence, the vector $v := \frac{1}{2}[2D(D+QM^1Q^T)^{-1} - I]Q\frac{1}{\lambda_n}f^1$ is independent of Q. For a given constant c_3, let us define the event

$$\mathcal{S} := \big\{\|v\|_2^2 \leq c_3^2 \, \frac{s^2}{n}\big[\frac{s}{n} + \frac{1}{\lambda_n^2 n}\big]\big\}.$$

We can then write

$$\mathbb{P}[\|Q^T v\|_\infty \geq \epsilon] \;\; \leq \;\; \mathbb{P}\big[\|Q^T v\|_\infty \geq \epsilon \;\big|\; \mathcal{S}\big] + \mathbb{P}[\mathcal{S}^c].$$

Note that we may consider the event that $\|\!|D|\!\|_2 = \mathcal{O}(1)$ and $[2D(D+QM^1Q^T)^{-1} - I] = \mathcal{O}(\sqrt{s/n})$. We claim that each of these events happens with high probability. Note that the former event occurs with high probability by Lemma 13. The latter event holds with high probability since,

$$[2D(D+QM^1Q^T)^{-1} - I] = [2D((D+QM^1Q^T)^{-1} - I/2) + D - I].$$

and, both $\|\!|D-I|\!\|_2 = \mathcal{O}(\sqrt{s/n})$ and $((D+QM^1Q^T)^{-1} - I/2) = \mathcal{O}(\sqrt{s/n})$ by equation (68a). Thus, the sum of the two random matrices is also $\mathcal{O}(\sqrt{s/n})$.

Recall the bound on the variance of each component of f^1 from equation (61) and note that each component is independent. Applying the concentration results from Lemma 12 for χ-squared random variables yields that $\|f^1\|_2^2 \leq \frac{1}{4}(1+\delta)\frac{s^2}{n} + \frac{1}{2}\frac{s}{n\lambda_n^2}$ with high probability. Hence, under the above conditions

$$\begin{aligned}\|\frac{1}{2}[2D(D+QM^1Q^T)^{-1} - I]Q\frac{1}{\lambda_n}f^1\|_2^2 \;\; &\leq \;\; \|\!|\frac{1}{2}[2D(D+QM^1Q^T)^{-1} - I]|\!\|_2^2 \|Q\frac{1}{\lambda_n}f^1\|_2^2 \\ &\leq \;\; c_3^2 \, \frac{s^2}{n}\big[\frac{s}{n} + \frac{1}{\lambda_n^2 n}\big],\end{aligned}$$

with high probabilty, which implies that $\mathcal{S}$ holds with high probability as well. Therefore, it immediately follows then that $\mathbb{P}[\mathcal{S}^c] \leq c_1 \exp(-c_2 s)$.

It remains to control the first term. We do so using the following lemma, which is proved in Appendix D.7:

Lemma 10. *Let $Q \in \mathbb{R}^{m\times m}$ be a matrix chosen uniformly at random from the space of orthogonal matrices. Let $v \in \mathbb{R}^m$ be a random vector independent of Q, such that $\|v\|_2 \leq v^*$ with probability one. Then we have*

$$\mathbb{P}\big[\|Q^T v\|_\infty \geq 2\, v^* \sqrt{\frac{\log m}{m}}\big] \;\; = \;\; o(1).$$

We now apply this lemma to the random vector v with $m = s$, and $v^* = c_3 \, \frac{s}{\sqrt{n}} \, \sqrt{\frac{s}{n} + \frac{1}{\lambda_n^2 n}}$. Note that

$$2v^* \sqrt{\frac{\log s}{s}} \;\; = \;\; 2c_3 \sqrt{\frac{\log s}{n}} \, \sqrt{\frac{s}{n} + \frac{1}{\lambda_n^2 n}} \;\; = \;\; o(1),$$

from which the second claim (57b) in Lemma 8 follows.

Finally, we turn to proving the third claim (57c) in Lemma 8. Following some algebra, we obtain

$$\|M^1\big[M^1+M^2\big]^{-1}\vec{1}-\frac{1}{2}\|_\infty \;=\; \frac{1}{4}(M^1-M^2)\vec{1}+\frac{1}{2}(M^1-M^2)(I/2-(M^1+M^2)^{-1})\vec{1}. \quad (62)$$

We diagonalize the matrix $M^1 = Q^T D Q$, where D is diagonal. Since the random matrix M^1 has a spherically symmetric distribution, the matrix Q has a uniform distribution over the space of orthogonal matrices and is independent of D. Using this decomposition, we can rewrite the second term in equation (62) as

$$\frac{1}{2}Q^T(D-QM^2Q^T)(\frac{I}{2}-(D+QM^2Q^T)^{-1})Q\vec{1} \;=\; Q^TRQ\vec{1} \quad (63)$$

where $R := \frac{1}{4}(D-QM^2Q^T)(I-2(D+QM^2Q^T)^{-1})$. We note that R is independent of Q, because D and M^2 are independent of Q. This independence follows from the spherical symmetry of M^2 and the fact that $M^2 \stackrel{d}{=} QM^2Q^T$.

