A Dirty Model for Multi-task Learning (Appendices)

Ali Jalali
University of Texas at Austin
alij@mail.utexas.edu

Pradeep Ravikumar
University of Texas at Asutin
pradeepr@cs.utexas.edu

Sujay Sanghavi
University of Texas at Austin
sanghavi@mail.utexas.edu

Chao Ruan
University of Texas at Austin
ruan@cs.utexas.edu

A Outline of Appendix

The proofs of our three main theorems are in appendices E, F and G respectively. We collate the machinery needed to prove these in earlier appendices. In the first appendix B, we restate our method for convenience and set up some notation. In the second appendix C, we state the optimality conditions characterizing the solution of the proposed optimization problem. The next appendix D is an important section of the appendix: it sets out our resume for proving our theorems via a primal-dual certificate construction. Such certificate proof techniques have been used in many such high-dimensional analyses [3, 5]. We have to provide a slightly more delicate construction because of the two interplaying parameter components in our optimization problem. Finally, in appendices E-G, we prove theorems 1-3 respectively by showing that the primal-dual witness construction succeeds under the conditions of the respective theorems. Appendix H describe our experimental setup and results in detail. Finally, Appendix I details the coordinate descent algorithm used for solving the dirty model optimization.

B Definitions and Setup

We now introduce the terms and notation we use throughout the proofs.

Notation. 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 $Q \in \mathbb{R}^{p\times r}$, the norm ℓ_ζ/ℓ_ρ is denoted as $\|Q\|_{\rho,\zeta} = \|\left(\|Q_1\|_\zeta, \cdots, \|Q_p\|_\zeta\right)\|_\rho$. 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 indecies $\mathcal{U} \subseteq \{1, \cdots, p\}$, the matrix $X_\mathcal{U} \in \mathbb{R}^{n\times|\mathcal{U}|}$ represents the sub-matrix of X consisting of X_j's where $j \in \mathcal{U}$.

Setup. The multiple regression problem is given as:

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

The optimization problem solved by our estimator:

$$(\hat{S}, \hat{B}) \in \arg\min_{S,B} \quad \frac{1}{2n}\sum_{k=1}^{K}\left\|y^{(k)} - X^{(k)}\left(S^{(k)} + B^{(k)}\right)\right\|_2^2 + \lambda_s\|S\|_{1,1} + \lambda_b\|B\|_{1,\infty}. \tag{2}$$

B.1 Splits and Transforms

We first define a d-split of a matrix for any $d \in \mathbb{N}$:

Definition 1. *A pair of matrices (B^*, S^*) is said to be a d-split of any matrix Θ^* if $\Theta^* = B^* + S^*$, and*

$$B_j^* = \begin{cases} \theta_j^* & \text{if } \|\theta_j^*\|_0 > d, \\ 0 & \text{otherwise}. \end{cases}$$

Thus, B contains those rows with greater than d elements, while S contains those with less than or equal to d elements; and are thus row-disjoint: for each row j we have $\|S_j^*\|_0 \|B_j^*\|_0 = 0$.

Remarks. One would expect that the solution $(\hat{S}, \hat{B})$ to the optimization problem (2) would be a d-split, for some $d \in \mathbb{N}$ of the true parameter Θ^*. It turns out however, that the solution is actually a *transformation* of such a d-split.

We define this *clipping transform* as follows.

Definition 2. *Given two matrices $B^*, S^* \in \mathbb{R}^{p \times r}$ with disjoint row support and a scalar $d \in \mathbb{N}$ with $\min_{j \in \text{RowSupp}(B)} \|B_j^*\|_0 \geq d+1$ and $d \geq \max_j \|S_j^*\|_0$, we define the clipping transform $(\bar{B}, \bar{S}) = \mathcal{H}_d(B^*, S^*)$ as:*

- *For $j \notin \text{RowSupp}(B^*)$, Set $\bar{B}_j = \mathbf{0}$.*
- *For each $j \in \text{RowSupp}(B^*)$, sort the largest magnitude $d+1$ non-zero entries as $\left|B_j^{*(k_1)}\right| \geq \left|B_j^{*(k_2)}\right| \geq \cdots \geq \left|B_j^{*(k_{d+1})}\right| > 0$ and let*

$$\bar{B}_j^{(k)} = \begin{cases} \left|B_j^{*(k_{d+1})}\right| \text{sign}\left(B_j^{*(k)}\right) & \left|B_j^{*(k)}\right| \geq \left|B_j^{*(k_{d+1})}\right| \\ B_j^{*(k)} & \text{ow}. \end{cases}$$

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

Remarks.

