A Dirty Model for Multiple Sparse Regression

Ali Jalali, Pradeep Ravikumar, and Sujay Sanghavi, *Member*

***Abstract*—The task of sparse linear regression consists of finding an unknown sparse vector from linear measurements. Solving this task even under "high-dimensional" settings, where the number of samples is fewer than the number of variables, is now known to be possible via methods such as the LASSO. We consider the *multiple* sparse linear regression problem, where the task consists of recovering several related sparse vectors at once. A simple approach to this task would involve solving independent sparse linear regression problems, but a natural question is whether one can reduce the overall number of samples required by leveraging *partial sharing* of the support sets, or non-zero patterns, of the signal vectors. A line of recent research has studied the use of ℓ_1/ℓ_q norm block-regularizations with $q > 1$ for such problems. However, depending on the level of sharing, these could actually perform *worse* in sample complexity when compared to solving each problem independently.**

We present a new "adaptive" method for multiple sparse linear regression that can leverage support and parameter overlap when it exists, but not pay a penalty when it does not. We show how to achieve this using a very simple idea: decompose the parameters into two components and *regularize these differently.* We show, theoretically and empirically, that our method strictly and noticeably outperforms both ℓ_1 or ℓ_1/ℓ_q methods, over the entire range of possible overlaps (except at boundary cases, where we match the best method), even under high-dimensional scaling.

***Index Terms*—Multi-task Learning, High-dimensional Statistics, Multiple Regression.**

I. Introduction: Motivation and Setup

High-dimensional scaling. In fields across science and engineering, we are increasingly faced with problems where the number of variables or features p is larger than the number of observations n. For any hope of statistically consistent estimation under such high-dimensional scaling, it becomes vital to leverage any potential structure in the problem such as sparsity (e.g. in compressed sensing [1] and LASSO [2]), low-rank structure [3, 4], or sparse graphical model structure [5]. It is in such high-dimensional contexts in particular that *multi-task learning* [6] could be most useful. Here, multiple tasks share some common structure such as sparsity, and estimating these tasks jointly by leveraging this common structure could be more statistically efficient.

Block-sparse Multiple Regression. A common multiple task learning setting, and which is the focus of this paper, is that of multiple regression, where we have $r > 1$ response variables, and a common set of p features or covariates. The r tasks could share certain aspects of their underlying distributions, such as common variance, but the setting we focus on in this paper is where the response variables have *shared* sparse structure: the index set of relevant features for each task is individually sparse; but there is also a large overlap of these relevant features across the different regression problems. Such "shared sparsity" arises in a variety of contexts; most applications of sparse signal recovery in contexts ranging from graphical model learning, kernel learning, and function estimation have natural extensions to the shared-sparse setting [5, 7, 8, 9].

It is conceptually useful to collate the multiple regression parameters into a matrix, with columns corresponding to tasks, and the rows corresponding to features. Having shared sparse structure then corresponds to this matrix being largely "block-sparse," where due to shared sparsity structure most rows are either exactly zero, or with a few non-zero entries, there are only a few rows with a large number of non-zero entries. A line of recent research in this setting has focused on ℓ_1/ℓ_q norm regularizations, for $q > 1$, which encourage the parameter matrix to be strictly row-sparse, starting from the work by Yuan and Lin [10] who termed the case with $q = 2$ as "Group Lasso". Examples of other recent results include those using the ℓ_1/ℓ_∞ norm [11, 12, 13], as well as the ℓ_1/ℓ_2 norm [10, 14, 15].

Our Model. Such block-regularization is "heavy-handed" in two ways. They strictly encourage block or row-sparsity, so that any row is either exactly zero or has all its entries being non-zero. This assumes that all relevant features are exactly shared, and hence suffers under settings, arguably more realistic, where each task depends on features specific to itself in addition to the ones that are common. The second concern with such block-sparse regularizers is that the ℓ_1/ℓ_q norms for $q > 2$ can be shown to encourage the entries in the non-sparse rows taking nearly identical *values*. Thus we are far away from the original goal of multitask learning: not only do the set of relevant features have to be exactly the same, but their values have to as well. Indeed recent research into such regularized methods [13, 15] caution against the use of block-regularization in regimes where the supports and values of the parameters for each task can vary widely. Since the true parameter values are unknown, that would be a worrisome caveat.

We thus ask the question: can we learn multiple regression models by leveraging whatever overlap of features there exist, and without requiring the parameter values to be near identical? Indeed this is an instance of a more general question on whether we can estimate statistical models where the data may not fall cleanly into any one structural bracket (sparse, block-sparse and so on). With the explosion of complex and *dirty* high-dimensional data in modern settings, it is vital to investigate estimation of corresponding *dirty* models, which might require new approaches to biased high-dimensional estimation. In this paper we take a first step, focusing on such dirty models for a specific problem: simultaneously sparse

The authors are with the Departments of Electrical and Computer Engineering (Jalali and Sanghavi) and Computer Science (Ravikumar), The University of Texas at Austin, Austin, TX 78712 USA email: (alij@mail.utexas.edu; pradeepr@cs.utexas.edu; sanghavi@mail.utexas.edu).
Preliminary results were published in NIPS 2010.

multiple regression.

Our approach uses a simple idea: while any one structure might not capture the data, a superposition of structural classes might. Our method thus searches for a parameter matrix that can be *decomposed* into a row-sparse matrix (corresponding to the overlapping or shared features) and an elementwise sparse matrix (corresponding to the non-shared features). As we show both theoretically and empirically, with this simple fix we are able to leverage any extent of shared features, while allowing disparities in support and values of the parameters, so that we are *always* better than both the Lasso or block-sparse regularizers, at times remarkably so.

The rest of the paper is organized as follows: In Section 2, we present basic definitions and the setup of the problem. We then discuss the main results of the paper in Section 3. Experimental results and simulations are demonstrated in Section 4.

Notation: For any matrix M, we denote its j^{th} row as m_j, and its k-th column as $m^{(k)}$. The set of all non-zero rows (i.e. all rows with at least one non-zero element) is denoted by RowSupp(M) and its support by Supp(M). Also, for any matrix M, let $\|M\|_{1,1} := \sum_{j,k} |m_j^{(k)}|$, i.e. the sums of absolute values of the elements, and $\|M\|_{1,\infty} := \sum_j \|m_j\|_\infty$ where, $\|m_j\|_\infty := \max_k |m_j^{(k)}|$.

II. Problem Set-up and Our Method

Multiple linear regression: We consider the following standard multiple linear regression model:

$$y^{(k)} = X^{(k)}\bar{\theta}^{(k)} + w^{(k)}, \quad k = 1, \ldots, r, \tag{1}$$

where, $y^{(k)} \in \mathbb{R}^n$ is the response for the k-th task, regressed on the design matrix $X^{(k)} \in \mathbb{R}^{n\times p}$ (possibly different across tasks), while $w^{(k)} \in \mathbb{R}^n$ is the noise vector. We assume each $w^{(k)}$ is drawn independently from $\mathcal{N}(0, \sigma^2)$. The total number of tasks or target variables is r, the number of features is p, while the number of samples we have for each task is n. For notational convenience, we collate these quantities into matrices $Y \in \mathbb{R}^{n\times r}$ for the responses, $\bar{\Theta} \in \mathbb{R}^{p\times r}$ for the regression parameters and $W \in \mathbb{R}^{n\times r}$ for the noise.

Our Model: In this paper we are interested in the setting where the true parameter $\bar{\Theta}$ from data $\{y^{(k)}, X^{(k)}\}$ has partially shared-sparsity, as detailed in the introduction. In particular, for any fixed integer d, suppose we denote rows of $\bar{\Theta}$ with greater than or equal to d non-zero entries, corresponding to features shared by several tasks, as "shared rows"; and those rows with less than d non-zero entries, corresponding to those features which are relevant for some tasks but not all, as "non-shared rows." The latter includes rows with all zero entries, corresponding to those features that are not relevant to any task. The true parameter can then be split as $\bar{\Theta} = \bar{B} + \bar{S}$, where, $\bar{B}$ contains the shared rows and $\bar{S}$ contains non-shared rows, with respect to the integer d. We are interested in estimators $(\check{B}, \check{S})$ that separate the shared and non-shared rows, and enjoy the following statistical guarantees.

Support recovery: We say an estimator $(\widehat{B}, \widehat{S})$ successfully recovers the true support if Supp$(\widehat{B}+\widehat{S})$ = Supp$(\bar{\Theta})$. We note that this is stronger than merely recovering the row-support of $\bar{\Theta}$, which is union of its supports for the different tasks. Support recovery is often also referred to as variable selection.

Error bounds: We are also interested in providing bounds on the element-wise ℓ_∞ norm error of the estimator $\widehat{\Theta} = \check{B} + \check{S}$ defined as

$$\|\widehat{\Theta} - \bar{\Theta}\|_\infty = \max_{j=1,\ldots,p} \max_{k=1,\ldots,r} \left|\widehat{\Theta}_j^{(k)} - \bar{\Theta}_j^{(k)}\right|.$$

Our Method: We model the unknown parameter Θ as a superposition of a block-sparse parameter matrix B (corresponding to the features shared across many tasks) and a sparse parameter matrix S (corresponding to the features shared across few tasks). We thus have two parameter matrices, B and S, and we regularize these two matrices differently, encouraging block-structured row-sparsity in B, and elementwise sparsity in S. This can be contrasted with the "clean" standard models that use a single parameter matrix, and either use just block-sparse regularizations [13, 15] or just elementwise sparsity regularizations [2, 16]. Interestingly, as we will see in the main results, by explicitly allowing to have both block-sparse and elementwise sparse components (see Algorithm II), we are able to *outperform both* classes of these "clean models", for *all* regimes of the parameter matrix $\bar{\Theta}$. Notice that our algorithm has a post processing step that combines the rows of $\widehat{S}$ and $\widehat{B}$ on the row support of $\widehat{B}$. This post processing does not change the sum of the two, i.e., $\widehat{\Theta} = \widehat{B}+\widehat{S} = \check{B}+\check{S}$.

III. Main Results and Their Consequences

We now provide precise statements of our main results. A number of recent results have shown that the Lasso [2, 16] and ℓ_1/ℓ_∞ block-regularization [13] methods succeed in model selection, i.e., recovering signed supports with controlled error bounds under high-dimensional scaling regimes. Our first two theorems extend these results to our M-estimator. In Theorem 1, we consider the case of deterministic design matrices $X^{(k)}$, and provide sufficient conditions guaranteeing signed support recovery, and elementwise ℓ_∞ norm error bounds. In Theorem 2, we specialize this theorem to the case where the rows of the design matrices are random from a general zero mean Gaussian distribution: this allows us to provide scaling on the number of observations required in order to guarantee signed support recovery and bounded elementwise ℓ_∞ norm error.

Our third result is the most interesting in that it explicitly quantifies the performance gains of our method vis-a-vis Lasso and the ℓ_1/ℓ_∞ block-regularization method. Since this entailed deriving precise constants underlying earlier theorems, and a correspondingly more delicate analysis, we follow Negahban and Wainwright [13] and focus on the case where there are two-tasks (i.e. $r = 2$), and where we have standard Gaussian design matrices as in Theorem 2. Further, while each of two tasks depends on s features, only a fraction α of these are common. It is then interesting to see how the behaviors of the different regularization methods vary with the extent of overlap α.

Algorithm 1 Dirty Multitask Learner

```
Pick $\lambda_s$ and $\lambda_b$ such that $\lfloor \frac{\lambda_s}{\lambda_b} \rfloor = d$ and $\frac{\lambda_s}{\lambda_b}$ is not integer.

Solve the following convex optimization problem:
    $$(\widehat{S}, \widehat{B}) \in \arg\min_{S,B} \frac{1}{2n} \sum_{k=1}^{r} \left\| y^{(k)} - X^{(k)}\left(s^{(k)} + b^{(k)}\right) \right\|_2^2 + \lambda_s \|S\|_{1,1} + \lambda_b \|B\|_{1,\infty}. \qquad (2)$$

For all $j \in \mathrm{RowSupp}(\widehat{B})$, let $\check{B}_j = \widehat{B}_j + \widehat{S}_j$.

Let $\check{S} = \widehat{B} + \widehat{S} - \check{B}$.

Output $(\check{B}, \check{S})$.
```

Comparisons: Negahban and Wainwright [13] show that there is actually a "phase transition" in the scaling of the probability of successful signed support-recovery with the number of observations. Consider the specific rescaling of the sample-size $\theta_{Lasso}(n,p,\alpha) := \frac{n}{s\log(p-s)}$. Then Wainwright [16] show that when the rescaled number of samples scales as $\theta_{Lasso} > 2+\delta$ for any $\delta > 0$, then Lasso succeeds in recovering the signed support of all columns with probability converging to one. But when the sample size scales as $\theta_{Lasso} < 2-\delta$ for any $\delta > 0$, Lasso *fails* with probability converging to one. For the ℓ_1/ℓ_∞-regularized multiple linear regression, define a similar rescaled sample size $\theta_{1,\infty}(n,p,\alpha) := \frac{n}{s\log(p-(2-\alpha)s)}$. Then as Negahban and Wainwright [13] show there is again a transition in probability of success from near zero to near one, as the rescaled sample size of $\theta_{1,\infty}$ is either less or greater than $(4-3\alpha)$. These phase transitions provide a natural means for comparing competing M-estimators. Thus, if $\theta_{Lasso}(n,p,\alpha) < \theta_{1,\infty}(n,p,\alpha)$, which can be shown to be equivalent to $\alpha < 2/3$, the phase transition for Lasso occurs at a smaller sample size than the ℓ_1/ℓ_∞ regularized method, so that the Lasso can be seen to be the more efficient method. Note that $\alpha < 2/3$ corresponds to the "less sharing" setting, so that it is not surprising that Lasso would perform better. Conversely, when $\theta_{Lasso}(n,p,\alpha) > \theta_{1,\infty}(n,p,\alpha)$, equivalent to $\alpha > 2/3$ and which corresponds to the "more sharing" setting, the ℓ_1/ℓ_∞ regularized method performs better in that its phase transition occurs at a smaller sample size.

As we show in our third theorem, the phase transition for our method occurs when the rescaled sample size $\theta_{1,\infty}$ is equal to $(2-\alpha)$, which is *strictly* before either the Lasso or the ℓ_1/ℓ_∞ regularized method except for the boundary cases: $\alpha = 0$, i.e. the case of no sharing, where we *match* Lasso, and for $\alpha = 1$, i.e. full sharing, where we *match* ℓ_1/ℓ_∞. Everywhere else, we *strictly outperform both* methods. Figure III shows the empirical performance of each of the three methods; as can be seen, they agree very well with the theoretical analysis. (Further details in the experiments Section IV).

A. *Sufficient Conditions for Deterministic Designs*

We first consider the case where the design matrices $X^{(k)}$ for $k = 1, \cdots, r$ are deterministic, and start by specifying the assumptions we impose on the model. We note that similar sufficient conditions for the deterministic $X^{(k)}$'s case were imposed in papers analyzing Lasso [16] and block-regularization methods [13, 15].

A0 Column Normalization: $\|X_j^{(k)}\|_2 \leq \sqrt{2n}$ for all $j = 1, \ldots, p$ and $k = 1, \ldots, r$.

A1 Incoherence Conditions:

$$\gamma_b := 1 - \max_{j \in \mathcal{U}^c} \sum_{k=1}^{r} \left\| \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\rangle \right\|_1 > 0,$$

where, $\mathcal{U}_k$ denotes the support of the k-th column of $\bar{\Theta}$, and $\mathcal{U} = \bigcup_k \mathcal{U}_k$ denotes the union of the supports of all tasks, and $\langle A, B \rangle = A^T B$. We also require

$$\gamma_s := 1 - \max_{1 \leq k \leq r} \max_{j \in \mathcal{U}_k^c} \left\| \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\rangle \right\|_1 > 0.$$

A2 Minimum Curvature Condition:

$$C_{min} := \min_{1 \leq k \leq r} \lambda_{min} \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right) > 0,$$

where, $\lambda_{\min}(\cdot)$ is the minimum eigenvalue of the matrix. Also, define $D_{max} := \max_{1 \leq k \leq r} \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \right\|_{\infty,1}$. As a consequence of **A2**, we have that $D_{\max}$ is finite.