Defining the event $\mathcal{T} := \big\{\|\!|R|\!\|_2 \leq 4s/n\big\}$, we claim that

$$\mathbb{P}[\mathcal{T}^c] \;\leq\; c_1\exp(-c_2 n) \;\to\; 0. \quad (64)$$

In order to establish this claim, we note that sub-multiplicativity and triangle inequality imply that

$$\begin{aligned}\|\!|R|\!\|_2 &\leq \frac{1}{4}\|\!|D-QM^2Q^T|\!\|_2\,\|\!|(D+QM^2Q^T)/2-I|\!\|_2\,\|\!|2(D+QM^2Q^T)^{-1}|\!\|_2 \\ &\leq 2(\|\!|D-I|\!\|_2+\|\!|I-QM^2Q^T|\!\|_2)\|\!|(D+QM^2Q^T)/2-I|\!\|_2,\end{aligned}$$

since $\|\!|2(Q^TDQ+QM^2Q^T)^{-1}|\!\|_2 \leq 2$ with probability greater than $1-c_1\exp(-c_2n)$, from the discussion following Lemma 13. Similarly, from this same result, we have $\mathcal{O}(\|\!|D-I|\!\|_2) = \mathcal{O}(\|\!|I-QM^2Q^T|\!\|_2) = \mathcal{O}(\|\!|(D+QM^2Q^T)/2-I|\!\|_2) \;\leq\; 2\sqrt{\frac{s}{n}}$, so that the claim (64) follows.

Using the decomposition (63) and the tail bound (64), we have

$$\begin{aligned}\mathbb{P}[\|Q^TRQ1\|_\infty \geq \epsilon] &= \mathbb{P}[\|Q^TRQ1\|_\infty \geq \epsilon \;\mid\; \mathcal{T}] + \mathbb{P}[\mathcal{T}^c] \\ &\leq \mathcal{O}\big(\frac{1}{s}\big) + \mathcal{O}(\exp(-c(\epsilon)n)),\end{aligned}$$

where Lemma 9 (proved in Appendix D.6) provides control on the first term in the inequality.

D.6 Proof of Lemma 9

We provide the proof for part (a) of the Lemma and note that part (b) is analogous.
By union bound, we have

$$\mathbb{P}[\|Q^TAQx\|_\infty \geq \epsilon] \;\leq\; s\max_{i=1,\ldots,s}\mathbb{P}[|e_i^TQ^TAQx| \geq \epsilon].$$

We will derive a bound on the probability $\mathbb{P}[|e_1^TQ^TAQx| \geq \epsilon]$ that holds for all $e_i, i = 1,\ldots,s$. We write $e_1^TQ^TAQx = x_1v_1^TAv_1 + v_1^TAv_2$, where v_1 denotes the first column of Q, and $v_2 = \sum_{k=2}^s x_kQ_k$ denotes the weighted sum of the remaining $(k-1)$ columns of Q. Since Q is orthogonal, the vector v_1 has unit norm $\|v_1\|_2 = 1$, the vector v_2 is orthogonal to v_1, and moreover $\|v_2\|_2^2 \leq \|x\|_\infty^2 s - 1$. Owing to the bound on the spectral norm of A, we have

$$|x_1v_1^TAv_1| \;\leq \|x\|_\infty\sqrt{\tfrac{s}{n}}$$

which is less than $\epsilon/2$ for (s, n) sufficiently large, since $s/n = o(1)$.

We now turn to the second term. Note that conditioned on v_2, the vector v_1 is uniformly distributed over an $(s-1)$-dimensional unit sphere, contained within the subspace orthogonal to v_2. Still conditioning on v_2, consider the function $f(v_1) = v_1^T A v_2$. For any pair of vectors v_1, v_1' on the unit sphere, we have

$$\begin{aligned} |f(v_1) - f(v_1')|^2 &= |(v_1 - v_1')^T A v_2|^2 \\ &\leq |\!|\!| A |\!|\!|_2^2 \, \|x\|_\infty^2 \, (s-1) \|v_1 - v_1'\|_2^2 \\ &= |\!|\!| A |\!|\!|_2^2 \, \|x\|_\infty^2 \, (s-1) \, \big[2\,(1 - \cos(d(v_1, v_1')))\big], \end{aligned}$$