1. The sum is maintained through the transformation, so that if $(\bar{B}, \bar{S}) = \mathcal{H}_d(B^*, S^*)$ then $\bar{B} + \bar{S} = B^* + S^*$.

2. If $(\hat{B}, \hat{S}) = \mathcal{H}_d(\widetilde{B}, \widetilde{S})$, then we can recover the arguments $(\widetilde{B}, \widetilde{S})$ from $(\hat{B}, \hat{S})$ using:

$$\widetilde{B}_j = \begin{cases} \hat{B}_j + \hat{S}_j & j \in \text{RowSupp}(\hat{B}) \\ \hat{B}_j & \text{ow} \end{cases},$$

and then using $\widetilde{S} = \hat{B} + \hat{S} - \widetilde{B}$.

We denote this mapping by $\mathcal{H}_d^{-1}$ and say $(\widetilde{B}, \widetilde{S}) = \mathcal{H}_d^{-1}(\hat{B}, \hat{S})$. Note that this map $\mathcal{H}_d^{-1}$ is independent of the parameter d.

B.2 Sparse Matrix Setup

Define $\text{Supp}(S) = \{(j,k) : S_j^{(k)} \neq 0\}$, and let $U_s(\bar{S}) = \{S \in \mathbb{R}^{p \times r} : \text{Supp}(S) \subseteq \text{Supp}(\bar{S})\}$ be the subspace of matrices whose their support is the subset of the matrix $\bar{S}$. To shorten the notation, we use U_s instead of $U_s(\bar{S})$. The orthogonal projection to the subspace U_s can be defined as follows:

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

We can define the orthogonal complement space of U_s to be $U_s^c(\bar{S}) = \{S \in \mathbb{R}^{p \times r} : \text{Supp}(S) \cap \text{Supp}(\bar{S}) = \phi\}$. The orthogonal projection to this space can be defined as $P_{U_s^c}(Q) = Q - P_{U_s}(Q)$. We define $D(S) = \max_{1 \leq j \leq p} \|S_j\|_0$ denoting the maximum number of non-zero elements in any row of the sparse matrix S.

B.3 Row-Sparse Matrix Setup

Define RowSupp$(B) = \{j : \exists k \quad \text{s.t.} \quad B_j^{(k)} \neq 0\}$, and let $U_b(\bar{B}) = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \subseteq \text{RowSupp}(\bar{B})\}$ be the subspace of matrices whose row support is the subset of the row support of the matrix $\bar{B}$. To shorten the notation, we use U_b instead of $U_b(\bar{B})$. The orthogonal projection to the subspace U_b can be defined as follows:

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

We can define the orthogonal complement space of U_b to be $U_b^c(\bar{B}) = \{B \in \mathbb{R}^{p\times r} : \text{RowSupp}(B) \cap \text{RowSupp}(\bar{B}) = \phi\}$. The orthogonal projection to this space can be defined as $P_{U_b^c}(Q) = Q - P_{U_b}(Q)$.

For a given matrix $B \in \mathbb{R}^{p\times r}$, let $M_j^+(B) = \left\{k : B_j^{(k)} = \|B_j\|_\infty > 0\right\}$ and $M_j^-(B) = \left\{k : -B_j^{(k)} = \|B_j\|_\infty > 0\right\}$ be the set of indecies that the corresponding elements achieve the maximum magnitude on the j^{th} row with positive and negative signs, respectively. To shorten the notation, let $M_j^\pm(B) = M_j^+(B) \cup M_j^-(B)$. Also, let $M(B) = \min_{1\leq j\leq p} |M_j^\pm(B)|$ be the minimum number of elements who achieve the maximum in each row of the matrix B.