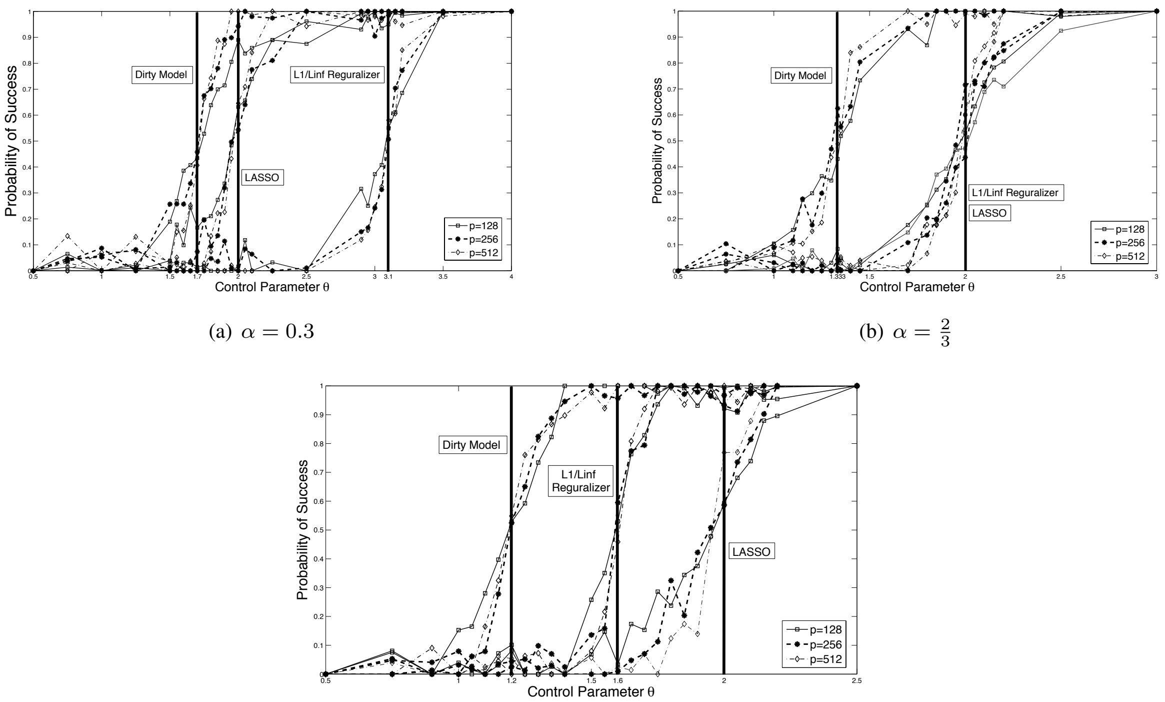
A3 Regularizers: We require the regularization parameters satisfy

A3-1 $\lambda_s > \frac{2(2-\gamma_s)\sigma\sqrt{\log(pr)}}{\gamma_s\sqrt{n}}$.

A3-2 $\lambda_b > \frac{2(2-\gamma_b)\sigma\sqrt{\log(pr)}}{\gamma_b\sqrt{n}}$.

A3-3 $1 \leq \frac{\lambda_b}{\lambda_s} \leq r$ and $\frac{\lambda_b}{\lambda_s}$ is not an integer (See Lemmas 2 and 3 for intuition on these conditions).

Theorem 1. *Consider the multiple linear regression model in* (1)*, and which satisfies assumptions* **A0-A3**. *Suppose we obtain estimate* $\widehat{\Theta} = \check{B} + \check{S}$ *from Algorithm AlgDirtyModel. Then, with probability at least* $1 - c_1\exp(-c_2 n)$*, we are guaranteed that the convex program (2) has a unique optimum, and that*



(a) $\alpha = 0.3$

(b) $\alpha = \frac{2}{3}$

(c) $\alpha = 0.8$

Fig. 1. Probability of success in recovering the true signed support using dirty model, Lasso and ℓ_1/ℓ_∞ regularizer. For a 2-task problem, the probability of success for different values of feature-overlap fraction α is plotted. As we can see in the regimes that Lasso is better than, as good as and worse than ℓ_1/ℓ_∞ regularizer ((a), (b) and (c) respectively), the dirty model outperforms both of the methods, i.e., it requires less number of observations for successful recovery of the true signed support compared to Lasso and ℓ_1/ℓ_∞ regularizer. Here $s = \lfloor \frac{p}{10} \rfloor$ always.

(a) *the estimate has no false inclusions, and has bounded ℓ_∞ norm error:*

$$\begin{aligned} &Supp(\check{S}) \subseteq Supp(\bar{S}), \quad and \\ &RowSupp(\check{B}) \subseteq RowSupp(\bar{B}), \quad and \\ &\|\widehat{\Theta} - \bar{\Theta}\|_{\infty,\infty} \leq \underbrace{\sqrt{\frac{4\sigma^2 \log(pr)}{n\, C_{min}}} + \lambda_s D_{max}}_{b_{\min}} . \end{aligned}$$

(b) *The estimate has no false exclusions, i.e., $sign(Supp(\check{S})) = sign\left(Supp(\bar{S})\right)$ and $RowSupp(\check{B}) = RowSupp(\bar{B})$ with the property that if $\bar{B}_j^{(k)} \neq 0$, then $sign(\check{B}_j^{(k)}) = sign\left(\bar{B}_j^{(k)}\right)$, provided that* $\min_{(j,k)\in Supp(\bar{\Theta})} \left|\bar{\theta}_j^{(k)}\right| > b_{\min}$.

The positive constants c_1, c_2 depend only on $\gamma_s, \gamma_b, \lambda_s, \lambda_b$ and σ, but are otherwise independent of n, p, r, the problem dimensions of interest.

Remark: Condition (a) guarantees that the estimate will have no *false inclusions*; i.e. all included features will be relevant. If in addition, we require that it have no *false exclusions* and that recover the support exactly, we need to impose the assumption in (b) that the non-zero elements are large enough to be detectable above the noise.

B. General Gaussian Designs

In many applications, the design matrices consist of samples from a Gaussian ensemble (e.g. in Gaussian graphical model structure learning). Suppose that for each task $k = 1, \dots, r$ the design matrix $X^{(k)} \in \mathbb{R}^{n\times p}$ is such that each row $X_i^{(k)} \in \mathbb{R}^p$ is a zero-mean Gaussian random vector with covariance matrix $\Sigma^{(k)} \in \mathbb{R}^{p\times p}$, and is independent of every other row. Let $\Sigma^{(k)}_{\mathcal{V},\mathcal{U}} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{U}|}$ be the sub-matrix of $\Sigma^{(k)}$ with corresponding rows to $\mathcal{V}$ and columns to $\mathcal{U}$. We require these covariance matrices to satisfy the following conditions:

C1 Incoherence Conditions:

$$\gamma_b := 1 - \max_{j\in\mathcal{U}^c} \sum_{k=1}^{r} \left\| \Sigma^{(k)}_{j,\mathcal{U}_k} \left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1} \right\|_1 > 0$$

and

$$\gamma_s := 1 - \max_{1\leq k\leq r} \max_{j\in\mathcal{U}_k^c} \left\| \Sigma^{(k)}_{j,\mathcal{U}_k} \left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1} \right\|_1 > 0 .$$

C2 Minimum Curvature Condition:

$$C_{min} := \min_{1\leq k\leq r} \lambda_{min}\left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right) > 0$$

and let $D_{max} := \left\| \left(\Sigma^{(k)}_{\mathcal{U}_k,\mathcal{U}_k}\right)^{-1} \right\|_{\infty,1}$.

These conditions are analogous to the sufficient conditions **A1-A2** in the previous theorem. Those earlier conditions were imposed on the design matrices themselves, whereas conditions **C1-C2** are imposed on the covariance matrix of the (randomly generated) rows of the design matrix.

C3 Regularizers: Defining $s := \max_k |\mathcal{U}_k|$, we require the regularization parameters satisfy

- **C3-1** $\lambda_s \geq \frac{\left(4\sigma^2 C_{min}\log(pr)\right)^{1/2}}{\gamma_s\sqrt{nC_{min}}-\sqrt{2s\log(pr)}}$.
- **C3-2** $\lambda_b \geq \frac{\left(4\sigma^2 C_{min} r(r\log(2)+\log(p))\right)^{1/2}}{\gamma_b\sqrt{nC_{min}}-\sqrt{2sr(r\log(2)+\log(p))}}$.
- **C3-3** $1 \leq \frac{\lambda_b}{\lambda_s} \leq r$ and $\frac{\lambda_b}{\lambda_s}$ is not an integer.

Theorem 2. *Suppose assumptions* **C1-C3** *hold, and that the number of samples scale as*

$$n > \max\left(\frac{2s\log(pr)}{C_{min}\gamma_s^2}, \frac{2sr\big(r\log(2)+\log(p)\big)}{C_{min}\gamma_b^2}\right).$$

Suppose we obtain estimate $\widehat{\Theta} = \check{B} + \check{S}$ *from our algorithm. Then, with probability at least*

$$1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right) - c_3\exp(-c_4\log(rs)) \to 1,$$

for some positive numbers $c_1 - c_4$*, we are guaranteed that the convex program (2) has a unique optimum and*

(a) The estimate has no false inclusions, and has bounded ℓ_∞ *norm error so that*

$$\textit{Supp}(\check{S}) \subseteq \textit{Supp}(\bar{S}),$$
$$\textit{RowSupp}(\check{B}) \subseteq \textit{RowSupp}(\bar{B}),$$
$$\|\widehat{\Theta} - \bar{\Theta}\|_{\infty,\infty} \leq \underbrace{\sqrt{\frac{50\sigma^2\log(rs)}{nC_{min}}} + \lambda_s\Big(\frac{4s}{C_{min}\sqrt{n}} + D_{max}\Big)}_{g_{\min}}.$$

(b) The estimate has no false exclusions, i.e., $\textit{sign}(\textit{Supp}(\check{S})) = \textit{sign}\left(\textit{Supp}(\bar{S})\right)$ *and* $\textit{RowSupp}(\check{B}) = \textit{RowSupp}(\bar{B})$ *with the property that if* $\bar{B}_j^{(k)} \neq 0$*, then* $\textit{sign}(\check{B}_j^{(k)}) = \textit{sign}\left(\bar{B}_j^{(k)}\right)$*, provided that* $\min_{(j,k)\in \textit{Supp}(\bar{\Theta})}\left|\bar{\theta}_j^{(k)}\right| > g_{\min}$.

C. Quantifying the gain for 2-Task Gaussian Designs

This is one of the most important results of this paper. Here, we perform a more delicate and finer analysis to establish precise quantitative gains of our method. We focus on the special case where $r = 2$ and the design matrix has rows generated from the standard Gaussian distribution $\mathcal{N}(0, I_{n\times n})$. As we will see both analytically and experimentally, our method strictly outperforms both Lasso and ℓ_1/ℓ_∞-block-regularization over for all cases, except at the extreme end-points of no support sharing (where it matches that of Lasso) and full support sharing (where it matches that of ℓ_1/ℓ_∞). We now present our analytical results; the empirical comparisons are presented next in Section IV. The results will be in terms of a particular rescaling of the sample size n as

$$\theta(n,p,s,\alpha) := \frac{n}{(2-\alpha)s\log\left(p-(2-\alpha)s\right)}.$$

We also require that the regularizers satisfy

F1 $\lambda_s > \dfrac{\Big(4\sigma^2(1-\sqrt{s/n})(\log(2)+\log(p-(2-\alpha)s))\Big)^{1/2}}{\sqrt{n}-\sqrt{s}-((2-\alpha)\ s\ (\log(2)+\log(p-(2-\alpha)s)))^{1/2}}$.

F2 $\lambda_b > \dfrac{\Big(8\sigma^2(1-\sqrt{s/n})(2\log(2)+\log(p-(2-\alpha)s))\Big)^{1/2}}{\sqrt{n}-\sqrt{s}-((2-\alpha)\ s\ (2\log(2)+\log(p-(2-\alpha)s)))^{1/2}}$.

F3 $\frac{\lambda_b}{\lambda_s} = \sqrt{2}$.

Notice that **F1** and **F2** only impose lower-bounds on λ_b and λ_s. Hence, while **F3** fixes the ratio of the two to be $\sqrt{2}$, there are always infinitely many pairs (λ_b, λ_s) that satisfy these conditions.

We also note that **F3** is not essential for the analysis, but it provides the tightest bounds. In the proofs, we actually analyze the case with any general value for the ratio $\kappa = \frac{\lambda_b}{\lambda_s}$, and provide the phase transition threshold for the number of samples in terms of this ratio; please see Theorem 4 on page 15. While the sample complexity threshold depends in a complicated way on the ratio, as we show there, it is minimized when $\kappa = \sqrt{2}$. However, in practice, when the assumptions in the theorem need not hold, or when we are interested in prediction error in contrast to support recovery as in Theorem 3, it might be useful to search for different ratios. The next theorem provides a sharp transition for the two task case with these assumptions.

Theorem 3. *Consider a* 2*-task regression problem* (n,p,s,α)*, where the design matrix has rows generated from the standard Gaussian distribution* $\mathcal{N}(0, I_{n\times n})$*. Suppose*

$$\max_{j\in B^*}\left|\,|\Theta_j^{*(1)}| - |\Theta_j^{*(2)}|\,\right| \leq c\lambda_s,$$

where, B^* *is the submatrix of* Θ^* *with rows where both entries are non-zero and* c *is a constant specified in Lemma 10. Then the estimate* $\widehat{\Theta} = \check{B} + \check{S}$ *of the problem (2) satisfies the following:*

(Success). *Suppose the regularization coefficients satisfy* **F1 − F3**. *Further, assume that the number of samples scales as* $\theta(n,p,s,\alpha) > 1$*. Then, with probability at least* $1 - c_1\exp(-c_2 n)$ *for some positive numbers* c_1 *and* c_2*, we are guaranteed that sign* $\Big(\textit{Supp}(\widehat{\Theta})\Big) = \textit{sign}\left(\textit{Supp}(\bar{\Theta})\right)$ *and* ℓ_∞ *error bound conditions (a-b) in Theorem 2 are hold.*

(Failure). *If* $\theta(n,p,s,\alpha) < 1$ *there is no solution* $(\check{B},\check{S})$ *for any choices of* λ_s *and* λ_b *such that sign* $\left(\textit{Supp}(\check{S})\right) = \textit{sign}\left(\textit{Supp}(\bar{S})\right)$ *and* $\textit{RowSupp}(\check{B}) = \textit{RowSupp}(\bar{B})$ *with the property that if* $\bar{B}_j^{(k)} \neq 0$ *then sign* $\left(\check{B}_j^{(k)}\right) = \textit{sign}\left(\bar{B}_j^{(k)}\right)$.

Remark: The assumption on the gap $\left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| \leq c\lambda_s$ requires that most values of Θ^* to be balanced on both tasks on the shared support. But as we show in a more general theorem (Theorem 4) in Section VI-C, even in the case where the gap is large, the dependence of the sample scaling on the gap is quite weak.

IV. Simulation Results

In this section, we provide some simulation results. First, using our synthetic data set, we investigate the consequences of Theorem 3 when we have $r = 2$ tasks to learn. As we see, the empirical result verifies our theoretical guarantees. Next, we apply our method to a real dataset: a hand-written digit classification dataset with $r = 10$ tasks (equal to the number of digits $0 - 9$). For this dataset, we show that our method outperforms both LASSO and ℓ_1/ℓ_∞ practically. For each method, the parameters are chosen via cross-validation; see supplemental material for more details.

A. Synthetic Data Simulation

Consider a two-task regression model, so that $r = 2$, as discussed in Theorem 3. As detailed in this section, we compare the performance our dirty M-estimator, ℓ_1/ℓ_∞ regularization based method, and LASSO in recovering the true signed support.

Data Generation: We ran the algorithms for multiple instances of the parameters (n, p, s, α). We used three different number of features $p \in \{128, 256, 512\}$, and five different values of the overlap ratio $\alpha \in \{0.05, 0.3, \frac{2}{3}, 0.8, 0.9\}$. For different values of p, we set $s = \lfloor 0.1p \rfloor$, and for different values of (s, p, α), we let $n = c\, s \log(p - (2 - \alpha)s)$ for different values of c. We generated the parameter matrix in two steps. We first generated a random sign matrix $\widetilde{\Theta}^* \in \mathbb{R}^{p\times 2}$ (each entry is either 0, 1 or -1) with column support size s and row support size $(2 - \alpha)s$ as required by Theorem 3. We then multiplied each row by a real random number with magnitude greater than the minimum required for sign support recovery by Theorem 3. We generate two sets of the matrix tuple $(X^{(1)}, X^{(2)}, W)$; we used one of them for training and the other for cross validation, subscripted Tr and Ts, respectively. Each entry of the noise matrices $W_{\text{Tr}}, W_{\text{Ts}} \in \mathbb{R}^{n\times 2}$ is drawn independently according to $\mathcal{N}(0, \sigma^2)$ where $\sigma = 0.1$. Each row of a design matrix $X_{\text{Tr}}^{(k)}, X_{\text{Ts}}^{(k)} \in \mathbb{R}^{n\times p}$ is sampled, independent of any other rows, from $\mathcal{N}(0, \mathbf{I}_{2\times 2})$ for all $k = 1, 2$. Having $X^{(k)}$, $Th\bar{e}ta$ and W in hand, we can calculate $Y_{\text{Tr}}, Y_{\text{Ts}} \in \mathbb{R}^{n\times 2}$ using the model $y^{(k)} = X^{(k)}\theta^{(k)} + w^{(k)}$ for all $k = 1, 2$ for both train and test set of variables.

Coordinate Descent Algorithm: Given the generated data $X_{\text{Tr}}^{(k)}$ for $k = 1, 2$ and Y_{Tr} in the previous section, we solve the M-estimation problem in (2) to obtain matrices $\hat{B}$ and $\hat{S}$. To numerically solve the problem, we use the coordinate descent algorithm outlined in Appendix D. The co-ordinate descent algorithm takes as input the tuple $(X_{\text{Tr}}^{(1)}, X_{\text{Tr}}^{(2)}, Y_{\text{Tr}}, \lambda_s, \lambda_b, \epsilon, \underline{B}, \underline{S})$ and outputs a matrix pair $(\hat{B}, \hat{S})$ as the solution of the M-estimation problem (2). The inputs $(\underline{B}, \underline{S})$ are initial guesses and can be set to zero. However, when we search over the regularizer coefficients, we can use the M-estimate for the previous set of coefficients (λ_b, λ_s) as a good initial guess for the corresponding M-estimation problem with the next set of coefficients $(\lambda_b + \xi, \lambda_s + \zeta)$. The parameter ϵ is the stopping criterion threshold of the co-ordinate descent algorithm, which is set to iterate until the relative update change of the objective function is less than ϵ.