where $d = \arccos(v_1^T v_1')$ is the geodesic distance. Using the inequality $\cos(d) \geq 1 - d^2/2$, valid for $d \in [0, \pi]$, and the assumption $|\!|\!| A |\!|\!|_2 \leq \sqrt{s/n}$, and taking square roots, we obtain

$$|f(v_1) - f(v_1')| \leq \sqrt{\frac{s}{n}} \, \|x\|_\infty \, \sqrt{(s-1)} \, d(v_1, v_v'),$$

so that f is a Lipschitz constant on the unit sphere (with dimension $s-1$) with constant $L = \|x\|_\infty \sqrt{\frac{s}{n}\,(s-1)}$. Consequently, by Levy's theorem [12], for any $\epsilon > 0$, we have

$$\mathbb{P}[|f(v_1)| \geq \epsilon] \leq 2\exp(-(s-2)\,\frac{n}{\|x\|_\infty^2 \, s(s-1)}\,\epsilon^2) \leq 2\,\exp\big(-c_1\,\frac{n}{\|x\|_\infty^2\,s}\epsilon^2\big).$$

As a final side remark, we note that under the scaling of Theorem 3(b), we have $\frac{n}{s}\epsilon^2 - \log(s) \to \infty$ as $n \to \infty$, so that the probability in question vanishes.

D.7 Proof of Lemma 10

By union bound and symmetry of the distribution Q, for any $t > 0$, we have

$$\begin{aligned} \mathbb{P}\big[\|Q^T v\|_\infty \geq t\big] &\leq m\,\mathbb{P}\big[|e_1^T Q^T v| \geq t\big] \\ &= m\,\mathbb{P}\big[|q_1^T v| \geq t\big], \end{aligned}$$

where q_1 is the first column of Q. Note that q_1 is a random vector distributed uniformly over the unit sphere S^{m-1} in m dimensions. Viewing the vector $v \in \mathbb{R}^m$ as fixed, consider the function $f(q) = q^T v$ defined over S^{m-1}. As in Lemma 9, some calculation shows the Lipschitz constant of g over S^{m-1} is at most $L = \|v\|_2$. Applying Levy's theorem [12], we conclude that for any $\epsilon > 0$,

$$m\,\mathbb{P}[|f(q_1)| \geq t] \leq 2\exp\big(-(m-1)\,\frac{t^2}{2\|v\|_2^2} + \log m\big).$$

Since $\|v\|_2 \leq v^*$ by assumption, it suffices to set $t = 2v^*\sqrt{\frac{\log m}{m}}$.

E Some large deviation bounds

In this appendix, we state some known large deviation bounds for the Gausssian variates, χ^2-variates, as well as the eigenvalues of random matrices. The following Gaussian tail bound is standard:

Lemma 11. *For a Gaussian variable $Z \sim N(0, \sigma^2)$, for all $t > 0$,*

$$\mathbb{P}[|Z| \geq t] \quad \leq \quad 2\exp\big(-\frac{t^2}{2\sigma^2}\big). \tag{65}$$

The following tail bounds on chi-squared variates are also useful:

Lemma 12. *Let X be a χ-squared random variable with d degrees of freedom. Then for all $t > 0$, we have*

$$\mathbb{P}[\frac{X}{d} \geq (1+t)^2] \quad \leq \quad \exp(-\frac{dt^2}{2}), \qquad \textit{and} \tag{66a}$$

$$\mathbb{P}[\frac{X}{d} \leq (1-2t)] \quad \leq \quad \exp(-dt^2). \tag{66b}$$

Proof. These tail bounds are immediate consequences of results due to Laurent and Massart [11], who prove that for all $x > 0$, we have

$$\mathbb{P}[X \geq x + (\sqrt{x} + \sqrt{d})^2] \quad \leq \quad \exp(-x), \quad \text{and} \tag{67a}$$

$$\mathbb{P}(X - d \leq -2\sqrt{dx}) \quad \leq \quad \exp(-x). \tag{67b}$$

Letting $x = dt^2/2$ in equation (67a), we have

$$\exp(-\frac{dt^2}{2}) \quad \geq \quad \mathbb{P}[\frac{X}{d} \geq \sqrt{2}t + 1 + t^2] \;\geq\; \mathbb{P}[\frac{X}{d} \geq (1+t)^2],$$

thereby establishing (66a). With the same choice of x, equation (67b) implies the bound (66b) immediately. □

Finally, the following type of large deviations bound on the eigenvalues of Gaussian random matrices is standard (e.g., [6]):