With this notation, we can now state some simple properties of the clipping transform:

Lemma 1. *If $(\bar{B}, \bar{S}) = \mathcal{H}_d(B^*, S^*)$ then the following properties hold*

(P1) $M(\bar{B}) \geq d+1$ *and* $D(\bar{S}) \leq d$.

(P2) $\text{sign}\left(\bar{S}_j^{(k)}\right) = \text{sign}\left(\bar{B}_j^{(k)}\right) = \text{sign}\left(B_j^{*(k)}\right)$ *for all* $j \in \text{RowSupp}(B^*)$ *and* $k \in M_j^\pm(B^*)$.

(P3) $\bar{S}_j^{(k)} = 0$ *for all* $j \in \text{RowSupp}(B^*)$ *and* $k \notin M_j^\pm(B^*)$.

Proof. The proof directly follows from the construction of $\bar{B}$ and $\bar{S}$. □

B.4 Sub-differential of ℓ_1/ℓ_∞ and ℓ_1/ℓ_1 Norms

In this section we detail the form of the sub-differentials of the regularization norms used in the convex program (2).

Lemma 2 (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 $j \in \text{RowSupp}(\widetilde{B})$, *we have* $\tilde{Z}_j^{(k)} = \begin{cases} t_j^{(k)}\, \text{sign}\left(\tilde{B}_j^{(k)}\right) & k \in M_j^\pm(\widetilde{B}) \\ 0 & \text{ow.} \end{cases}$, *where,* $t_j^{(k)} \geq 0$ *and* $\sum_{k=1}^r t_j^{(k)} = 1$.

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

Lemma 3 (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 $(j,k) \in \text{Supp}(\widetilde{S})$, *we have* $\tilde{Z}_j^{(k)} = \text{sign}\left(\tilde{S}_j^{(k)}\right)$.

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

Throughout the proof we use four pairs of matrices:

(B^*, S^*) : The d-split of the true parameter matrix Θ for a fixed integer d.

$(\bar{B}, \bar{S})$: The clipped-transform of the d-split; $(\bar{B}, \bar{S}) = \mathcal{H}_d(B^*, S^*)$.

$(\hat{B}, \hat{S})$: The solution to the original convex optimization problem 2.

$(\tilde{B}, \tilde{S})$: The solution to an oracle convex optimization problem 4.

Our proof outline in a nutshell: We first show that the solution of the oracle problem and the original problem are the same, that is $(\hat{B}, \hat{S}) = (\tilde{B}, \tilde{S})$. Then, we conclude that $\text{Supp}(\hat{S}) = \text{Supp}(\bar{S})$ and $\text{Supp}(\hat{B}) = \text{Supp}(\bar{B})$. Finally, since $\bar{B} + \bar{S} = B^* + S^*$, we have $\text{Supp}(\hat{B} + \hat{S}) = \text{Supp}(\Theta^*)$.

C Optimality Conditions

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

The first lemma states necessary conditions for any solution of the problem (2).

Lemma 4. *If $(\hat{S}, \hat{B})$ is a solution (uniqueness is NOT required) of (2) then the following properties hold*

(P1) $\text{sign}(\hat{S}_j^{(k)}) = \text{sign}(\hat{B}_j^{(k)})$ *for all* $(j,k) \in \text{Supp}(\hat{S})$ *with* $j \in \text{RowSupp}(\hat{B})$.

(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* $(j,k) \in \text{Supp}(\hat{S})$.

(P4) if $\frac{\lambda_b}{\lambda_s}$ *is not an integer,* $\forall j\, \exists k$ *such that* $(j,k) \notin \text{Supp}(\hat{S})$ *and* $\left|\hat{B}_j^{(k)}\right| = \left\|\hat{B}_j\right\|_\infty$.

(P5) if $d = \lfloor \frac{\lambda_b}{\lambda_s} \rfloor < \frac{\lambda_b}{\lambda_s}$ *then* $(\hat{B}, \hat{S}) = \mathcal{H}_d(\hat{B}, \hat{S})$.

Remarks. This lemma motivated the definition of the clipping transform $\mathcal{H}_d$. Note that (P5) states that the solution of the optimization problem (2) has the form of the output of the clipping transform.