Choosing penalty regularizer coefficients: Our optimality conditions entail that $1 > \frac{\lambda_s}{\lambda_b} > \frac{1}{2}$. Thus, given one of the regularization coefficients, the search-range for the other is bounded and known. We set $\lambda_b = c\sqrt{\frac{rlog(p)}{n}}$ and search for the constant c over a logarithmic partition of the interval $[0.01, 100]$. For any pair (λ_b, λ_s), we first compute our M-estimate $(\hat{B}, \hat{S})$ from the coordinate descent algorithm run over the training data; and then compute the unregularized parameter estimate $\widehat{\Theta}^\lambda$, that minimizes the un-regularized squared error loss function over the training data, but with support restricted to that of $\hat{B} + \hat{S}$. We then pick the pair (λ_b, λ_s) for which the corresponding parameter $\widehat{\Theta}^\lambda$ has the least unregularized loss over the test data $\{Y_{\text{Ts}}, X_{\text{Ts}}^{(k)}\}_{k=1,2}$. Finally we let $\hat{\Theta} = \hat{B} + \hat{S}$ for the M-estimate $(\hat{B}, \hat{S})$ corresponding to the optimal (λ_b, λ_s).

Performance Analysis: For any instance of the problem (n, p, s, α), we generate 100 batches of samples from the corresponding problem instance. We then solve these problem instances using our "dirty" M-estimator, the ℓ_1/ℓ_∞ regularized method, and LASSO, where we set the penalty regularizer coefficients independently for each one of these programs via cross validation. For any method, if the recovered matrix $\hat{\Theta}$ has the same sign support as the true $\bar{\Theta}$, then we count it as a success, or as a failure otherwise (note that even if one element has different sign, we count it as failure).

As Theorem 3 predicts and Fig III in Section III shows, the number of observations rescaled as $\frac{n}{s\log(p-(2-\alpha)s)}$ is the key control parameter driving the probability of success of our method, since the curves for different problem sizes p stack on the top of each other. It can also be seen that the number of observations required by our method for true signed support recovery is always less than both the LASSO and the ℓ_1/ℓ_∞ regularized method. Fig 1(a) shows the probability of success for the case $\alpha = 0.3$, where LASSO is better than the ℓ_1/ℓ_∞ regularized method, while our dirty M-estimator outperforms both methods. Fig 1(b) shows the case with $\alpha = \frac{2}{3}$, where the LASSO and the ℓ_1/ℓ_∞ regularized method performs the same, but our method require almost 33% less observations for the same probability of success. As α grows toward 1, e.g. $\alpha = 0.8$ as shown in Fig 1(c), ℓ_1/ℓ_∞ regularization performs better than the LASSO. Our M-estimator performs better than both methods in this case as well.

Scaling Verification: To verify that the phase transition threshold changes linearly with α as predicted by Theorem 3, we plot the phase transition threshold versus α. For five different values of $\alpha \in \{0.05, 0.3, \frac{2}{3}, 0.8, 0.95\}$ and three different values of $p \in \{128, 256, 512\}$, we first compute the phase-transition sample-size n as the point where the probability of success in recovery of signed support exceeds 50% (which we find by interpolating the closest two points). In Fig 2, we then plot the rescaled phase-transition sample-size $\theta = \frac{n}{s\log(p-(2-\alpha)s)}$ vs α, for three methods;

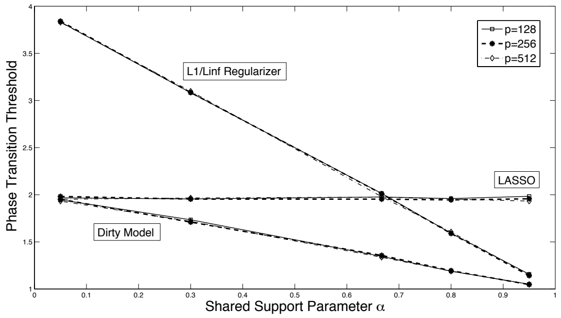


Fig. 2. Verification of the result of the Theorem 3 on the behavior of phase transition threshold by changing the parameter α in a 2-task (n, p, s, α) problem for our method, LASSO and ℓ_1/ℓ_∞ regularizer. The y-axis is $\frac{n}{s\log(p-(2-\alpha)s)}$, where n is the number of samples at which threshold was observed. Here $s = \lfloor \frac{p}{10} \rfloor$. Our method shows a gain in sample complexity over the entire range of sharing α. The pre-constant in Theorem 3 is also validated.

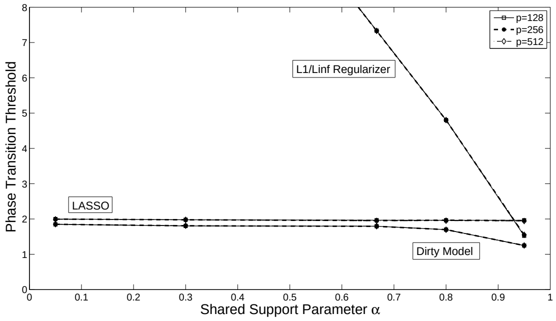


Fig. 3. Phase transition threshold by changing the parameter α in a 10-task (n, p, s, α) problem for our method, LASSO and ℓ_1/ℓ_∞ regularizer. Here, we assume that each of 10 tasks has a support of size s and α portion of that support is shared across all 10 tasks and the rest is distributed randomly. The y-axis is $\frac{n}{s\log(p)}$, where n is the number of samples at which threshold was observed. Here $s = \lfloor \frac{p}{10} \rfloor$. Our method shows a gain in sample complexity over the entire range of sharing α.

our M-estimator, LASSO and the ℓ_1/ℓ_∞ regularized method. As the figure shows, the phase transition threshold for our method is always lower than the phase transition for the other two methods.

10-task Experiment: Although we do not have a theoretical analysis for sharp phase transitions in the problem beyond $r = 2$, we now present some empirical observations of the behavior of our method for $r > 2$. We run the same experiment as the earlier two task case for this 10-task case, where we assume each task has a support of size s and α portion of this support is shared across all tasks. The non-shared portion of the task supports is distributed randomly for each task. Fig 3 shows the phase transition for different methods. It can be seen that our algorithm outperforms other methods for all regimes of α.

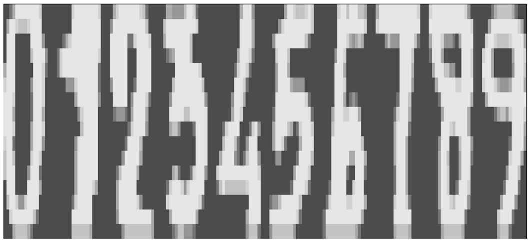


Fig. 4. An instance of images of the ten digits extracted from the dataset

B. Handwritten Digits Dataset

We use a handwritten digit dataset to illustrate the performance of our method. According to the description of the dataset, this dataset consists of features of handwritten numerals (0-9) extracted from a collection of Dutch utility maps [17]. This dataset has been used in a number of papers [18, 19] as a reliable dataset for handwriting recognition algorithms.

Structure of the Dataset: The dataset has 200 handwritten instances of the digits 0-9, so that there are 2000 digit instances in total. Each instance of each digit is scanned to an image of the size 30×48 pixels. Instead of the raw image, the dataset provides six different classes of features drawn from the full resolution image of each digit. A total of 649 features are provided for each instance of each digit. The information about each class of features is provided in Table I. The combined set of handwritten images of record number 100 are shown in Fig 4 (ten images are concatenated together with space between any two).

Fitting the dataset to our model: We have 649 features for each of 200 instance of each digit. We need to learn $K = 10$ different tasks corresponding to ten different digits. To make the associated numbers of features comparable, we shrink the dynamic range of each feature to the interval -1 and 1.

Out of the 200 samples provided for each digit, we select $n \leq 200$ samples for the training dataset. We then follow the typical binary classification setup for this problem. For any $0 \leq k \leq 9$, let $X^{(k)} = X \in \mathbb{R}^{10n \times 649}$ be the matrix whose first n rows correspond to the features of the digit 0, the second n rows correspond to the features of the digit 1 and so on. Let the vector $y^{(k)} \in \{0,1\}^{10n}$ be the vector such that $y^{(k)}_j = 1$ if and only if the j^{th} row of the feature matrix X corresponds to the digit k.

We then solve the M-estimation problem (2) to get a block-sparse matrix $\hat{B} \in \mathbb{R}^{649\times 10}$, and a sparse matrix $\hat{S} \in \mathbb{R}^{649\times 10}$. Given any feature vector $\mathbf{x} \in \mathbb{R}^{649}$ extracted from the image of a handwritten digit, we then classify the image as digit $k^* = \arg\max_{k\in\{0,\ldots,9\}}[\mathbf{x}^T\left(\hat{B} + \hat{S}\right)]_k$. We set the regularization parameters λ_b and λ_s as before. We first solve (2) for each pair of regularization parameters, and then minimize the unregularized loss function on the support

	Feature	Size	Type	Dynamic Range
1	Pixel Shape (15 × 16)	240	Integer	0-6
2	2D Fourier Transform Coefficients	74	Real	0-1
3	Karhunen-Loeve Transform Coeficients	64	Real	-17:17
4	Profile Correlation	216	Integer	0-1400
5	Zernike Moments	46	Real	0-800
6	Morphological Features	3	Integer	0-6
		1	Real	100-200
		1	Real	1-3
		1	Real	1500-18000

TABLE I
SIX DIFFERENT CLASSES OF FEATURES PROVIDED IN THE DATASET. THE DYNAMIC RANGES ARE APPROXIMATE NOT EXACT. THE DYNAMIC RANGE OF DIFFERENT MORPHOLOGICAL FEATURES ARE COMPLETELY DIFFERENT. FOR THOSE 6 MORPHOLOGICAL FEATURES, WE PROVIDE THEIR DIFFERENT DYNAMIC RANGES SEPARATELY.

recovered by that choice of parameters. We then evaluate the prediction error using this reoptimized solution over the test set. Since we have 10 tasks, we search for $\frac{\lambda_s}{\lambda_b} \in \left[\frac{1}{10}, 1\right]$ and let $\lambda_b = c\sqrt{\frac{2log(649)}{n}} \approx \frac{5c}{\sqrt{n}}$, where we search over the constant c in the interval $[0.01, 10]$

Performance Analysis: Table II shows the results of our analysis for different sizes of the training set n. We measure the classification error on the test set for each digit to get the error vector of length ten. We then find the average and variance of the error vector to show how the error is distributed over all tasks. We compare our method with the ℓ_1/ℓ_∞ reguralized method and LASSO.

V. PROOF OF MAIN RESULTS

In this section, we first recall some notations and definitions, and then provide a proof outline of all three theorems, which follow along similar lines. We then follow with the detailed proofs in the next Section VI.

A. Definitions and Preliminaries

We first collate the terms and notation we use throughout the proofs.

General Notations: For a vector v, the norms ℓ_1, ℓ_2 and ℓ_∞ are denoted as $\|v\|_1 = \sum_k \left|v^{(k)}\right|$, $\|v\|_2 = \sqrt{\sum_k \left|v^{(k)}\right|^2}$ and $\|v\|_\infty = \max_k \left|v^{(k)}\right|$, respectively. Also, for a matrix $Q \in \mathbb{R}^{p\times r}$ with rows denoted by $q^{(i)}$, the norm ℓ_ζ/ℓ_ρ is denoted as $\|Q\|_{\rho,\zeta} = \|\left(\|q^{(1)}\|_\zeta, \cdot\cdot\cdot, \|q^{(p)}\|_\zeta\right)\|_\rho$. The maximum singular value of Q is denoted as $\lambda_{max}(Q)$. For a matrix $X \in \mathbb{R}^{n\times p}$ and a set of indices $\mathcal{U} \subseteq \{1, \cdot\cdot\cdot, p\}$, the matrix $X_{\mathcal{U}} \in \mathbb{R}^{n\times|\mathcal{U}|}$ represents the sub-matrix of X consisting of X_j's where $j \in \mathcal{U}$.

Sparse Matrix Notations: For any matrix S, define $\text{Supp}(S) = \{(j,k) : s_j^{(k)} \neq 0\}$, and let $U_s = \{S \in \mathbb{R}^{p\times r} : \text{Supp}(S) \subseteq \text{Supp}(S^*)\}$ be the subspace of matrices whose support is the subset of the matrix S^*. The orthogonal projection to the subspace U_s can be defined as follows:

$$\left(P_{U_s}(S)\right)_{j,k} = \begin{cases} s_j^{(k)} & (j,k) \in \text{Supp}(S^*) \\ 0 & \text{ow.} \end{cases}$$

We can define the orthogonal complement space of U_s to be $U_s^c = \{S \in \mathbb{R}^{p\times r} : \text{Supp}(S) \cap \text{Supp}(S^*) = \emptyset\}$. The orthogonal projection to this space can be defined as $P_{U_s^c}(S) = S - P_{U_s}(S)$. Since the type of the block-sparsity we consider is a block-sparsity assumption on the rows of matrices, we need to characterize the sparsity of the rows of the matrix S^*.

As an important piece of notation, we denote $D(S) = \max_{1\leq j\leq p} \|s_j\|_0$ denoting the maximum number of non-zero elements in any row of the sparse matrix S.

Row-Sparse Matrix Notations: For any matrix B, define $\text{RowSupp}(B) = \{j : \exists k \quad \text{s.t.} \quad b_j^{(k)} \neq 0\}$, and let $U_b = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \subseteq \text{RowSupp}(B^*)\}$ be the subspace of matrices whose their row support is the subset of the row support of the matrix B^*. The orthogonal projection to the subspace U_b can be defined as follows:

$$\left(P_{U_b}(B)\right)_j = \begin{cases} b_j & j \in \text{RowSupp}(B^*) \\ \mathbf{0} & \text{ow.} \end{cases}$$

We can define the orthogonal complement space of U_b to be $U_b^c = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \cap \text{RowSupp}(B^*) = \phi\}$. The orthogonal projection to this space can be defined as $P_{U_b^c}(B) = B - P_{U_b}(B)$.
As an important piece of notation, we denote $M_j(B) = \{k : |b_j^{(k)}| = \|b_j\|_\infty > 0\}$, for any matrix $B \in \mathbb{R}^{p\times r}$, as the set of indices corresponding to elements that achieve the maximum magnitude on the j^{th} row with positive or negative signs. We set $M_j(B) = \emptyset$ if $j \notin \text{RowSupp}(B^*)$. Also, let $M(B) = \min_{1\leq j\leq p} |M_j(B)|$ be the minimum number of elements who achieve the maximum absolute value in each row of the matrix B.

Splitting a Matrix: We now develop some machinery for analyzing the splits of any matrix into sparse and block-sparse components. For (2), let $d = \lfloor\frac{\lambda_b}{\lambda_s}\rfloor$; we will always ensure $1 \leq d \leq r$, where r is the number of tasks. Given this d, we now define two matrices B^*, S^*, such that $B^* + S^* = \bar{\Theta}$, as follows.

1) In each row $\bar{\Theta}_j$, let v_j be the $(d+1)^{th}$ largest magnitude of the elements in Θ_j. Then, set the $(j,k)^{th}$ element

$\frac{n}{200}$		**Our Model**	ℓ_1/ℓ_∞	**LASSO**	ℓ_1/ℓ_2
5%	Average Classification Error	8.8%	11.6%	12.8%	10.8%
	Variance of Error	0.58%	0.69%	0.58%	0.61%
	Average Row Support Size	B:159 $B+S$:170	170	123	145
	Average Support Size	S:18 $B+S$:1631	1700	539	832
10%	Average Classification Error	3.3%	4.8%	5.7%	4.7%
	Variance of Error	0.43%	0.52%	0.58%	0.49%
	Average Row Support Size	B:211 $B+S$:226	217	173	195
	Average Support Size	S:34 $B+S$:2118	2165	821	1432
20%	Average Classification Error	2.2%	3.2%	4.2%	2.9%
	Variance of Error	0.26%	0.38%	0.35%	0.24%
	Average Row Support Size	B:270 $B+S$:299	368	354	350
	Average Support Size	S:67 $B+S$:2761	3669	2053	2213

TABLE II
SIMULATION RESULTS FOR OUR MODEL, ℓ_1/ℓ_∞ AND LASSO.

$s_j^{*(k)}$ of the matrix S^* as

$$s_j^{*(k)} \;=\; \operatorname{sign}(\theta_j^{(k)}) \max\left\{0, \left|\theta_j^{(k)}\right| - v_j\right\}$$

2) Given the matrix S^*, set B^* as the residual

$$B^* = \bar{\Theta} - S^*.$$

We use the transform $\mathcal{H}_d(\bar{\Theta}) = (B^*, S^*)$ to denote the output of this procedure.

Note that for each row of the matrix $\bar{\Theta}$, we set the corresponding row in S^* by taking the *clipped excess* over the $(d+1)^{th}$ largest magnitude element in that row of $\bar{\Theta}$. S^* will thus have at most d non-zero elements in each row. Correspondingly, each row of B^* is either identically 0, or has at least d non-zero elements of the same magnitude (equal to $(d+1)^{th}$ largest magnitude element in that row of $\bar{\Theta}$). Note also that if any element (j,k) is non-zero in both S^* and B^*, then its sign is the same in both. S^* thus takes on the role of the "true sparse matrix", and B^* the role of the "true block-sparse matrix". As we will see, such a split (B^*, S^*) has the following significance: our results will imply that if we have infinite samples, then (B^*, S^*) will be the solution to (2).

The following technical lemma is useful in the proof of all three theorems and summarizes the properties of $\mathcal{H}_d(\cdot)$.