Lemma 13. *Let $X \in \mathbb{R}^{n \times s}$ be a random matrix from the standard Gaussian ensemble (i.e., $X_{ij} \sim N(0,1)$, i.i.d). Then with probability greater than $1 - c_1 \exp(-c_2 n)$, for any $\delta > 0$, its eigenspectrum satisfies the bounds*

$$(1-\delta)\,\Big[1 - \sqrt{\frac{s}{n}}\Big]^2 \;\leq\; \Lambda_{\min}\big(\frac{X^T X}{n}\big) \leq \Lambda_{\max}\big(\frac{X^T X}{n}\big) \;\leq\; (1+\delta)\,\Big[1 + \sqrt{\frac{s}{n}}\Big]^2.$$

Note that this lemma implies similar bounds for eigenvalues of the inverse:

$$\frac{1}{(1+\delta)\,\big[1+\sqrt{\frac{s}{n}}\big]^2} \;\leq\; \Lambda_{\min}\big((\frac{X^T X}{n})^{-1}\big) \leq \Lambda_{\max}\big((\frac{X^T X}{n})^{-1}\big) \;\leq\; \frac{1}{(1-\delta)\,\big[1-\sqrt{\frac{s}{n}}\big]^2}.$$

From the above two sets of inequalities, we conclude for $s/n \leq 1$, we have with probability greater than $1 - c_1 \exp(-c_2 n)$

$$\|\frac{1}{n} X^T X - I\|_2 \quad \leq \quad 4\sqrt{\frac{s}{n}}, \quad \text{and} \tag{68a}$$

$$\|(\frac{1}{n} X^T X)^{-1} - I\|_2 \quad \leq \quad 4\sqrt{\frac{s}{n}}. \tag{68b}$$

For random matrices where each row is distributed $N(0, \Sigma)$ and $\Lambda_{\min}(\Sigma) > C_{\min}$ and $\Lambda_{\max}(\Sigma) \leq C_{\max}$, we have

$$\|\frac{1}{n} X^T X - \Sigma\|_2 \quad \leq \quad \lambda_{\max}(\Sigma) 4\sqrt{\frac{s}{n}}, \quad \text{and} \tag{69a}$$

$$\|(\frac{1}{n} X^T X)^{-1} - \Sigma^{-1}\|_2 \quad \leq \quad \frac{4}{\lambda_{\min}(\Sigma)}\,\sqrt{\frac{s}{n}}. \tag{69b}$$

References

[1] F. Bach. Consistency of the group Lasso and multiple kernel learning. Technical report, INRIA - Département d'Informatique, Ecole Normale Supérieure, 2008.

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

[3] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2009. To appear.

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

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

[6] 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.

[7] D. L. Donoho and J. M. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. Technical report, Stanford University, 2006. Submitted to Journal of the AMS.

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

[9] J. Huang and T. Zhang. The benefit of group sparsity. Technical Report arXiv:0901.2962, Rutgers University, January 2009.

[10] M. Jordan, editor. *Learning in graphical models*. MIT Press, Cambridge, MA, 1999.

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

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

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

[14] H. Liu, J. Lafferty, and L. Wasserman. Nonparametric regression and classification with joint sparsity constraints. In *Neural Info. Proc. Systems (NIPS) 22*, Vancouver, Canada, December 2008.

[15] H. Liu and J. Zhang. On $\ell_1 - \ell_q$ regularized regression. Technical Report arXiv:0802.1517v1, Carnegie Mellon University, 2008.

[16] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[17] S. G. Mallat. *A wavelet tour of signal processing*. Academic Press, New York, 1998.

[18] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008.

[19] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(2):1436–1462, 2006.

[20] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 2008. To appear.

[21] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[22] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection for grouped classification. Technical report, Statistics Department, UC Berkeley, 2007.

[23] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical report, Department of Statistics, UC Berkeley, August 2008.

[24] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. Technical Report arXiv:0711.4555v2, Carnegie Mellon University, 2008.

[25] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional graph selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 2008.

[26] G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[27] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. In P. Müller and B. Vidakovic, editors, *Bayesian Inference in Wavelet Based Models*, chapter 18, pages 291–308. Springer-Verlag, New York, June 1999. Lecture Notes in Statistics, vol. 141.

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

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

[30] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, April 2006. Special issue on "Sparse approximations in signal and image processing".

[31] B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005.

[32] 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.

[33] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using using ℓ_1-constrained quadratic programs. *IEEE Transactions on Information Theory*, In press. Appeared as Tech. Report 709, Department of Statistics, UC Berkeley. May 2006.

[34] 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.

[35] Kim Y., Kim J., and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16(2), 2006.

[36] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical report, Statistics Department, UC Berkeley, 2007. To appear in Annals of Statistics.

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