Proof. We provide the proof of each property separately.

(P1) Suppose there exists $(j_0, k_0) \in \text{Supp}(\hat{S})$, such that $\text{sign}(\hat{S}_j^{(k)}) = -\text{sign}(\hat{B}_j^{(k)})$. 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 $(j_0, k_0) \notin \text{Supp}(\check{S})$.
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 $\text{sign}\left(\check{B}_{j_0}^{(k_0)}\right) = \text{sign}\left(\check{S}_{j_0}^{(k_0)}\right)$.

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. In contrary, suppose there exists a row $j_0 \in \text{RowSupp}(\hat{B})$ such that $\left|M_{j_0}^{\pm}(\hat{B})\right| \leq \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor$. 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 $\text{sign}\left(\check{S}_{j_0}^{(k)}\right) = \text{sign}\left(\check{B}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\check{S}_{j_0}\right)$ since $\text{sign}\left(\hat{S}_{j_0}^{(k)}\right) = \text{sign}\left(\hat{B}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\hat{S}_{j_0}\right)$ 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. In contrary, suppose there exists a row $j_0 \in \text{RowSupp}(\hat{S})$ such that $\left\|\hat{S}_{j_0}\right\|_0 \geq \left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil$. 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 $\text{sign}\left(\check{S}_{j_0}^{(k)}\right) = \text{sign}\left(\check{B}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\check{S}_{j_0}\right)$ since $\text{sign}\left(\hat{S}_{j_0}^{(k)}\right) = \text{sign}\left(\hat{B}_{j_0}^{(k)}\right)$ for all $(j_0, k) \in \text{Supp}\left(\hat{S}_{j_0}\right)$. 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 $j \notin \text{RowSupp}(\hat{B})$ then the result is trivial. Suppose there exists $(j_0, k_0) \in \text{Supp}(\hat{S})$ with $j_0 \in \text{RowSupp}(\hat{S})$ 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 $\left\|\check{B}_{j_0}\right\|_\infty = \left\|\hat{B}_{j_0}\right\|_\infty$ and $\left\|\check{S}_{j_0}\right\|_1 < \left\|\hat{S}_{j_0}\right\|_1$, 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 $\left\|\check{B}_{j_0}\right\|_\infty = \left|\hat{B}_{j_0}^{(k^*)}\right|$ and by (P2) and (P3), $\left\|\check{S}_{j_0}\right\|_1 \leq \left\|\hat{S}_{j_0}\right\|_1 + \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right)\left(\left\|\hat{B}_{j_0}\right\|_\infty - \left|\hat{B}_{j_0}^{(k^*)}\right|\right)$, this is a contradiction to the optimality of $(\hat{B}, \hat{S})$, due to the fact that $\lambda_s \left(\left\lceil \frac{\lambda_b}{\lambda_s} \right\rceil - 1\right) < \lambda_s \left\lfloor \frac{\lambda_b}{\lambda_s} \right\rfloor < \lambda_b$.

(P5) This result follows from the definition of $\mathcal{H}_d$ and (P1)-(P4).

This concludes the proof of the lemma.

□

Lemma 5 (Convex Optimality)**.** *$((\hat{B},\hat{S}),\hat{Z})$ is an optimal primal-dual solution pair of* (2) *if it satisfies:*

1. **(Stationary Condition).**
$$\frac{1}{n}\left\langle X^{(k)}, X^{(k)}\right\rangle\left(\hat{S}^{(k)}+\hat{B}^{(k)}\right)-\frac{1}{n}(X^{(k)})^T y^{(k)}+\hat{Z}^{(k)}=0. \tag{3}$$

2. **(Dual Feasibility).** *$\hat{Z}\in\mathbb{R}^{p\times r}$ satisfies: $\hat{Z}\in\lambda_s\partial\|\hat{S}\|_{1,1}$ and $\hat{Z}\in\lambda_b\partial\|\hat{B}\|_{1,\infty}$ and for all $k=1,\ldots,r$.*

Proof. The result follows from the standard optimality condition of convex programs.