Lemma 1. *If $(B,S) = \mathcal{H}_d(\Theta)$ then*

(P1) $M(B) \geq d+1$ *and* $D(S) \leq d$.
(P2) *sign*$(s_j^{(k)})$ = *sign*$(b_j^{(k)})$ *for all* $j \in$ *RowSupp*(B) *and* $k \in M_j(B)$.
(P3) $s_j^{(k)} = 0$ *for all* $j \in$ *RowSupp*(B) *and* $k \notin M_j(B)$.

Proof: The proof follows from the definition of $\mathcal{H}$. ■

Necessary Conditions for Optimality: Before we proceed, we characterize the properties of the solution of 2 in the following lemma.

Lemma 2. *If $(\hat{S}, \hat{B})$ is a solution (uniqueness is <u>NOT</u> required) of (2) then the following properties hold*

(P1) *sign*$(\hat{s}_j^{(k)})$ = *sign*$(\hat{b}_j^{(k)})$ *for all* $(j,k) \in$ *Supp*$(\hat{S})$ *with* $j \in$ *RowSupp*$(\hat{B})$.
(P2) *if* $\frac{\lambda_b}{\lambda_s}$ *is not an integer,* $\frac{1}{D(\hat{S})} > \frac{\lambda_s}{\lambda_b} > \frac{1}{M(\hat{B})}$.
(P3) $\left|\hat{b}_j^{(k)}\right| = \left\|\hat{b}_j\right\|_\infty$ *for all* $(j,k) \in$ *Supp*$(\hat{S})$.
(P4) *if* $\frac{\lambda_b}{\lambda_s}$ *is not an integer,* $\forall j\, \exists k$ *such that* $(j,k) \notin$ *Supp*$(\hat{S})$ *and* $\left|\hat{b}_j^{(k)}\right| = \left\|\hat{b}_j\right\|_\infty$.

This lemma shows that $(\hat{S}, \hat{B}) = \mathcal{H}_d(\widehat{\Theta})$, for $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$, which was our motivation behind the definition of the transformation $\mathcal{H}_d(\cdot)$. The next lemma shows why the assumption that the ratio of penalty regularizer parameters is crucial for our analysis.

Lemma 3. *If $(\hat{S}, \hat{B})$ with $\hat{B} \neq \mathbf{0}$ is a solution to (2) and $d = \frac{\lambda_b}{\lambda_s}$ is an integer then $(\hat{S}, \hat{B})$ is not the unique solution.*

For the sake of completeness, we revisit the necessary first-order optimality condition in the next lemma.

Lemma 4 (Convex Optimality). *If $(\hat{B}, \hat{S})$ is a solution of* (2)*, then there exists a dual matrix $\hat{Z} \in \mathbb{R}^{p \times r}$, such that $\hat{Z} \in \lambda_s \partial \|\hat{S}\|_{1,1}$ and $\hat{Z} \in \lambda_b \partial \|\hat{B}\|_{1,\infty}$ and for all $k = 1, \ldots, r$,*

$$\frac{1}{n}\left\langle X^{(k)}, X^{(k)} \right\rangle \left(\hat{s}^{(k)} + \hat{b}^{(k)}\right) - \frac{1}{n}(X^{(k)})^T y^{(k)} + \hat{z}^{(k)} = 0. \tag{3}$$

B. Proof Overview

The proofs of all three of our theorems follow a primal-dual witness technique, and consist of two steps, as detailed in this section. The first step constructs a primal-dual witness candidate, and is common to all three theorems. The second step consists of showing that the candidate constructed in the first step is indeed a dual witness. The theorem proofs differ only in this second step, and show that under the respective conditions imposed in the theorems, the construction succeeds with high probability. These steps are as follows:

STEP 1: Considering Lemma 2, it is clear that the solution of (2) $\widehat{\Theta} = \hat{S} + \hat{B}$ for $(\hat{S}, \hat{B})$ satisfies $(\hat{S}, \hat{B}) = \mathcal{H}_d(\widehat{\Theta})$, where $\mathcal{H}_d(\cdot)$ is defined in Section V-A for $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$. Let $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$ with properties summarized in Lemma 1. Now, if we *construct* a primal pair $(\tilde{S}, \tilde{B})$ whose signs agree with those of (S^*, B^*), and show that this is a unique solution of the M-estimation problem in (2) with high probability, then it follows that the sparsity patterns of the M-estimate (2) and (S^*, B^*) agree, and hence and so do the sparsity patterns of $\widehat{\Theta}$ and $\bar{\Theta}$. Thus, for the rest of the proof, our focus is to construct such a primal pair $(\tilde{S}, \tilde{B})$.

Primal Candidate: We design a candidate optimal solution $(\tilde{S}, \tilde{B})$ with the desired sparsity pattern using a restricted support optimization problem, called *oracle problem*:

$$(\tilde{S}, \tilde{B}) \in \arg \min_{S \in U_s, B \in U_b} \frac{1}{2n} \sum_{k=1}^{r} \left\| y^{(k)} - X^{(k)} \left(s^{(k)} + b^{(k)} \right) \right\|_2^2 + \lambda_s \|S\|_{1,1} + \lambda_b \|B\|_{1,\infty}. \tag{4}$$

This pair $(\tilde{S}, \tilde{B})$ has support constrained to lie within that of (S^*, B^*). We still need to make sure that the signs agree and this is the unique pair with these properties.

Sufficient Optimality Conditions: The following lemma specifies a set of sufficient (stationary) optimality conditions for the $(\tilde{S}, \tilde{B})$ from (4) to be the unique solution of the (unrestricted) optimization problem (2) while having the same sign as (S^*, B^*):

Lemma 5. *Under our assumptions on the design matrices $X^{(k)}$, the matrix pair $(\tilde{S}, \tilde{B})$ is the unique solution of the problem* (2) *if there exists a matrix $\widetilde{Z} \in \mathbb{R}^{p \times r}$ such that*

(C1) $P_{U_s}(\widetilde{Z}) = \lambda_s sign\left(\tilde{S}\right)$.

(C2) $[P_{U_b}(\widetilde{Z})]_{jk} = \begin{cases} t_j^{(k)}\, sign\left(\tilde{b}_j^{(k)}\right), & k \in M_j(\tilde{B}) \\ 0 & o.w.. \end{cases}$,
where, $t_j^{(k)} \geq 0$ such that $\sum_{k \in M_j(\tilde{B})} t_j^{(k)} = \lambda_b$.

(C3) $\left\| P_{U_s^c}(\widetilde{Z}) \right\|_{\infty,\infty} < \lambda_s$.

(C4) $\left\| P_{U_b^c}(\widetilde{Z}) \right\|_{\infty,1} < \lambda_b$.

(C5) $\frac{1}{n} \left\langle X^{(k)}, X^{(k)} \right\rangle \left(\tilde{b}^{(k)} + \tilde{s}^{(k)} \right) - \frac{1}{n} (X^{(k)})^T y^{(k)} + \tilde{z}^{(k)} = 0$
for all $1 \leq k \leq r$.

Proof: By conditions (C1) and (C3), $\frac{1}{\lambda_s}\widetilde{Z} \in \partial \|\tilde{S}\|_{1,1}$ and by conditions (C2) and (C4), $\frac{1}{\lambda_b}\widetilde{Z} \in \partial \|\tilde{B}\|_{1,\infty}$. Thus, $(\tilde{S}, \tilde{B}, \widetilde{Z})$ is a feasible primal-dual pair of (2) by the first-order optimality condition (C5). It remains to show the uniqueness to conclude that $(\tilde{S}, \tilde{B}) = (\hat{S}, \hat{B})$.

Let $\mathbb{B}$ and $\mathbb{S}$ to be balls of ℓ_∞/ℓ_1 and ℓ_∞/ℓ_∞ with radiuses λ_b and λ_s, respectively. Considering the fact that $\lambda_b \|B\|_{1,\infty} = \sup_{Z \in \mathbb{B}} \langle Z, B \rangle$ and $\lambda_s \|S\|_{1,1} = \sup_{Z \in \mathbb{S}} \langle Z, S \rangle$, the problem (2) can be written as

$$(\hat{S}, \hat{B}) = \arg \inf_{S,B} \sup_{Z \in \mathbb{B} \cap \mathbb{S}} \left\{ \frac{1}{2n} \sum_{k=1}^{r} \left\| y^{(k)} - X^{(k)} \left(b^{(k)} + s^{(k)} \right) \right\|_2^2 + \langle Z, S \rangle + \langle Z, B \rangle \right\}.$$

This saddle-point problem is strictly feasible and convex-concave. Given the dual variable $\widetilde{Z}$, and the primal optimal pair $(\hat{S}, \hat{B})$, we have $\lambda_b \|\hat{B}\|_{1,\infty} = \left\langle \widetilde{Z}, \hat{B} \right\rangle$ and $\lambda_s \|\hat{S}\|_{1,1} = \left\langle \widetilde{Z}, \hat{S} \right\rangle$. This implies that $\hat{b}_j = \mathbf{0}$ if $\|\tilde{z}_j\|_1 < \lambda_b$ (because $\lambda_b \sum_j \|\hat{b}_j\|_\infty \leq \sum_j \|\tilde{z}_j\|_1 \|\hat{b}_j\|_\infty$ and if $\|\tilde{z}_{j_0}\|_1 < \lambda_b$ for some j_0, then others can not compensate for that in the sum due to the fact that $\widetilde{Z} \in \mathbb{B}$, i.e., $\|\tilde{z}_j\|_1 \leq \lambda_b$). It also implies that $\hat{s}_j^{(k)} = 0$ if $\left| \tilde{z}_j^{(k)} \right| < \lambda_s$ for a similar reason. Hence, $P_{U_b^c}(\hat{B}) = 0$ and $P_{U_s^c}(\hat{S}) = 0$. This argument rules out the possibility of having a non-sparse solution. Thus, solving the restricted problem (4) is equivalent to solving the problem (2), because the oracle problem only restricts $\tilde{S}$ and $\tilde{B}$ to be zero outside the support of (S^*, B^*) and existence of $\widetilde{Z}$ implies that $\hat{S}$ and $\hat{B}$ are zero outside the support of (S^*, B^*).

The uniqueness follows from our (stationary) assumptions on design matrices $X^{(k)}$ that the matrix $\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle$ is invertible for all $1 \leq k \leq r$. Using this assumption, the problem (4) is *strictly* convex and the solution is unique. Consequently, the solution of (2) is also unique, since we showed that these two problems are equivalent. This concludes the proof of the lemma.

■

Dual Candidate: We need to construct a dual candidate $\widetilde{Z}$ that satisfies (C1)-(C5) in Lemma 5. Specifically, we construct $\widetilde{Z}$ as the superposition of three components with disjoint supports as follows

$$\widetilde{Z} = \widetilde{Z}_s + \widetilde{Z}_b + \widetilde{Z}_{\mathcal{U}^c},$$

where, $\widetilde{Z}_s = \lambda_s \text{sign}(\tilde{S})$ is supported on $\text{Supp}(\tilde{S})$, and $\widetilde{Z}_b$ is supported on $\text{Supp}(\tilde{B}) - \text{Supp}(\tilde{S})$ defined as

$$(\tilde{z}_b)_j^{(k)} = \begin{cases} \dfrac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left| M_j(\tilde{B}) \right| - \|\tilde{s}_j\|_0} \text{sign}\left(\tilde{b}_j^{(k)} \right) & \\ \qquad\qquad k \in M_j(\tilde{B}) \quad \& \quad (j,k) \notin \text{Supp}(\tilde{S}) & , \\ 0 & \text{ow} \end{cases}$$

and finally, $\widetilde{Z}_{\mathcal{U}^c}$, supported on $j \in \mathcal{U}^c$, is set as

$$(\tilde{z}_{\,\mathcal{U}^c})_j^{(k)} = \frac{1}{n} \left(X_j^{(k)} \right)^T w^{(k)} - \frac{1}{n} \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(\frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} - \tilde{z}_{\mathcal{U}_k}^{(k)} \right). \tag{5}$$

It is easy to check that conditions (C1) and (C2) in Lemma 5 are satisfied. To check condition (C5), let $\Delta = \tilde{B} + \tilde{S} - B^* - S^*$. From the first-order optimality conditions for the oracle problem (4), we have

$$\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \Delta_{\mathcal{U}_k}^{(k)} - \frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} + \tilde{z}_{\mathcal{U}_k}^{(k)} = 0.$$

and consequently,

$$\Delta_{\mathcal{U}_k}^{(k)} = \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \left(\frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} - \tilde{z}_{\mathcal{U}_k}^{(k)} \right). \tag{6}$$

Solving for $\tilde{z}_{\mathcal{U}^c}^{(k)}$, for all $j \in \mathcal{U}_k^c$, we get

$$(\tilde{z}_{\,\mathcal{U}^c})_j^{(k)} = -\frac{1}{n} \left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \Delta_{\mathcal{U}_k}^{(k)} + \frac{1}{n} \left(X_j^{(k)} \right)^T w^{(k)}.$$

Substituting for the value of $\Delta_{\mathcal{U}_k}^{(k)}$, we get (5). Thus, condition (C5) in Lemma 5 is also satisfied. It remains to show that the conditions (C3) and (C4) are also satisfied.

STEP 2: This step consists of showing that the pair $(\tilde{S}, \tilde{B}, \widetilde{Z})$ constructed in the earlier step is actually a *feasible* primal-dual pair of (2).*It only remains to guarantee (C3) and (C4) separately for each of the theorems.*

Indeed, this is where the proofs of the theorems differ. Specifically, Lemmas 6, 8 and 11 ensure these conditions are satisfied with given sample complexities in Theorems 1, 2 and 3, respectively.

VI. Proofs

The proofs of our three main theorems are in sections VI-A, VI-B and VI-C respectively.

A. Proof of Theorem 1

Let $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor$ and $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$. Then, the result follows from Proposition 1 below.

Proposition 1 (Structure Recovery)**.** *Under assumptions of Theorem 1, with probability* $1 - c_1 \exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2*, we are guaranteed that the following properties hold:*

(P1) *Problem (2) has unique solution $(\hat{S}, \hat{B})$ such that Supp$(\hat{S}) \subseteq$ Supp(S^*) and RowSupp$(\hat{B}) \subseteq$ RowSupp(B^*).*

(P2) $$\left\| \hat{B} + \hat{S} - B^* - S^* \right\|_{\infty,\infty} \leq \underbrace{\sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max}}_{\delta_{\min}}.$$

(P3) *sign (Supp$(\hat{s}_j)$) = sign (Supp(s_j^*))*
for all $j \notin$ RowSupp(B^*) provided that

$$\min_{\substack{j \notin RowSupp(B^*) \\ (j,k) \in Supp(S^*)}} \left| s_j^{*(k)} \right| > \delta_{\min}.$$

(P4) *sign $\left(\text{Supp}(\hat{s}_j + \hat{b}_j)\right)$ = sign $\left(\text{Supp}(s_j^* + b_j^*)\right)$*
for all $j \in$ RowSupp(B^*) provided that

$$\min_{(j,k) \in Supp(B^*)} \left| b_j^{*(k)} + s_j^{*(k)} \right| > \delta_{\min}.$$

Proof: We prove the result separately for each part.

(P1) Considering the constructed primal-dual pair, it suffices to show that (C3) and (C4) in Lemma 5 are satisfied with high probability. By Lemma 6, with probability at least $1 - c_1 \exp(-c_2 n)$ those two conditions hold and hence, $(\hat{S}, \hat{B}) = (\tilde{S}, \tilde{B})$ is the unique solution of (2) and the property (P1) follows.

(P2) Using (6), we have

$$\begin{aligned}
\max_{j \in \mathcal{U}_k} \left| \Delta_j^{(k)} \right| &\leq \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \frac{1}{n} \left(X_{\mathcal{U}_k}^{(k)} \right)^T w^{(k)} \right\|_\infty \\
&\quad + \left\| \left(\frac{1}{n} \left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)} \right\rangle \right)^{-1} \tilde{z}_{\mathcal{U}_k}^{(k)} \right\|_\infty \\
&\leq \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max},
\end{aligned}$$

where, the second inequality holds with high probability as a result of Lemma 7 for $\alpha = \epsilon \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}}$ for some $\epsilon > 1$, considering the fact that Var $\left(\Delta_j^{(k)} \right) \leq \frac{\sigma^2}{C_{min} n}$.

(P3) Using (P1) in Lemma 2, this event is equivalent to the event that for all $j \notin$ RowSupp(B^*) with $(j,k) \in$ Supp(S^*), we have $\left(\Delta_j^{(k)} + s_j^{*(k)} \right)$ sign $\left(s_j^{*(k)} \right) > 0$. By Hoeffding inequality, we have

$$\begin{aligned}
\mathbb{P} &\left[\left(\Delta_j^{(k)} + s_j^{*(k)} \right) \text{sign} \left(s_j^{*(k)} \right) > 0 \right] \\
&= \mathbb{P} \left[-\Delta_j^{(k)} \text{sign} \left(s_j^{*(k)} \right) < \left| s_j^{*(k)} \right| \right] \\
&\geq \mathbb{P} \left[\left| \Delta_j^{(k)} \right| < \left| s_j^{*(k)} \right| \right].
\end{aligned}$$

By part (P2), this event happens with high probability if $\min_{\substack{j \notin \text{RowSupp}(B^*) \\ (j,k) \in \text{Supp}(S^*)}} \left| s_j^{*(k)} \right| > b_{\min}$.

(P4) Using (P1) in Lemma 2, this event is equivalent to the event that for all $j \in$ RowSupp(B^*), we have $\left(\Delta_j^{(k)} + b_j^{*(k)} + s_j^{*(k)} \right)$ sign $\left(b_j^{*(k)} + s_j^{*(k)} \right) > 0$. By Hoeffding inequality, we have

$$\begin{aligned}
\mathbb{P} &\left[\left(\Delta_j^{(k)} + b_j^{*(k)} + s_j^{*(k)} \right) \text{sign} \left(b_j^{*(k)} + s_j^{*(k)} \right) > 0 \right] \\
&= \mathbb{P} \left[-\Delta_j^{(k)} \text{sign} \left(b_j^{*(k)} + s_j^{*(k)} \right) < \left| b_j^{*(k)} + s_j^{*(k)} \right| \right] \\
&\geq \mathbb{P} \left[\left| \Delta_j^{(k)} \right| < \left| b_j^{*(k)} + s_j^{*(k)} \right| \right].
\end{aligned}$$

By part (P2), this event happens with high probability if $\min_{(j,k) \in \text{Supp}(B^*)} \left| b_j^{*(k)} + s_j^{*(k)} \right| > b_{\min}$.

■

Lemma 6. *Under conditions of Proposition 1, the conditions (C3) and (C4) in Lemma 5 hold for the constructed primal-dual pair with probability at least* $1 - c_1 \exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2.

Proof: First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice that

$$\left\| P_{U_s^c}(\widetilde{Z}) \right\|_{\infty,\infty} \leq \max \left(\frac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left| M_j(\tilde{B}) \right| - \|\tilde{s}_j\|_0}, \left| (\tilde{z}_{\mathcal{U}^c})_j^{(k)} \right| \right).$$

By our assumption on the ratio of the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s \|\tilde{s}_j\|_0}{\left| M_j(\tilde{B}) \right| - \|\tilde{s}_j\|_0} < \lambda_s$. Moreover, we have

$$\left|(\tilde{z}_{\mathcal{U}^c})_j^{(k)}\right| \leq \max_{1\leq k\leq r}\max_{j\in\mathcal{U}_k^c}\left\|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\|_1 \left(\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty + \left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty\right)$$
$$+\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty$$
$$\leq (2-\gamma_s)\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty + (1-\gamma_s)\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$\leq (2-\gamma_s)\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty + (1-\gamma_s)\lambda_s.$$

Thus, the event $\|P_{U_s^c}(\widetilde{Z})\|_{\infty,\infty} < \lambda_s$ is equivalent to the event $\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \frac{\gamma_s}{2-\gamma_s}\lambda_s$. By Lemma 7, this event happens with probability at least $1-2\exp\left(-\frac{\gamma_s^2 n\lambda_s^2}{4(2-\gamma_s)^2\sigma^2}+\log(pr)\right)$. This probability goes to 1 if $\lambda_s > \frac{2(2-\gamma_s)\sigma\sqrt{\log(pr)}}{\gamma_s\sqrt{n}}$ as stated in the assumptions.

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} \leq \max\left(\lambda_s\|\tilde{s}_j\|_0, \sum_{k=1}^r\left|(\tilde{z}_{\mathcal{U}^c})_j^{(k)}\right|\right).$$

We have $\lambda_s\|\tilde{s}_j\|_0 \leq \lambda_s D(S^*) < \lambda_b$ by our assumption on the ratio of the penalty regularizer coefficients. We can establish the following bound:

$$\sum_{k=1}^r\left|\tilde{z}_j^{(k)}\right|$$
$$\leq \max_{j\in\mathcal{U}^c}\sum_{k=1}^r\left\|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\|_1 \left(\max_{j\in\bigcup_{k=1}^r\mathcal{U}_k}\left\|\tilde{z}_j^{(k)}\right\|_1 + \max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty\right)$$
$$+\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty$$
$$\leq (1-\gamma_b)\lambda_b + (2-\gamma_b)\max_{1\leq k\leq K}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty.$$

Thus, the event $\|P_{U_b^c}(\widetilde{Z})\|_{\infty,1} < \lambda_b$ is equivalent to the event $\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \frac{\gamma_b}{2-\gamma_b}\lambda_b$. By Lemma 7, this event happens with probability at least $1-2\exp\left(-\frac{\gamma_b^2 n\lambda_b^2}{4(2-\gamma_b)^2\sigma^2}+\log(pr)\right)$. This probability goes to 1 if $\lambda_b > \frac{2(2-\gamma_b)\sigma\sqrt{\log(pr)}}{\gamma_b\sqrt{n}}$ as stated in the assumptions.

Hence, with probability at least $1-c_1\exp(-c_2 n)$ conditions (C3) and (C4) in Lemma 5 are satisfied. ∎

Lemma 7.

$$\mathbb{P}\left[\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \alpha\right] \geq 1-2\exp\left(-\frac{\alpha^2 n}{2\sigma^2}+\log(pr)\right).$$

Proof: Since $w_j^{(k)}$'s are distributed as $\mathcal{N}(0,\sigma^2)$, we have $\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}$ distributed as $\mathcal{N}\left(0, \frac{\sigma^2}{n}\left(X^{(k)}\right)^T X_{\mathcal{U}_k}^{(k)}\right)$. Using concentration of Gaussian variables, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty \geq \alpha\right] \leq \sum_{j=1}^p \mathbb{P}\left[\left|\frac{1}{n}\left(X_j^{(k)}\right)^T w^{(k)}\right| \geq \alpha\right]$$
$$\leq \sum_{j=1}^p 2\exp\left(-\frac{\alpha^2 n}{2\sigma^2\left(X_j^{(k)}\right)^T X_j^{(k)}}\right)$$
$$\leq 2p\exp\left(-\frac{\alpha^2 n}{2\sigma^2}\right).$$

By union bound, the result follows. ∎

B. Proof of Theorem 2

Let $d = \lfloor\frac{\lambda_b}{\lambda_s}\rfloor$ and $(B^*, S^*) = \mathcal{H}_d(\bar{\Theta})$. Then, the result follows from the next proposition.

Proposition 2. *Under assumptions of Theorem 2, if*

$$n > \max\left(\frac{Bs\log(pr)}{C_{min}\gamma_s^2}, \frac{Bsr\big(r\log(2)+\log(p)\big)}{C_{min}\gamma_b^2}\right)$$

then with probability at least $1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right) - c_3\exp(-c_4\log(rs))$ *for some positive constants* $c_1 - c_4$*, we are guaranteed that the following properties hold:*

(P1) *The solution* $(\hat{B},\hat{S})$ *to (2) is unique and RowSupp*$(\hat{B}) \subseteq$ *RowSupp*(B^*) *and Supp*$(\hat{S}) \subseteq$ *Supp*(S^*).

(P2) $$\left\|\hat{B}+\hat{S}-B^*-S^*\right\|_\infty \leq \underbrace{\sqrt{\frac{50\sigma^2\log(rs)}{nC_{min}}}+\lambda_s\left(\frac{Ds}{C_{min}\sqrt{n}}+D_{max}\right)}_{g_{\min}}.$$

(P3) *sign* $\left(\textit{Supp}(\hat{s}_j)\right) =$ *sign* $\left(\textit{Supp}(s_j^*)\right)$ *for all* $j \notin$ *RowSupp*(B^*) *provided that*

$$\min_{\substack{j\notin RowSupp(B^*)\\(j,k)\in Supp(S^*)}}\left|s_j^{*(k)}\right| > g_{\min}.$$

(P4) *sign* $\left(\textit{Supp}(\hat{s}_j+\hat{b}_j)\right) =$ *sign* $\left(\textit{Supp}(s_j^*+b_j^*)\right)$ *for all* $j \in$ *RowSupp*(B^*) *provided that*

$$\min_{(j,k)\in Supp(B^*)}\left|b_j^{*(k)}+s_j^{*(k)}\right| > g_{\min}.$$

Proof: We provide the proof of each part separately.

(P1) Considering the constructed primal-dual pair $(\tilde{S},\tilde{B},\widetilde{Z})$, it suffices to show that the conditions (C3) and (C4) in Lemma 5 are satisfied under these assumptions. Lemma 8 guarantees that with probability at least $1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right)$ those conditions are satisfied. Hence, $(\hat{B},\hat{S}) = (\tilde{B},\tilde{S})$ are the unique solution to (2) and (P1) follows.

(P2) From (6), we have

$$\max_{j\in\mathcal{U}_k}\left|\Delta_j^{(k)}\right| \le \underbrace{\left\|\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\frac{1}{n}\left(X_{\mathcal{U}_k}^{(k)}\right)^T w^{(k)}\right\|_\infty}_{\mathcal{W}^{(k)}}$$
$$+\left\|\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$\le \left\|\mathcal{W}^{(k)}\right\|_\infty + \left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$+\left\|\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty .$$

We need to bound these three quantities. Notice that

$$\left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty \le \left\|\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right\|_{\infty,1}\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$\le D_{max}\lambda_s.$$

Also, we have

$$\left\|\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$\le \lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_2$$
$$\le \lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)\sqrt{s}\lambda_s$$
$$\le \frac{4}{C_{min}}\sqrt{\frac{s}{n}}\sqrt{s}\lambda_s,$$

where, the last inequality holds with probability at least $1 - c_1\exp\left(-c_2\left(\sqrt{n}-\sqrt{s}\right)^2\right)$ for some positive constants c_1 and c_2 as a result of Davidson and Szarek [20] on eigenvalues of Gaussian random matrices. Conditioned on $X_{\mathcal{U}_k}^{(k)}$, the vector $\mathcal{W}^{(k)} \in \mathbb{R}^{|\mathcal{U}_k|}$ is a zero-mean Gaussian random vector with covariance matrix $\frac{\sigma^2}{n}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}$. Thus, we have

$$\frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right)$$
$$\le \frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1} - \left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)$$
$$+\frac{1}{n}\lambda_{max}\left(\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right)$$
$$\le \frac{1}{n}\left(\frac{4}{C_{min}}\sqrt{\frac{s}{n}} + \frac{1}{C_{min}}\right)$$
$$\le \frac{5}{nC_{min}}.$$

From the concentration of Gaussian random variables (Lemma 7) and using the union bound, we get

$$\mathbb{P}\left[\max_{1\le k\le r}\left\|\mathcal{W}^{(k)}\right\|_\infty \ge t\right] \le 2\exp\left(-\frac{t^2 nC_{min}}{50\sigma^2} + \log(rs)\right).$$

For $t = \epsilon\sqrt{\frac{50\sigma^2\log(rs)}{nC_{min}}}$ for some $\epsilon > 1$, the result follows.

(P3),(P4) The results are immediate consequence of (P2).

■

Lemma 8. *Under the assumptions of Proposition 2, the conditions (C3) and (C4) in Lemma 5 hold for the constructed primal-dual pair with probability at least* $1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right)$ *for some positive constants* c_1 *and* c_2.

Proof: First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice that

$$\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty} \le \max\left(\frac{\lambda_b - \lambda_s\|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0}, \left|(\tilde{z}_{\mathcal{U}^c})_j^{(k)}\right|\right).$$

By our assumptions on the ratio of the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s\|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0} < \lambda_s$. For all $j \in \bigcap_{k=1}^r \mathcal{U}_k$ and $R \in \mathbb{R}^{p\times r}$ with i.i.d. standard Gaussian entries (see Lemma 4 in [13]), we have

$$\left|(\tilde{z}_{\mathcal{U}^c})_j^{(k)}\right|$$
$$\le \max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I} - \frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}}$$
$$+\max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle\tilde{z}_{\mathcal{U}_k}^{(k)}\right|$$
$$\le \max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left\|\Sigma_{j,\mathcal{U}_k}^{(k)}\left(\Sigma_{\mathcal{U}_k,\mathcal{U}_k}^{(k)}\right)^{-1}\right\|_1\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_\infty$$
$$+\max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\underbrace{\left|\frac{1}{n}\left\langle R_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle\tilde{z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{R}_j^{(k)}}$$
$$\le (1-\gamma_s)\lambda_s + \max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|,$$

The second inequality follows from a simple application of the triangle inequality, following the line of argument in Appendix B of [13]. By Lemma 9, if $n \ge \frac{2}{2-\sqrt{3}}\log(pr)$ then with high probability $\left\|X_j^{(k)}\right\|_2^2 \le 2n$ and hence $\mathrm{Var}\left(\mathcal{W}_j^{(k)}\right) \le \frac{2\sigma^2}{n}$. Using the concentration results for the zero-mean Gaussian random variable $\mathcal{W}_j^{(k)}$ and using the union bound, we get

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| \ge t\right] \le 2\exp\left(-\frac{t^2 n}{4\sigma^2} + \log(p)\right) \qquad \forall t \ge 0.$$

Conditioning on $\left(X_{\mathcal{U}_k}^{(k)}, w^{(k)}, \tilde{z}^{(k)}\right)$'s, we have that $\mathcal{R}_j^{(k)}$ is a zero-mean Gaussian random variable with

$$\mathrm{Var}\left(\mathcal{R}_j^{(k)}\right) \le \frac{\left\|\tilde{z}_{\mathcal{U}_k}^{(k)}\right\|_2^2}{nC_{min}} \le \frac{s\lambda_s^2}{nC_{min}}.$$

By concentration of Gaussian random variables, we have

$$\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r\mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right| \ge t\right] \le 2\exp\left(-\frac{t^2 nC_{min}}{Bs\lambda_s^2} + \log(p)\right) \quad \forall t \ge 0.$$

Using these bounds, we get

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty}<\lambda_s\right]\\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| < \gamma_s\lambda_s \qquad \forall 1\leq k\leq r\right]\\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{R}_j^{(k)}\right| < t_0 \quad \forall 1\leq k\leq r\right]\\
&\qquad\qquad \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| < \gamma_s\lambda_s - t_0 \quad \forall 1\leq k\leq r\right]\\
&\geq \left(1-2\exp\left(-\frac{t_0^2 nC_{min}}{Bs\lambda_s^2}+\log(pr)\right)\right)\\
&\qquad\qquad \left(1-2\exp\left(-\frac{(\gamma_s\lambda_s-t_0)^2 n}{4\sigma^2}+\log(pr)\right)\right).
\end{aligned}$$

This probability goes to 1 for $t_0 = \frac{\sqrt{Bs}\lambda_s}{\sqrt{Bs}\lambda_s+2\sigma\sqrt{C_{min}}}\gamma_s\lambda_s$ (the solution to $\frac{t_0^2 C_{min}}{Bs\lambda_s^2} = \frac{(\gamma_s\lambda_s-t_0)^2}{4\sigma^2}$), if the regularization parameter $\lambda_s > \frac{\sqrt{4\sigma^2 C_{min}\log(pr)}}{\gamma_s\sqrt{nC_{min}}-\sqrt{Bs\log(pr)}}$ provided that $n > \frac{Bs\log(pr)}{C_{min}\gamma_s^2}$ as stated in the assumptions.

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} \leq \max\left(\lambda_s\|\tilde{s}_j\|_0, \sum_{k=1}^r \left|(\tilde{z}_{\mathcal{U}^c})_j^{(k)}\right|\right).$$

We have $\lambda_s\|\tilde{s}_j\|_0 \leq \lambda_s D(S^*) < \lambda_b$ by our assumption on the ratio of the penalty regularizer coefficients. For all $j \in \bigcap_{k=1}^r \mathcal{U}_k^c$, we have

$$\begin{aligned}
&\sum_{k=1}^r \left|\tilde{z}_j^{(k)}\right|\\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}}\\
&\quad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r \left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right|\\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r \left|\mathcal{W}_j^{(k)}\right|\\
&\quad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r \left\|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle\right\|_1\\
&\qquad\qquad\qquad\qquad\qquad\qquad \max_{j\in\bigcup_{k=1}^r \mathcal{U}_k}\left\|\tilde{z}_j^{(k)}\right\|_1\\
&\quad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r \underbrace{\left|\frac{1}{n}\left\langle R_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{R}_j^{(k)}}\\
&\leq (1-\gamma_b)\lambda_b + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{W}_j^{(k)}\right|.
\end{aligned}$$

We first note that for any $\mathbf{v} \in \{-1,+1\}^r$, we have

$$\mathrm{Var}\left(\sum_{k=1}^r v_k\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2 r}{n}.$$

Using the union bound and previous discussion, we get

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{W}_j^{(k)}\right| \geq t\right]\\
&\quad = \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^r v_k\mathcal{W}_j^{(k)} \geq t\right]\\
&\quad \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2 r}+r\log(2)+\log(p)\right) \qquad \forall t\geq 0.
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{k=1}^r\left|\mathcal{R}_j^{(k)}\right|\right) &= \mathrm{Var}\left(\sum_{k=1}^r v_k\mathcal{R}_j^{(k)}\right)\\
&\leq \frac{\sum_{k=1}^r\left\|\tilde{z}_j^{(k)}\right\|_2^2}{nC_{min}} \leq \frac{rs\lambda_s^2}{nC_{min}} < \frac{rs\lambda_b^2}{nC_{min}}
\end{aligned}$$

and consequently by concentration of Gaussian variables,

$$\begin{aligned}
&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^K\left|\mathcal{R}_j^{(k)}\right| \geq t\right]\\
&\quad = \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^r v_k\mathcal{R}_j^{(k)} \geq t\right]\\
&\quad \leq 2\exp\left(-\frac{t^2 nC_{min}}{2rs\lambda_b^2}+r\log(2)+\log(p)\right) \qquad \forall t\geq 0.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1}<\lambda_b\right]\\
&\quad\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{R}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{W}_j^{(k)}\right| < \gamma_b\lambda_b\right]\\
&\quad\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{R}_j^{(k)}\right| < t_0\right]\\
&\qquad\qquad \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\sum_{k=1}^r\left|\mathcal{W}_j^{(k)}\right| < \gamma_b\lambda_b - t_0\right]\\
&\quad\geq \left(1-2\exp\left(-\frac{t_0^2 nC_{min}}{2rs\lambda_b^2}+r\log(2)+\log(p)\right)\right)\\
&\qquad\qquad \left(1-2\exp\left(-\frac{(\gamma_b\lambda_b-t_0)^2 n}{4\sigma^2 r}+r\log(2)+\log(p)\right)\right).
\end{aligned}$$

This probability goes to 1 for $t_0 = \frac{\sqrt{Bs}\lambda_b}{\sqrt{Bs}\lambda_b+2\sigma\sqrt{C_{min}}}\gamma_b\lambda_b$ (the solution to $\frac{(\gamma_b\lambda_b-t_0)^2 n}{4\sigma^2 r} = \frac{t_0^2 nC_{min}}{2rs\lambda_b^2}$), if

$$\lambda_b > \frac{\sqrt{4\sigma^2 C_{min} r\Big(r\log(2)+\log(p)\Big)}}{\gamma_b\sqrt{nC_{min}}-\sqrt{Bsr\Big(r\log(2)+\log(p)\Big)}},$$

provided that $n > \frac{Bsr(r\log(2)+\log(p))}{\gamma_b^2 C_{min}}$ as stated in the assumptions. Hence, with probability at least $1 - c_1\exp\left(-c_2\left(r\log(2)+\log(p)\right)\right)$ the conditions of the Lemma 5 are satisfied.