□

D Primal-Dual Construction

In the light of Lemma 4, our goal is then to recover the clipped transform $(\bar{B},\bar{S})=\mathcal{H}_d(B^*,S^*)$ in our regression model, for some $d\in\mathbb{N}$. Accordingly, we construct the primal-dual pair $((\tilde{S},\tilde{B}),\widetilde{Z})$ as follows:

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

2. Let $\widetilde{Z}_{\bigcup_{k=1}^r\mathcal{U}_k}=\left(\widetilde{Z}_s\right)_{\bigcup_{k=1}^r\mathcal{U}_k}+\left(\widetilde{Z}_b\right)_{\bigcup_{k=1}^r\mathcal{U}_k}$, where, $\widetilde{Z}_s=\lambda_s\text{sign}(\tilde{S})$, and for all $j\in\bigcup_{k=1}^r\mathcal{U}_k$,
$$(\tilde{Z}_b)_j^{(k)}=\begin{cases}\dfrac{\lambda_b-\lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right|-\|\tilde{S}_j\|_0}\text{sign}\left(\tilde{B}_j^{(k)}\right) & k\in M_j^{\pm}(\tilde{B})\quad\&\quad(j,k)\notin\text{Supp}(\tilde{S})\\ 0 & \text{ow}\end{cases}.$$

3. Set $\tilde{Z}^{(k)}_{\bigcap_{k=1}^r\mathcal{U}_k^c}$ from the stationary condition (3).

Remarks. Note that by construction $((\tilde{S},\tilde{B}),\widetilde{Z})$ satisfy:

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

(C2) $P_{U_b(\bar{B})}(\widetilde{Z})=\begin{cases}t_j^{(k)}\,\text{sign}\left(\tilde{B}_j^{(k)}\right) & k\in M_j^{\pm}(\bar{B})\\ 0 & \text{ow.}\end{cases}$, where, $t_j^{(k)}\geq 0$ such that $\sum_{k\in M_j^{\pm}(\bar{B})}t_j^{(k)}=\lambda_b$.

(C3) $\frac{1}{n}\left\langle X^{(k)},X^{(k)}\right\rangle\left(\hat{S}^{(k)}+\hat{B}^{(k)}\right)-\frac{1}{n}(X^{(k)})^T y^{(k)}+\tilde{Z}^{(k)}=0\qquad\forall 1\leq k\leq r.$

The next lemma states that $(\tilde{B},\tilde{S})$ is equal to the optimal solution $(\hat{B},\hat{S})$ to (2) provided that the dual candidate $\widetilde{Z}$ is feasible.

Lemma 6. *Under our assumptions on the design matrices $X^{(k)}$, the candidate pair $(\tilde{S},\tilde{B})$ is unique solution to the problem* (2) *if the dual candidate $\widetilde{Z}$ satisfies*

(C4) $\left\|P_{U_s^c(\bar{S})}(\widetilde{Z})\right\|_{\infty,\infty}<\lambda_s.$

(C5) $\left\|P_{U_b^c(\bar{B})}(\widetilde{Z})\right\|_{\infty,1} < \lambda_b$.

Proof. By construction, and assumption (C4) , $\widetilde{Z} \in \partial\|\tilde{S}\|_{1,1}$. Similarly, by construction and assumptions (C5), $\widetilde{Z} \in \partial\|\tilde{B}\|_{1,\infty}$. Thus, from Lemma 5 $((\tilde{S},\tilde{B}),\widetilde{Z})$ is an optimal primal-dual pair of (2).

Uniqueness. By our assumptions on design matrices $X^{(k)}$, the matrix $\frac{1}{n}\left\langle X^{(k)}_{\mathcal{U}_k}, X^{(k)}_{\mathcal{U}_k}\right\rangle$ is invertible for all $1 \le k \le r$. Thus, as a function of the sum of the two components, the problem (2) is *strictly* convex, so that the sum of the two components is unique. Now, by Lemma 4, we know that $\tilde{B}$ and $\tilde{S}$ satisfy $(\tilde{B},\tilde{S}) = \mathcal{H}_d(\tilde{B},\tilde{S})$ where $d = \lfloor\frac{\lambda_b}{\lambda_s}\rfloor$. It can also be verified that $(\mathcal{H}_d)^{-1}(\tilde{B},\tilde{S})$ is the d-split of the sum of $\tilde{B}+\tilde{S}$, which is unique. Thus, the component pair $(\tilde{B},\tilde{S}) = (\mathcal{H}_d)^{-1}(\tilde{B},\tilde{S})$ is unique. □