■

Lemma 9.

$$\mathbb{P}\left[\max_{1\leq k\leq r}\max_{1\leq j\leq p}\left\|X_j^{(k)}\right\|_2^2 \leq 2n\right] \geq 1-\exp\left(-(1-\frac{\sqrt{3}}{2})n+\log(pr)\right).$$

Proof: Notice that $\|X_j^{(k)}\|_2^2$ is a χ^2 random variable with n degrees of freedom. According to [21], we have

$$\mathbb{P}\left[\left\|X_j^{(k)}\right\|_2^2 \geq t + (\sqrt{t}+\sqrt{n})^2\right] \leq \exp(-t) \qquad \forall t \geq 0.$$

Letting $t = \left(\frac{\sqrt{3}-1}{2}\right)^2 n$ and using the union bound, the result follows. ■

C. Proof of Theorem 3

We will actually prove a more general theorem, from which Theorem 3 would follow as a corollary. Among shared features (with size αs), we say a fraction τ has different magnitudes on $\bar{\Theta}$ (i.e., a fraction $1-\tau$ of shared features have approximately same magnitude on both tasks. See Theorem 4 for exact definition): Let τ_1 be the fraction with larger magnitude on the first task and τ_2 the fraction with larger magnitude on the second task (so that $\tau = \tau_1 + \tau_2$). Moreover, let $\frac{\lambda_b}{\lambda_s} = \kappa$ and

$$f(\kappa) = f(\kappa,\tau,\alpha) = 2 - 2(1-\tau)\alpha - 2\tau\alpha\kappa + \left(\frac{1+\tau}{2}\right)\alpha\kappa^2,$$

and

$$g(\kappa,\tau,\alpha) = \max\left(\frac{2\,f(\kappa)}{\kappa^2}, f(\kappa)\right).$$

Theorem 4. *Under the assumptions of the Theorem 3, if*

$$\left|\left\{j \in \mathit{RowSupp}(B^*) : \left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| \leq c\lambda_s\right\}\right| = (1-\tau)\alpha s,$$

then, the result of Theorem 3 holds for

$$\theta(n,s,p,\alpha) = \frac{n}{g(\kappa,\tau,\alpha)\, s \log\left(p-(2-\alpha)s\right)}.$$

Corollary 4. *Under the assumptions of the Theorem 4, if the regularization penalties are set as $\kappa = \lambda_b/\lambda_s = \sqrt{2}$, then the result of Theorem 3 holds for $\theta(n,s,p,\alpha) = \frac{n}{\left(2-\alpha+(3-2\sqrt{2})\tau\alpha\right)s\log(p-(2-\alpha)s)}$.*

Proof: Follows trivially by substituting $\kappa = \sqrt{2}$ in Theorem 4. Indeed, this setting of κ can also be shown to minimize $g(\kappa,\tau,\alpha)$:

$$\begin{aligned}
&\min_{1<\kappa<2} \max\left(\frac{2\,f(\kappa)}{\kappa^2}, f(\kappa)\right)\\
&= \min\left(\min_{1<\kappa\leq\sqrt{2}} \frac{2}{\kappa^2}\left(f(\kappa)\right), \min_{\sqrt{2}<\kappa<2} f(\kappa)\right)\\
&= 2 - \alpha + (3-2\sqrt{2})\,\tau\,\alpha.
\end{aligned}$$

■

Proof of Theorem 3: The proof follows from Corollary 4 by setting $\tau = 0$ and $\kappa = \sqrt{2}$.

We will now set out to prove Theorem 4. We will first need the following lemma.

Lemma 10. *For any $j \in \mathit{RowSupp}(B^*)$, if $\left|S_j^{*(k)}\right| < c\lambda_s$ for some constant c specified in the proof, then $\tilde{S}_j^{(k)} = 0$ with probability at least $1 - c_1\exp(-c_2 n)$.*

Proof: Let $\check{S}$ be a matrix equal to $\tilde{S}$ except that $\check{S}_j^{(k)} = 0$. Using the concentration of Gaussian random variables and optimality of $\tilde{S}$, we get

$$\begin{aligned}
&\mathbb{P}\left[\left|\tilde{S}_j^{(k)}\right| > 0\right]\\
&\leq \mathbb{P}\Bigg[2n\lambda_s\left|\tilde{S}_j^{(k)}\right| < \left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right\|_2^2\\
&\qquad\qquad - \left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \tilde{S}^{(k)})\right\|_2^2\Bigg]\\
&= \mathbb{P}\Bigg[2n\lambda_s < \Bigg(\frac{\left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right\|_2^2}{\left\|\tilde{S}_j^{(k)}X_j^{(k)}\right\|_2}\\
&\qquad - \frac{\left\|y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)}) - \tilde{S}_j^{(k)}X_j^{(k)}\right\|_2^2}{\left\|\tilde{S}_j^{(k)}X_j^{(k)}\right\|_2}\Bigg)\left\|X_j^{(k)}\right\|_2\Bigg]\\
&= \mathbb{P}\left[2n\lambda_s < \frac{2\tilde{S}_j^{(k)}X_j^{(k)T}\left(y^{(k)} - X^{(k)}(\tilde{B}^{(k)} + \check{S}^{(k)})\right)}{\left\|\tilde{S}_j^{(k)}X_j^{(k)}\right\|_2}\left\|X_j^{(k)}\right\|_2\right]\\
&= \mathbb{P}\left[n\lambda_s < X_j^{(k)T}\left(X^{(k)}(B^{*(k)} + S^{*(k)} - \tilde{B}^{(k)} - \check{S}^{(k)}) + w^{(k)}\right)\right]
\end{aligned}$$

Since X's and w's are independent, then $X_j^{(k)T}w^{(k)} \leq \epsilon_1$ with high probability. Moreover, the vector $X_j^{(k)T}X^{(k)}$ is smaller than some ϵ_2 on entries different from j and is equal to $\|X_j^{(k)}\|_2^2$ on the j^{th} entry. Using the ℓ_∞ bound in Theorem 2, we have $\|B^{*(k)} + S^{*(k)} - \tilde{B}^{(k)} - \check{S}^{(k)}\|_\infty \leq \max(b_{\min}, S_j^{*(k)}) \leq S_j^{*(k)}$. Now, for a small constant $c < 1$ given by

$$c = \frac{1}{1 + \frac{\epsilon_2(s-1)+\epsilon_1/|S_j^{*(k)}|}{\|X_j^{(k)}\|_2^2}},$$

we have

$$\begin{aligned}
\mathbb{P}\left[\left|\tilde{S}_j^{(k)}\right| > 0\right] &\leq \mathbb{P}\left[n\lambda_s < \left|S_j^{*(k)}\right|\left(\epsilon_2(s-1) + \left\|X_j^{(k)}\right\|_2^2\right) + \epsilon_1\right]\\
&= \mathbb{P}\left[n\lambda_s < \frac{1}{c}\left|S_j^{*(k)}\right|\left\|X_j^{(k)}\right\|_2^2\right]\\
&= \mathbb{P}\left[\frac{c\lambda_s}{\left|S_j^{*(k)}\right|}n < \left\|X_j^{(k)}\right\|_2^2\right].
\end{aligned}$$

Notice that $\mathbb{E}[\|X_j^{(k)}\|_2^2] = n$. Using the concentration of χ^2 random variables (see [21]), this probability vanishes exponentially fast in n for $\left|S_j^{*(k)}\right| < c\lambda_s$.

■

D. Proof of Theorem 4

We will now provide the proofs of different parts separately.

Proof: **(Success):** Recall the constructed primal-dual pair $(\tilde{B}, \tilde{S}, \tilde{Z})$. It suffices to show that the dual variable $\tilde{Z}$ satisfies the conditions (C3) and (C4) of Lemma 5. By Lemma 11, these conditions are satisfied with probability at least $1 - c_1\exp(-c_2 n)$ for some positive constants c_1 and c_2. Hence, $(\hat{B}, \hat{S}) = (\tilde{B}, \tilde{S})$ is the unique optimal solution. The rest are direct consequences of Proposition 2 for $C_{min} = 1$ and $D_{max} = 1$.

(Failure): We prove this result by contradiction. Suppose there exist a solution to (2), say $(\hat{B}, \hat{S})$ such that $\operatorname{sign}\left(\operatorname{Supp}(\hat{B}+\hat{S})\right) = \operatorname{sign}\left(\operatorname{Supp}(B^*+S^*)\right)$. By Lemma 2, this is equivalent to having $\operatorname{sign}\left(\operatorname{Supp}(\hat{B})\right) = \operatorname{sign}\left(\operatorname{Supp}(B^*)\right)$ and $\operatorname{sign}\left(\operatorname{Supp}(\hat{s}_j)\right) = \operatorname{sign}\left(\operatorname{Supp}(S_j^*)\right)$ if $j \notin \operatorname{RowSupp}(B^*)$ and $\frac{\lambda_b}{\lambda_s} = \kappa$. Moreover for $j \in \operatorname{RowSupp}(B^*)$, if $\hat{s}_j^{(k)} \neq 0$ and $s_j^{*(k)} \neq 0$, then by Lemma 11, we have $\operatorname{sign}(\hat{s}_j^{(k)}) = \operatorname{sign}(s_j^{*(k)})$.

Now, suppose $n < (1-\nu)\max\left(\frac{2f(\kappa)}{\kappa^2}, f(\kappa)\right) s\log(p-(2-\alpha)s)$, for some $\nu > 0$. This entails that
either (i) $n < (1-\nu)f(\kappa)s\log(p-(2-\alpha)s)$,
or (ii) $n < (1-\nu)\left(\frac{2f(\kappa)}{\kappa^2}\right) s\log(p-(2-\alpha)s)$.

Case (i): We will show that with high probability, there exists k for which, there exists $j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c$ such that $\left|\tilde{Z}_j^{(k)}\right| > \lambda_s$. This is a contradiction to Lemma 4.

Using (5) and conditioning on $(X_{\mathcal{U}_k}^{(k)}, w^{(k)}, \tilde{Z}_{\mathcal{U}_k}^{(k)})$, for all $j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c$ we have that the random variables $\tilde{Z}_j^{(k)}$ are i.i.d. zero-mean Gaussian random variables with

$$\begin{aligned}
&\operatorname{Var}\left(\tilde{Z}_j^{(k)}\right)\\
&=\Bigg\|\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\tilde{Z}_{\mathcal{U}_k}^{(k)}\\
&\qquad+\frac{1}{n}\left(\mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right)w^{(k)}\Bigg\|_2^2\\
&=\left\|\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\tilde{Z}_{\mathcal{U}_k}^{(k)}\right\|_2^2\\
&\qquad+\left\|\frac{1}{n}\left(\mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right)w^{(k)}\right\|_2^2
\end{aligned}$$

The second equality holds by orthogonality of projections. We thus have

$$\begin{aligned}
&\operatorname{Var}\left(\tilde{Z}_j^{(k)}\right)\\
&\geq\max\Bigg(\lambda_{min}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right)\frac{\left\|\tilde{Z}_{\mathcal{U}_k}^{(k)}\right\|_2^2}{n}\\
&\qquad,\frac{\left\|\left(\mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right)w^{(k)}\right\|_2^2}{n^2}\Bigg)\\
&\geq\frac{\left\|\tilde{Z}_{\mathcal{U}_k}^{(k)}\right\|_2^2}{(\sqrt{n}+\sqrt{s})^2}
\end{aligned}$$

The first term in the maximum satisfies the second inequality with probability at least $1-c_1\exp\left(-c_2\left(\sqrt{n}+\sqrt{s}\right)^2\right)$ as a result of Theorem II.13 in [20] on the eigenvalues of Gaussian matrices. The second term in the maximum satisfies the second inequality holds with probability at least $1-c_3\exp(-c_4 n)$ as a result of Lemma 1 in [21]. Considering $\tilde{B}+\tilde{S}$, assume that among shared features (with size αs), a portion of τ_1 has larger magnitude on the fist task and a portion of τ_2 has larger magnitude on the second task (and consequently a portion of $1-\tau_1-\tau_2$ has equal magnitude on both tasks). Assuming $\lambda_b = \kappa\lambda_s$ for some $\kappa \in (1,2)$, we get

$$\begin{aligned}
\widetilde{\sigma}_1^2 &:= \operatorname{Var}\left(\tilde{Z}_j^{(1)}\right)\\
&= \frac{(1-\alpha)s\lambda_s^2+\tau_1\alpha s\lambda_s^2+\tau_2\alpha s(\lambda_b-\lambda_s)^2+(1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{(\sqrt{n}+\sqrt{s})^2}\\
&=: \frac{f_1(\kappa)s\lambda_s^2}{n\left(1+\sqrt{\frac{s}{n}}\right)^2}.
\end{aligned}$$

The first equality follows from the construction of the dual matrix and the fact that we have recovered the sign support correctly. The last strict inequality follows from the assumption that $\theta(n,p,s,\alpha) < 1$. Similarly, we have

$$\begin{aligned}
\widetilde{\sigma}_2^2 &:= \operatorname{Var}\left(\tilde{Z}_j^{(2)}\right)\\
&> \frac{(1-\alpha)s\lambda_s^2+\tau_2\alpha s\lambda_s^2+\tau_1\alpha s(\lambda_b-\lambda_s)^2+(1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{n\left(1+\sqrt{\frac{s}{n}}\right)^2}\\
&=: \frac{f_2(\kappa)s\lambda_s^2}{n\left(1+\sqrt{\frac{s}{n}}\right)^2}.
\end{aligned}$$

Given these lower bounds on the variance, by results on Sudakov minoration (see Theorem 3.15 in [22]), for any $\delta > 0$, there exists $N(\delta)$ such that if $p-(2-\alpha)s \geq N(\delta)$, with high probability we have

$$\begin{aligned}
&\max_{1\leq k\leq r}\max_{j\in\bigcup_{k=1}^{r}\mathcal{U}_k}\left|\tilde{Z}_j^{(k)}\right|\\
&\qquad\geq(1-\delta)\sqrt{(\widetilde{\sigma}_1^2+\widetilde{\sigma}_2^2)\log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}.
\end{aligned}$$

This in turn can be bound as

$$\begin{aligned}
&(1-\delta)\,(\widetilde{\sigma}_1^2+\widetilde{\sigma}_2^2)\log\Big(r\Big(p-(2-\alpha)s\Big)\Big)\\
&\qquad\geq(1-\delta)\frac{(f_1(\kappa)+f_2(\kappa))\ s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2.\\
&\qquad\geq(1-\delta)\frac{f(\kappa)\,s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2.
\end{aligned}$$

Consider two cases:

1) $\frac{s}{n} = \Omega(1)$: In this case, we have $s > cn$ for some constant $c > 0$. Then,

$$\begin{aligned}
&(1-\delta)\frac{(f(\kappa))\ s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\ \left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2\\
&\qquad=(1-\delta)\frac{(f(\kappa))\ (s/n)\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{\Big(1+\sqrt{s/n}\Big)^2}\lambda_s^2\\
&\qquad>c'f(\kappa)\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)\ \lambda_s^2\\
&\qquad>(1+\epsilon)\lambda_s^2,
\end{aligned}$$

for any fixed $\epsilon > 0$, as $p \to \infty$.

2) $\frac{s}{n} \to 0$: In this case, we have $s/n = o(1)$. Here we will use that the sample size scales as $n < (1-$

$\nu)\,(f(\kappa))\,s\log(p-(2-\alpha)s)$.

$$
\begin{aligned}
(1-\delta)&\frac{(f(\kappa))\ s\ \log\Big(r\Big(p-(2-\alpha)s\Big)\Big)}{n\ \left(1+\sqrt{\frac{s}{n}}\right)^2}\lambda_s^2 \\
&\geq \frac{(1-\delta)(1-o(1))}{1-\nu}\lambda_s^2 \\
&> (1+\epsilon)\lambda_s^2,
\end{aligned}
$$

for some $\epsilon > 0$ by taking δ small enough.

Thus with high probability, $\exists k \exists j \in \bigcap_{k=1}^r \mathcal{U}_k^c$ such that $\left|\tilde{Z}_j^{(k)}\right| > \lambda_s$. This is a contradiction to Lemma 4.

Case (ii): We need to show that with high probability, there exists a row that violates the sub-gradient condition of ℓ_∞-norm: $\exists j \in \bigcap_{k=1}^r \mathcal{U}_k^c$ such that $\left\|\tilde{Z}_j^{(k)}\right\|_1 > \lambda_b$. This is a contradiction to Lemma 4.

Following the same proof technique, notice that $\sum_{k=1}^r \tilde{Z}_j^{(k)}$ is a zero-mean Gaussian random variable with $\operatorname{Var}\left(\sum_{k=1}^r \tilde{Z}_j^{(k)}\right) \geq r(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2)$. Thus, with high probability

$$
\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \left\|\tilde{Z}_j^{(k)}\right\|_1 \geq (1-\delta)\sqrt{r(\widetilde{\sigma}_1^2+\widetilde{\sigma}_2^2)\log\Big(p-(2-\alpha)s\Big)}.
$$

Following the same line of argument for this case, yields the required bound $\left\|\tilde{Z}_j^{(k)}\right\|_1 > (1+\epsilon)\lambda_b$.

This concludes the proof of the theorem. ■

Lemma 11. *Under assumptions of Theorem 3, the conditions (C3) and (C4) in Lemma 5 hold with probability at least* $1 - c_1\exp(-c_2 n)$ *for some positive constants* c_1 *and* c_2.

Proof: First, we need to bound the projection of $\widetilde{Z}$ into the space U_s^c. Notice that

$$
\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty} \leq \max\left(\frac{\lambda_b - \lambda_s\|\tilde{s}_j\|_0}{\left|M_j(\tilde{B})\right| - \|\tilde{s}_j\|_0}, \left|(\tilde{z}_{\,\mathcal{U}^c})_j^{(k)}\right|\right).
$$

By our assumption on the penalty regularizer coefficients, we have $\frac{\lambda_b-\lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right|-\|\tilde{S}_j\|_0} < \lambda_s$. Moreover, we have

$$
\begin{aligned}
&\left|(\tilde{z}_{\,\mathcal{U}^c})_j^{(k)}\right| \\
&\leq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, \mathbf{I} - \frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}} \\
&\qquad + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)}, X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle \tilde{Z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{Z}_j^{(k)}} \\
&\triangleq \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|.
\end{aligned}
$$

By Lemma 9, if $n \geq \frac{2}{2-\sqrt{3}}\log(pr)$ then with high probability $\left\|X_j^{(k)}\right\|_2^2 \leq 2n$ and hence $\operatorname{Var}\left(\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2}{n}$. Notice that $\mathbb{E}\left[\left\|X_j^{(k)}\right\|_2^2\right] = n$ and we added the factor of 2 arbitrarily to use the concentration theorems. Using the concentration results for the zero-mean Gaussian random variable $\mathcal{W}_j^{(k)}$ (conditioned on X's) and using the union bound, for all $t > 0$, we get

$$
\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right| \geq t\right] \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2} + \log\left(p-(2-\alpha)s\right)\right).
$$

Conditioning on $\left(X_{\mathcal{U}_k}^{(k)}, w^{(k)}, \tilde{Z}^{(k)}\right)$'s, we have that $\mathcal{Z}_j^{(k)}$ is a zero-mean Gaussian random variable with

$$
\operatorname{Var}\left(\mathcal{Z}_j^{(k)}\right) \leq \frac{1}{n}\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right)\left\|\tilde{Z}_{\mathcal{U}_k}^{(k)}\right\|_2^2.
$$

According to the result of [20] on singular values of Gaussian matrices, for the matrix $X_{\mathcal{U}_k}^{(k)}$, for all $\delta > 0$, we have

$$
\mathbb{P}\left[\sigma_{min}\left(X_{\mathcal{U}_k}^{(k)}\right) \leq (1-\delta)\left(\sqrt{n}-\sqrt{s}\right)\right] \leq \exp\left(-\frac{\delta^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2}\right),
$$

and since $\lambda_{max}\left(\left(\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right) = \sigma_{min}\left(X_{\mathcal{U}_k}^{(k)}\right)^{-2}$, we get

$$
\begin{aligned}
\mathbb{P}&\left[\lambda_{max}\left(\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right) \geq \frac{(1+\delta)}{\left(1-\sqrt{\frac{s}{n}}\right)^2}\right] \\
&\leq \exp\left(-\frac{\left(\sqrt{\delta+1}-1\right)^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2(1+\delta)}\right).
\end{aligned}
$$

According to Lemma 10, if $\left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| = \mathcal{O}(\lambda_s)$, then with high probability $\tilde{S}_j = 0$, so that $|\tilde{\Theta}_j^{(1)}| = |\tilde{\Theta}_j^{(2)}|$. Thus, among shared features (with size αs), a fraction τ have differing magnitudes on $\tilde{\Theta}$. Let τ_1 be the fraction with larger magnitude on the first task and τ_2 the fraction with larger magnitude on the second task (so that $\tau = \tau_1 + \tau_2$). Then, with high probability, recalling that $\lambda_b = \kappa\lambda_s$ for some $1 < \kappa < 2$, we get

$$
\begin{aligned}
\operatorname{Var}\left(\mathcal{Z}_j^{(1)}\right) &\leq \frac{\left\|\tilde{Z}_{\mathcal{U}_1}^{(1)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{(1-\alpha)s\lambda_s^2 + \tau_1\alpha s\lambda_s^2 + \tau_2\alpha s(\lambda_b-\lambda_s)^2 + (1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{\left(1-(1-\tau_1-\tau_2)\alpha - 2\tau_2\alpha\kappa + \left(\tau_2 + \frac{1-\tau_1-\tau_2}{4}\right)\alpha\kappa^2\right)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&\triangleq \frac{f_1(\kappa)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\operatorname{Var}\left(\mathcal{Z}_j^{(2)}\right) &\leq \frac{\left\|\tilde{Z}_{\mathcal{U}_2}^{(2)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&= \frac{\left(1-(1-\tau_1-\tau_2)\alpha - 2\tau_1\alpha\kappa + \left(\tau_1 + \frac{1-\tau_1-\tau_2}{4}\right)\alpha\kappa^2\right)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2} \\
&\triangleq \frac{f_2(\kappa)s\lambda_s^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.
\end{aligned}
$$

By concentration of Gaussian random variables, we have

$$\begin{aligned}&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|\geq t\right]\\&\leq 2\exp\left(-\frac{t^2\left(\sqrt{n}-\sqrt{s}\right)^2}{2f_k(\kappa)s\lambda_s^2}+\log\left(p-(1-\alpha)s\right)\right)\qquad\forall t\geq 0.\end{aligned}$$

Using these bounds, we get

$$\begin{aligned}&\mathbb{P}\left[\left\|P_{U_s^c}(\widetilde{Z})\right\|_{\infty,\infty}<\lambda_s\right]\\&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|+\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\lambda_s\qquad\forall 1\leq k\leq r\right]\\&\geq\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{Z}_j^{(k)}\right|<t_0\quad\forall 1\leq k\leq r\right]\\&\qquad\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\left|\mathcal{W}_j^{(k)}\right|<\lambda_s-t_0\quad\forall 1\leq k\leq r\right]\\&\geq\left(1-2\exp\left(-\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_s^2}+\log\left(p-(2-\alpha)s\right)+\log(r)\right)\right)\\&\qquad\left(1-2\exp\left(-\frac{(\lambda_s-t_0)^2n}{4\sigma^2}+\log\left(p-(2-\alpha)s\right)+\log(r)\right)\right).\end{aligned}$$

This probability goes to 1 for

$$t_0=\frac{\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_s}{\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)ns}\lambda_s+2\sigma(\sqrt{n}-\sqrt{s})}\lambda_s$$

(the solution to $\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{(f_1(\kappa)+f_2(\kappa))s\lambda_s^2}=\frac{(\lambda_s-t_0)^2n}{4\sigma^2}$), if

$$\lambda_s>\frac{\sqrt{4\sigma^2\left(1-\sqrt{\frac{s}{n}}\right)^2\left(\log(r)+\log\left(p-(2-\alpha)s\right)\right)}}{\sqrt{n}-\left(\sqrt{s}+\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)s\left(\log(r)+\log\left(p-(2-\alpha)s\right)\right)}\right)}$$

provided that (substituting $r=2$),

$$\begin{aligned}n>&\left(f_1(\kappa)+f_2(\kappa)\right)s\log\left(p-(2-\alpha)s\right)\\&+\left(1+\left(f_1(\kappa)+f_2(\kappa)\right)\log(2)\right.\\&\qquad\left.+2\sqrt{\left(f_1(\kappa)+f_2(\kappa)\right)\left(\log(2)+\log\left(p-(2-\alpha)s\right)\right)}\right)s.\end{aligned}$$

Since $f_1(\kappa)+f_2(\kappa)=f(\kappa)$ by definition, for large enough p with $\frac{s}{p}=\mathbf{o}(1)$, we require

$$n>f(\kappa)s\log\left(p-(2-\alpha)s\right).\tag{7}$$

Next, we need to bound the projection of $\widetilde{Z}$ into the space U_b^c. Notice that

$$\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1}\leq\max\left(\lambda_s\|\tilde{s}_j\|_0,\sum_{k=1}^{r}\left|\left(\tilde{z}_{\ \mathcal{U}^c}\right)_j^{(k)}\right|\right).$$

We have $\lambda_s\|\tilde{S}_j\|_0\leq\lambda_s D(S^*)<\lambda_b$ by our assumption on the ratio of penalty regularizer coefficients. For all $j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c$, we have

$$\begin{aligned}&\sum_{k=1}^{r}\left|\tilde{Z}_j^{(k)}\right|\\&\leq\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)},\mathbf{I}-\frac{1}{n}X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)},X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\left(X_{\mathcal{U}_k}^{(k)}\right)^T\right\rangle w^{(k)}\right|}_{\mathcal{W}_j^{(k)}}\\&\qquad+\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\underbrace{\left|\frac{1}{n}\left\langle X_j^{(k)},X_{\mathcal{U}_k}^{(k)}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)},X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}\right\rangle\tilde{Z}_{\mathcal{U}_k}^{(k)}\right|}_{\mathcal{Z}_j^{(k)}}\\&=\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right|+\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|.\end{aligned}$$

Let $\mathbf{v}\in\{-1,+1\}^r$ be a vector of signs such that $\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|=\sum_{k=1}^{r}v_k\mathcal{W}_j^{(k)}$. Thus,

$$\text{Var}\left(\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|\right)=\text{Var}\left(\sum_{k=1}^{r}v_k\mathcal{W}_j^{(k)}\right)\leq\frac{2\sigma^2r}{n}.$$

Using the union bound and previous discussion, for all $t>0$, we get

$$\begin{aligned}&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{W}_j^{(k)}\right|\geq t\right]\\&\qquad=\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r}v_k\mathcal{W}_j^{(k)}\geq t\right]\\&\qquad\leq 2\exp\left(-\frac{t^2n}{4\sigma^2r}+r\log(2)+\log\left(p-(2-\alpha)s\right)\right).\end{aligned}$$

Also from the previous analysis, assuming $\lambda_b=\kappa\lambda_s$ for some $1<\kappa<2$, we get

$$\begin{aligned}&\text{Var}\left(\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right|\right)=\text{Var}\left(\sum_{k=1}^{r}v_k\mathcal{Z}_j^{(k)}\right)\leq\frac{\sum_{k=1}^{r}\left\|\tilde{Z}_j^{(k)}\right\|_2^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}\\&=\frac{2(1-\alpha)s\lambda_s^2+(\tau_1+\tau_2)\alpha s\lambda_s^2+(\tau_1+\tau_2)\alpha s(\lambda_b-\lambda_s)^2+2(1-\tau_1-\tau_2)\alpha s\frac{\lambda_b^2}{4}}{\left(\sqrt{n}-\sqrt{s}\right)^2}\\&=\frac{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2}{\left(\sqrt{n}-\sqrt{s}\right)^2}.\end{aligned}$$

and consequently for all $t>0$,

$$\begin{aligned}&\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right|\geq t\right]\\&=\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r}v_k\mathcal{Z}_j^{(k)}\geq t\right]\\&\leq 2\exp\left(-\frac{t^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2}+r\log(2)+\log\left(p-(2-\alpha)s\right)\right).\end{aligned}$$

Finally, we have

$$
\begin{aligned}
&\mathbb{P}\left[\left\|P_{U_b^c}(\widetilde{Z})\right\|_{\infty,1} < \lambda_b\right] \\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{Z}_j^{(k)}\right| + \max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{W}_j^{(k)}\right| < \lambda_b\right] \\
&\geq \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{Z}_j^{(k)}\right| < t_0\right] \\
&\qquad\qquad \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^r \mathcal{U}_k^c} \sum_{k=1}^r \left|\mathcal{W}_j^{(k)}\right| < \lambda_b - t_0\right] \\
&\geq \left(1 - 2\exp\left(-\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right)\right) \\
&\qquad\qquad \left(1-2\exp\left(-\frac{(\lambda_b-t_0)^2 n}{4\sigma^2 r} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right)\right).
\end{aligned}
$$

This probability goes to 1 for

$$
t_0 = \frac{\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)ns\lambda_b}}{\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)ns\lambda_b} + 2\sigma(\sqrt{n}-\sqrt{s})}\lambda_b
$$

(the solution to $\frac{(\lambda_b-t_0)^2 n}{4\sigma^2 r} = \frac{t_0^2(\sqrt{n}-\sqrt{s})^2}{\frac{1}{\kappa^2}(f_1(\kappa)+f_2(\kappa))s\lambda_b^2}$), if

$$
\lambda_b > \frac{\sqrt{4\sigma^2\left(1-\sqrt{\frac{s}{n}}\right)^2 r\left(r\log(2)+\log\left(p-(2-\alpha)s\right)\right)}}{\sqrt{n}-\left(\sqrt{s}+\sqrt{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)sr\left(r\log(2)+\log\left(p-(2-\alpha)s\right)\right)}\right)}
$$

provided that (substituting $r = 2$),

$$
\begin{aligned}
n &> \frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\log\left(p-(2-\alpha)s\right) \\
&\quad + \left(1 + \frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)2\log(2)\right. \\
&\quad \left. + 2\sqrt{\frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)\left(2\log(2)+\log\left(p-(2-\alpha)s\right)\right)}\right)s.
\end{aligned}
$$

For large enough p with $\frac{s}{p} = \mathbf{o}(1)$, we require

$$
n > \frac{2}{\kappa^2}f(\kappa)s\log\left(p-(2-\alpha)s\right).
$$

Combining this result with (7), the lemma follows. ■

References

[1] R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

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

[3] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. In *Allerton Conference, Allerton House, Illinois*, 2007.

[4] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *27th International Conference on Machine learning (ICML)*, 2010.

[5] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2009.

[6] R. Caruana. Multitask learning. *Machine Learning*, 28: 41–75, 1997.

[7] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[8] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B*, 71(5):1009–1030, 2009.

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

[10] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.

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

[12] C.Zhang and J.Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

[13] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *IEEE Transaction on Information Theory*, volume 57, pages 3841–3863, 2011.

[14] K. Lounici, A. B. Tsybakov, M. Pontil, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *22nd Conference On Learning Theory (COLT)*, 2009.

[15] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47, 2010.

[16] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[17] R. P.W. Duin. Department of Applied Physics, Delft University of Technology, Delft, The Netherlands, 2002.

[18] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, and J.E. den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998.

[19] X. He and P. Niyogi. Locality preserving projections. In *Neural Information Processing Systems (NIPS)*, 2003.

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

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

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

[23] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *26th International Conference on Machine Learning (ICML)*, 2009.

[24] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

Appendix A
Deterministic Necessary Optimality Conditions

In this appendix, we investigate deterministic necessary conditions for the optimality of the solutions $(\hat{B}, \hat{S})$ of the problem (2).

A. Sub-differential of ℓ_1/ℓ_∞ and ℓ_1/ℓ_1 Norms

In this section we state the sub-differential characterization of the norms we used in out convex program. The results can be directly derived from the definition of sub-differential of a function.

Lemma 12 (Sub-differential of ℓ_1/ℓ_∞-Norm)**.** *The matrix $\widetilde{Z} \in \mathbb{R}^{p\times r}$ belongs to the sub-differential of ℓ_1/ℓ_∞-norm of matrix $\widetilde{B}$, denoted as $\widetilde{Z} \in \partial \left\|\widetilde{B}\right\|_{1,\infty}$ iff*

(i) *for all $j \in$ RowSupp$(\widetilde{B})$, we have $\tilde{z}_j^{(k)} = \begin{cases} t_j^{(k)}\, \text{sign}\left(\tilde{b}_j^{(k)}\right) & k \in M_j(\widetilde{B}) \\ 0 & \text{ow.} \end{cases}$, where, $t_j^{(k)} \geq 0$ and $\sum_{k=1}^r t_j^{(k)} = 1$.*

(ii) *for all $j \notin$ RowSupp$(\widetilde{B})$, we have $\sum_{k=1}^r \left|\tilde{z}_j^{(k)}\right| \leq 1$.*

Lemma 13 (Sub-differential of ℓ_1/ℓ_1-Norm)**.** *The matrix $\widetilde{Z} \in \mathbb{R}^{p\times r}$ belongs to the sub-differential of ℓ_1/ℓ_1-norm of matrix $\widetilde{S}$, denoted as $\widetilde{Z} \in \partial \left\|\widetilde{S}\right\|_{1,1}$ iff*

(i) *for all $(j,k) \in$ Supp$(\widetilde{S})$, we have $\tilde{z}_j^{(k)} = \text{sign}\left(\tilde{s}_j^{(k)}\right)$.*

(ii) *for all $(j,k) \notin$ Supp$(\widetilde{S})$, we have $\left|\tilde{z}_j^{(k)}\right| \leq 1$.*

Appendix B
Proof of Lemma 2

We provide the proof of each property separately.