Let $\Delta = \tilde{B}+\tilde{S}-\bar{B}-\bar{S}$. From the optimality conditions for the oracle problem (4), we have

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

and consequently,

$$\Delta^{(k)}_{\mathcal{U}_k} = \left(\frac{1}{n}\left\langle X^{(k)}_{\mathcal{U}_k}, X^{(k)}_{\mathcal{U}_k}\right\rangle\right)^{-1}\left(\frac{1}{n}\left(X^{(k)}_{\mathcal{U}_k}\right)^T w^{(k)} - \tilde{Z}^{(k)}_{\mathcal{U}_k}\right). \tag{5}$$

Solving for $\tilde{Z}^{(k)}_{\bigcap_{k=1}^r \mathcal{U}_k^c}$ from (3), for all $j \in \bigcap_{k=1}^r \mathcal{U}_k^c$, we get

$$\tilde{Z}^{(k)}_j = -\frac{1}{n}\left\langle X^{(k)}_j, X^{(k)}_{\mathcal{U}_k}\right\rangle \Delta^{(k)}_{\mathcal{U}_k} + \frac{1}{n}\left(X^{(k)}_j\right)^T w^{(k)}.$$

Substituting for the value of $\Delta^{(k)}_{\mathcal{U}_k}$, we get

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

E Proof of Theorem 1

Proof of Theorem 1: Let $d = \lfloor\frac{\lambda_s}{\lambda_b}\rfloor$ and $(\bar{B},\bar{S}) = \mathcal{H}(B^*,S^*)$, where, (B^*,S^*) is the d-split (see Appendix B) of Θ^*. Then, the result follows from Proposition 1.

Proposition 1 (Sufficient Conditions for 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*$(\bar{S})$ *and RowSupp*$(\hat{B}) \subseteq$ *RowSupp*$(\bar{B})$.

(P2) $\left\|\hat{B}+\hat{S}-\bar{B}-\bar{S}\right\|_\infty \le \sqrt{\frac{4\sigma^2\log(pr)}{C_{min}n}} + \lambda_s D_{max} =: \mathbf{T}$.

(P3) sign $\left(\textit{Supp}(\hat{S}_j)\right) =$ *sign* $\left(\textit{Supp}(\bar{S}^*_j)\right)$ *for all* $j \notin$ *RowSupp*$(\bar{B})$ *provided that* $\min_{\substack{j\notin \textit{RowSupp}(\bar{B})\\(j,k)\in \textit{Supp}(\bar{S})}} \left|\bar{S}^{(k)}_j\right| > \mathbf{T}$.

(P4) sign $\left(\textit{Supp}(\hat{S}_j+\hat{B}_j)\right) =$ *sign* $\left(\textit{Supp}(\bar{S}^*_j+\bar{B}_j)\right)$ *for all* $j \in$ *RowSupp*$(\bar{B})$ *provided that* $\min_{(j,k)\in \textit{Supp}(\bar{B})} \left|\bar{B}^{(k)}_j + \bar{S}^{(k)}_j\right| > \mathbf{T}$.

Proof. We prove the result separately for each part.

(P1) Consider the primal-dual constructed in Appendix D, it suffices to show that (C3) and (C4) in Lemma 6 are satisfied with high probability. By Lemma 7, with probability at least $1 - c_1 \exp(-c_2 n)$ those two conditions are hold and hence, $(\hat{S}, \hat{B}) = (\tilde{S}, \tilde{B})$ is the unique solution of (2) and the property (P1) follows.

(P2) Using (5), we have

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

where, the second inequality holds with high probability as a result of Lemma 8 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 4, this event is equivalent to the event that for all $j \notin$ RowSupp$(\bar{B})$ with $(j,k) \in$ Supp$(\bar{S})$, we have $\left(\Delta_j^{(k)} + \bar{S}_j^{(k)}\right)$ sign $\left(\bar{S}_j^{(k)}\right) > 0$. By Hoeffding inequality, we have

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

By part (P2), this event happens with high probability if $\min_{\substack{j \notin \text{RowSupp}(\bar{B}) \\ (j,k) \in \text{Supp}(\bar{S})}} \left|\bar{S}_j^{(k)}\right| > \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max}$.

(P4) Using (P1) in Lemma 4, this event is equivalent to the event that for all $j \in$ RowSupp$(\bar{B})$, we have $\left(\Delta_j^{(k)} + \bar{B}_j^{(k)} + \bar{S}_j^{(k)}\right)$ sign $\left(\bar{B}_j^{(k)} + \bar{S}_j^{(k)}\right) > 0$. By Hoeffding inequality, we have

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

By part (P2), this event happens with high probability if $\min_{(j,k) \in \text{Supp}(\bar{B})} \left|\bar{B}_j^{(k)} + \bar{S}_j^{(k)}\right| > \sqrt{\frac{4\sigma^2 \log(pr)}{C_{min} n}} + \lambda_s D_{max}$.

□

Lemma 7. *Under conditions of Proposition 1, the conditions (C3) and (C4) in Lemma 6 hold for the primal-dual pair constructed in Appendix D 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

$$\left|\left(P_{U_s^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right| - \|\tilde{S}_j\|_0} & j \in \text{RowSupp}(\tilde{B}) \quad \& \quad (j,k) \notin \text{Supp}(\tilde{S}) \\ \left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c \\ 0 & \text{ow.} \end{cases}.$$

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

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

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 8, 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

$$\sum_{k=1}^{r}\left|\left(P_{U_b^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \lambda_s\|\tilde{S}_j\|_0 & j \in \bigcup_{k=1}^{r}\mathcal{U}_k - \text{RowSupp}(\bar{B}) \\ \sum_{k=1}^{r}\left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

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

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

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 8, 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 6 are satisfied.

□

Lemma 8. $\mathbb{P}\left[\max_{1\leq k\leq r}\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty < \alpha\right] \geq 1 - 2\exp\left(-\frac{\alpha^2 n}{4\sigma^2} + \log(pr)\right)$

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 $\mathcal{N}\left(0, \frac{\sigma^2}{n}\left(X^{(k)}\right)^T X^{(k)}_{\mathcal{U}_k}\right)$. Using Hoeffding inequality, we have

$$\begin{aligned}\mathbb{P}\left[\left\|\frac{1}{n}\left(X^{(k)}\right)^T w^{(k)}\right\|_\infty \geq \alpha\right] &\leq \sum_{j=1}^{p}\mathbb{P}\left[\left|\frac{1}{n}\left(X_j^{(k)}\right)^T w^{(k)}\right| \geq \alpha\right] \\ &\leq \sum_{j=1}^{p} 2\exp\left(-\frac{\alpha^2 n}{2\sigma^2\left(X_j^{(k)}\right)^T X_j^{(k)}}\right) \\ &\leq 2p\exp\left(-\frac{\alpha^2 n}{4\sigma^2}\right).\end{aligned} \tag{7}$$

By union bound, the result follows. □

F Proof of Theorem 2

Proof of Theorem 2: Let $d = \lfloor \frac{\lambda_s}{\lambda_b} \rfloor$ and $(\bar{B}, \bar{S}) = \mathcal{H}(B^*, S^*)$, where, (B^*, S^*) is the d-split (see Appendix B) of Θ^*. Then, the result follows from Proposition 1.

Proposition 2 (Sufficient Conditions for Gaussian Design Matrices)**.** *Under assumptions of Theorem 2, if* $n > \max\left(\frac{Bs\log(pr)}{C_{min}\gamma_s^2}, \frac{Bsr\left(r\log(2)+\log(p)\right)}{C_{min}\gamma_b^2}\right)$ *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*$(\bar{B})$ *and Supp*$(\hat{S}) \subseteq$ *Supp*$(\bar{S})$.