(P1) Suppose there exists $(j_0, k_0) \in$ Supp$(\hat{S})$, such that sign$(\hat{s}_j^{(k)}) = -$sign$(\hat{b}_j^{(k)})$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except at (j_0, k_0). Consider the following two cases

1) $\left|\hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)}\right| \leq \left\|\hat{b}_{j_0}\right\|_\infty$: Let $\check{b}_{j_0}^{(k_0)} = \hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}$ and $\check{s}_{j_0}^{(k_0)} = 0$. Notice that $(j_0, k_0) \notin$ Supp$(\check{S})$.
2) $\left|\hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)}\right| > \left\|\hat{b}_{j_0}\right\|_\infty$: Let $\check{b}_{j_0}^{(k_0)} = -\text{sign}\left(\hat{b}_{j_0}^{(k_0)}\right)\left\|\hat{b}_{j_0}\right\|_\infty$ and $\check{s}_{j_0}^{(k_0)} = \hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)} - \check{b}_{j_0}^{(k_0)}$. Notice that sign$\left(\check{b}_{j_0}^{(k_0)}\right) =$ sign$\left(\check{s}_{j_0}^{(k_0)}\right)$.

Since $\check{B} + \check{S} = \hat{B} + \hat{S}$ and $\|\check{b}_{j_0}\|_\infty \leq \|\hat{b}_{j_0}\|_\infty$ and $\|\check{s}_{j_0}\|_1 < \|\hat{s}_{j_0}\|_1$, it is a contradiction to the optimality of $(\hat{B}, \hat{S})$.

(P2) We prove the result in two steps by establishing 1. $M(\hat{B}) > \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$ and 2. $D(\hat{S}) < \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$.

1) On the contrary, suppose there exists a row $j_0 \in$ RowSupp$(\hat{B})$ such that $\left|M_{j_0}(\hat{B})\right| \leq \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$. Let k^* be the index of the element whose magnitude is ranked $\left(\left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor + 1\right)$ among the element of the vector $\hat{b}_{j_0} + \hat{s}_{j_0}$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except on the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \text{sign}\left(\hat{b}_{j_0}^{(k)}\right) & \\ & \left|\hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)}\right| \geq \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \\ \hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)} & \text{ow,} \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Notice that $M(\check{B}) > \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$ and sign$\left(\check{s}_{j_0}^{(k)}\right) =$ sign$\left(\check{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in$ Supp$\left(\check{s}_{j_0}\right)$ since sign$\left(\hat{s}_{j_0}^{(k)}\right) =$ sign$\left(\hat{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in$ Supp$\left(\hat{S}_{j_0}\right)$ by (P1). Further, since $\check{S} + \check{B} = \hat{S} + \hat{B}$ and $\|\check{b}_{j_0}\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right| + \left|\hat{s}_{j_0}^{(k^*)}\right|$ and $\|\check{s}_{j_0}\|_1 \leq \|\hat{s}_{j_0}\|_1 + \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor \left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\check{b}_{j_0}^{(k^*)}\right| - \left|\check{s}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$ due to the fact that $\lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

2) On the contrary, suppose there exists a row $j_0 \in$ RowSupp$(\hat{S})$ such that $\|\hat{s}_{j_0}\|_0 \geq \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$. Let k^* be the index of the element whose magnitude is ranked $\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$ among the elements of the vector $\hat{b}_{j_0} + \hat{s}_{j_0}$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices respectively equal to $\hat{B}$ and $\hat{S}$ in all entries except on the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \text{sign}\left(\hat{b}_{j_0}^{(k)}\right) & \\ & \left|\hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)}\right| \geq \left|\hat{b}_{j_0}^{(k^*)} + \hat{s}_{j_0}^{(k^*)}\right| \\ \hat{b}_{j_0}^{(k)} + \hat{s}_{j_0}^{(k)} & \text{ow,} \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Notice that $D(\check{S}) < \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$ and sign$\left(\check{s}_{j_0}^{(k)}\right) =$ sign$\left(\check{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in$ Supp$\left(\check{s}_{j_0}\right)$ since sign$\left(\hat{s}_{j_0}^{(k)}\right) =$ sign$\left(\hat{b}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in$ Supp$\left(\hat{s}_{j_0}\right)$. Since $\check{S} + \check{B} = \hat{S} + \hat{B}$ and $\|\check{b}_{j_0}\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right| + \left|\hat{s}_{j_0}^{(k^*)}\right|$ and $\|\check{s}_{j_0}\|_1 \leq \|\hat{s}_{j_0}\|_1 + \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right)\left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\check{b}_{j_0}^{(k^*)}\right| - \left|\check{s}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$, due to the fact that $\lambda_s\left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right) < \lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

(P3) If $j \notin$ RowSupp$(\hat{B})$ then the result is trivial. Suppose there exists $(j_0, k_0) \in$ Supp$(\hat{S})$ with $j_0 \in$ RowSupp$(\hat{S})$ such that $\left|b_{j_0}^{(k_0)}\right| < \|\hat{b}_{j_0}\|_\infty$. Let $\check{B}, \check{S} \in \mathbb{R}^{p\times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except for the entry corresponding to the index (j_0, k_0). Let $\check{b}_{j_0}^{(k_0)} = \left\|\hat{b}_{j_0}\right\|_\infty \text{sign}\left(\hat{b}_{j_0}^{(k_0)}\right)$ if $\left|\hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}\right| \geq \|b_{j_0}\|_\infty$ and $\check{b}_{j_0}^{(k_0)} = \hat{b}_{j_0}^{(k_0)} + \hat{s}_{j_0}^{(k_0)}$

otherwise. Let $\check{s}_{j_0}^{(k_0)} = \hat{s}_{j_0}^{(k_0)} + \hat{b}_{j_0}^{(k_0)} - \check{b}_{j_0}^{(k_0)}$. Since $\check{B} + \check{S} = \hat{B} + \hat{S}$ and $\|\check{b}_{j_0}\|_\infty = \left\|\hat{b}_{j_0}\right\|_\infty$ and $\|\check{s}_{j_0}\|_1 < \|\hat{s}_{j_0}\|_1$, it is a contradiction to the optimality of $(\hat{B}, \hat{S})$.

(P4) If $j \notin \text{RowSupp}(\hat{B})$ or $j \notin \text{RowSupp}(\hat{S})$ the result is trivial. Suppose there exists a row $j_0 \in \text{RowSupp}(\hat{B}) \cap \text{RowSupp}(\hat{S})$ such that the result does not hold for that. Let $k^* = \arg\max_{\{k:(j,k) \notin \text{Supp}(\hat{S})\}} \left|\hat{b}_j^{(k)}\right|$. Let $\check{B}, \check{S} \in \mathbb{R}^{p \times r}$ be matrices equal to $\hat{B}, \hat{S}$ in all entries except for the row j_0 and

$$\hat{b}_{j_0}^{(k)} = \begin{cases} \left|\hat{b}_{j_0}^{(k^*)}\right| \text{sign}\left(\hat{b}_{j_0}^{(k)}\right) & (j_0, k) \in \text{Supp}(\hat{S}) \\ \hat{b}_{j_0}^{(k)} & \text{ow}, \end{cases}$$

and $\check{s}_{j_0} = \hat{s}_{j_0} + \hat{b}_{j_0} - \check{b}_{j_0}$. Since $\check{B} + \check{S} = \hat{S} + \hat{B}$ and $\|\check{b}_{j_0}\|_\infty = \left|\hat{b}_{j_0}^{(k^*)}\right|$ and by (P2) and (P3), $\|\check{s}_{j_0}\|_1 \leq \|\hat{s}_{j_0}\|_1 + \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right) \left(\left\|\hat{b}_{j_0}\right\|_\infty - \left|\hat{b}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$, due to the fact that $\lambda_s \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right) < \lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

This concludes the proof of the lemma.

Appendix C
Proof of Lemma 3

On the contrary, assume that $(\hat{S}, \hat{B})$ is the unique solution. Take a non-zero row $\hat{b}_{j_0}$ with $j_0 \in \text{RowSupp}(\hat{B})$. If $\left|M_{j_0}(\hat{B})\right| < d$, then let $\check{B}, \check{S} \in \mathbb{R}^{p \times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except on the row j_0 and let $\check{b}_{j_0} = \mathbf{0}$ and $\check{s}_{j_0} = \hat{b}_{j_0} + \hat{s}_{j_0}$. Then, $(\check{B}, \check{S})$ are *strictly* better solutions than $(\hat{B}, \hat{S})$. This contradicts the optimality of $(\hat{B}, \hat{S})$. Hence, $\left|M_{j_0}(\hat{B})\right| \geq d$. with similar argument we can conclude that $\left\|\hat{S}_{j_0}\right\|_0 \leq d$.

If $\left\|\hat{S}_{j_0}\right\|_0 = d$, then let $0 < \delta \leq \min_{(j_0,k) \in \text{Supp}(\hat{S})} \left|\hat{s}_{j_0}^{(k)}\right|$ and $\check{B}(\delta), \check{S}(\delta) \in \mathbb{R}^{p \times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except for the entries indexed $(j_0, k) \in \text{Supp}(\hat{S})$ and let $\check{b}_{j_0}^{(k)} = \hat{b}_{j_0}^{(k)} + \delta \text{sign}\left(\hat{b}_{j_0}^{(k)}\right)$ and $\check{s}_{j_0}^{(k)} = \hat{s}_{j_0}^{(k)} - \delta \text{sign}\left(\hat{s}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}(\hat{S})$. Then, $(\check{B}(\delta), \check{S}(\delta))$ is another solution to (2). This contradicts the uniqueness of $(\hat{B}, \hat{S})$.

If $\left\|\hat{S}_{j_0}\right\|_0 < d$, then using Lemma 2 and Equation 6, we have

$$\begin{aligned}
&\mathbb{P}\left[\left|M_{j_0}(\hat{B})\right| \geq d+1\right] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\left[\left|M_{j_0}(\hat{B})\right| = d+i\right] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l = 1, \ldots, i+1 : \\
&\qquad\qquad \|\hat{b}_{j_0}^{(k_l)} + \underbrace{\hat{s}_{j_0}^{(k_l)}}_{0}| = \left\|\hat{b}_{j_0}\right\|_\infty\Bigg] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l = 1, \ldots, i+1 : \\
&\qquad\qquad \left|\Delta_{j_0}^{(k_l)}\right| = \left|b_j^{*(k_l)} + s_j^{*(k_l)}\right| + \left\|\hat{b}_j\right\|_\infty\Bigg] \\
&= \sum_{i=1}^{r-d} \mathbb{P}\Bigg[\exists k_1, \ldots, k_{i+1} \in M_{j_0}(\hat{B}) \quad \forall l, m = 1, \ldots, i+1 : \\
&\qquad\qquad \left|\Delta_{j_0}^{(k_l)}\right| = C_{k_l,k_m} + \left|\Delta_{j_0}^{(k_m)}\right|\Bigg] = 0.
\end{aligned}$$

In above equation C_{k_l,k_m} are some constants. The last conclusion follows from the fact that $\Delta_{j_0}^{(k_l)}$'s are continuous Gaussian variables and the cardinality of this event is less than the cardinality of the space they lie in. Hence, $\left|M_{j_0}(\hat{B})\right| = d$.

Let $0 < \delta < \|b_{j_0}\|_\infty$ and $\check{B}(\delta), \check{S}(\delta) \in \mathbb{R}^{p \times r}$ be two matrices equal to $\hat{B}, \hat{S}$ except for the entries indexed (j_0, k) for $k \in M_{j_0}(\hat{B})$ and let $\check{b}_{j_0}^{(k)} = \hat{b}_{j_0}^{(k)} - \delta$ and $\check{s}_{j_0}^{(k)} = \hat{s}_{j_0}^{(k)} + \delta$ for all $k \in M_{j_0}(\hat{B})$. Then, $(\check{B}(\delta), \check{S}(\delta))$ is another solution to (2). This contradicts the uniqueness of $(\hat{B}, \hat{S})$.

Appendix D
Coordinate Descent Algorithm

We use the coordinate descendent algorithm described as follows. The algorithm takes the tuple $(X, Y, \lambda_s, \lambda_b, \epsilon, B, S)$ as input, and outputs $(\hat{B}, \hat{S})$. Note that X and Y are given to this algorithm, while B and S are our initial guess or the warm start of the regression matrices. ϵ is the precision parameter which determines the stopping criterion.

We update elements of the sparse matrix S using the subroutine $UpdateS$, and update elements in the block sparse matrix B using the subroutine $UpdateB$, respectively, until the regression matrices converge. The pseudo code is in Algorithm 1 to Algorithm 3.

A. Correctness of Algorithms

In this algorithm, B is the block sparse matrix and S is the sparse matrix. We alternatively update B and S until they converge. When updating S, we cycle through each element of S while holding all the other elements of S and B unchanged; When updating B, we update each block B_j (the coefficient vector of the j^{th} feature for r tasks) as a whole, while keeping S and other coefficient vector of B fixed.

Algorithm 2 Our Model Solver

```
Input: $X$, $Y$, $\lambda_b$, $\lambda_s$, $B$, $S$ and $\varepsilon$
Output: $\hat{S}$ and $\hat{B}$

  Initialization:
  for $j = 1 : p$ do
    for $k = 1 : r$ do
      $c_j^{(k)} \leftarrow \left\langle X_j^{(k)}, y^{(k)} \right\rangle$
      for $i = 1 : p$ do
        $d_{i,j}^{(k)} \leftarrow \left\langle X_i^{(k)}, X_j^{(k)} \right\rangle$
      end for
    end for
  end for

  Updating:
  loop
    $S \leftarrow UpdateS(c; d; \lambda_s; B; S)$
    $B \leftarrow UpdateB(c; d; \lambda_b; B; S)$
    if Relative Update $< \epsilon$ then
      BREAK
    end if
  end loop
  RETURN $\hat{B} = B,\ \hat{S} = S$
```

Algorithm 3 UpdateB

```
Input: c, d, $\lambda_b$, $B$ and $S$
Output: $B$
  Update $B$ using the cyclic coordinate descent algorithm for
  $\ell_1/\ell_\infty$ while keeping $S$ unchanged.

  for $j = 1 : p$ do
    for $k = 1 : r$ do
      $\alpha_j^{(k)} \leftarrow \left( c_j^{(k)} - \sum_{i \neq j} (b_i^{(k)} + s_i^{(k)}) d_{i,j}^{(k)} \right) / d_{j,j}^{(k)} - s_i^{(k)}$
      if $\sum_{k=1}^{r} |\alpha_j^{(k)}| \leq \lambda_b$ then
        $b_j \leftarrow 0$
      else
        Sort $\alpha$ to be $|\alpha_j^{(k_1)}| \geq |\alpha_j^{(k_2)}| \geq \cdots \geq |\alpha_j^{(k_r)}|$
        $m^* = \arg\max_{1 \leq m \leq r} (\sum_{k=1}^{r} |\alpha_j^{(k_m)}| - \lambda_b)/m$
        for $i = 1 : r$ do
          if $i > m^*$ then
            $b_j^{(k_i)} \leftarrow \alpha_j^{(k_i)}$
          else
            $b_j^{(k_i)} \leftarrow \frac{\text{sign}(\alpha_j^{(k_i)})}{m^*} \left( \sum_{l=1}^{m^*} |\alpha_j^{(k_l)}| - \lambda_b \right)$
          end if
        end for
      end if
    end for
  end for
  RETURN $B$
```

Algorithm 4 Update-S

```
Input: c, d, $\lambda_s$, $B$ and $S$
Output: $S$
  Update $S$ using the cyclic coordinate descent algorithm for
  LASSO while keeping $B$ unchanged.
  for $j = 1 : p$ do
    for $k = 1 : r$ do
      $\alpha_j^{(k)} \leftarrow \left( c_j^{(k)} - \sum_{i \neq j} (b_i^{(k)} + s_i^{(k)}) d_{i,j}^{(k)} \right) / d_{j,j}^{(k)} - s_i^{(k)}$
      if $|\alpha_j^{(k)}| \leq \lambda_s$ then
        $s_j^k \leftarrow 0$
      else
        $s_j^k \leftarrow \alpha_j^{(k)} - \lambda_s \text{sign}(\alpha_j^{(k)})$
      end if
    end for
  end for
  RETURN $S$
```

For updating B, the subproblem is updating B_j

$$\hat{b}_j = \arg\min_{b_j} \qquad \frac{1}{2}\sum_{k=1}^{r} \left\| r_j^{(k)} - b_j^{(k)} X_j^{(k)} \right\|_2^2 + \lambda_b \|b_j\|_\infty . \quad (8)$$

If we take the partial residual vector $r_j^{(k)} = y^{(k)} - \sum\limits_{l \neq j} (b_l^{(k)} X_l^{(k)}) - \sum_l (s_l^{(k)} X_l^{(k)})$, the correctness of this algorithm will directly follow from the correctness of coordinate descent algorithm of ℓ_1/ℓ_∞ in [23, 24]. With the same argument, the correctness of the Algorithm 3 can be proven.