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

(P3) sign $\left(\text{Supp}(\hat{S}_j)\right) = \text{sign}\left(\text{Supp}(\bar{S}_j^*)\right)$ *for all* $j \notin$ *RowSupp*$(\bar{B})$ *provided that* $\min_{\substack{j\notin \text{RowSupp}(\bar{B})\\(j,k)\in\text{Supp}(\bar{S})}}\left|\bar{S}_j^{(k)}\right| > \mathbf{T}$.

(P4) sign $\left(\text{Supp}(\hat{S}_j + \hat{B}_j)\right) = \text{sign}\left(\text{Supp}(\bar{S}_j^* + \bar{B}_j)\right)$ *for all* $j \in$ *RowSupp*$(\bar{B})$ *provided that* $\min_{(j,k)\in\text{Supp}(\bar{B})}\left|\bar{B}_j^{(k)} + \bar{S}_j^{(k)}\right| > \mathbf{T}$.

Proof. We provide the proof of each part separately.

(P1) Consider the primal-dual pair $(\tilde{S}, \tilde{B}, \widetilde{Z})$ constructed in Appendix D. It suffices to show that the conditions (C3) and (C4) in Lemma 6 are satisfied under these assumptions. Lemma 9 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 (5), we have

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

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 [4] 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 $\frac{\sigma^2}{n}\left(\frac{1}{n}\left\langle X_{\mathcal{U}_k}^{(k)}, X_{\mathcal{U}_k}^{(k)}\right\rangle\right)^{-1}$.

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

From the concentration of Gaussian random variables (Lemma 8) 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 $t = (1+\epsilon)\sqrt{\frac{50\sigma^2\log(rs)}{nC_{min}}}$ for some $\epsilon > 0$, the result follows.

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

□

Lemma 9. *Under the assumptions of Proposition 2, the conditions (P3) and (P4) in Lemma 6 hold for the primal-dual pair constructed in Appendix D 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

$$\left|\left(P_{U_s^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right| - \|\tilde{S}_j\|_0} & j \in \text{RowSupp}(\tilde{B}) \quad \& \quad (j,k) \notin \text{Supp}(\tilde{S}) \\ \left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^r \mathcal{U}_k^c \\ 0 & \text{ow.} \end{cases}.$$

By our assumptions on the ratio of the penalty regularizer coefficients, we have $\frac{\lambda_b - \lambda_s\|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right| - \|\tilde{S}_j\|_0} < \lambda_s$. 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 [3]), we have

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

The second inequality follows from the triangle inequlaity on the distributions. By Lemma 10, 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 Var $\left(\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2}{n}$. Using the concentration results for the zero-mean Gaussian random variable $\mathcal{W}_j^{(k)}$ and using the union bound, we get

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

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

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

By concentration of Gaussian random variables, we have

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

Using these bounds, we get

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

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 $\lambda_s > \frac{\sqrt{4\sigma^2 C_{min}\log(pr)}}{\gamma_s\sqrt{nC_{min}} - \sqrt{Bs\log(pr)}}$ 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

$$\sum_{k=1}^{r}\left|\left(P_{U_b^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \lambda_s\|\tilde{S}_j\|_0 & j \in \bigcup_{k=1}^{r}\mathcal{U}_k - \text{RowSupp}(\bar{B}) \\ \sum_{k=1}^{r}\left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r}\mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

We have $\lambda_s\|\tilde{S}_j\|_0 \leq \lambda_s D(\bar{S}) < \lambda_b$ 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

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

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

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

Using the union bound and previous discussion, we get

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

We have

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

and consequently by concentration of Gaussian variables,

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

Finally, we have

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

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 $\frac{(\gamma_b\lambda_b-t_0)^2n}{4\sigma^2r}=\frac{t_0^2nC_{min}}{2rs\lambda_b^2}$), if

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

provided that $n>\frac{Bsr(r\log(2)+\log(p))}{\gamma_b^2C_{min}}$ 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 6 are satisfied.

□

Lemma 10. $\mathbb{P}\left[\max_{1\leq k\leq r}\max_{1\leq j\leq p}\left\|X_j^{(k)}\right\|_2^2\leq 2n\right]\geq 1-\exp\left(-\left(1-\frac{\sqrt{3}}{2}\right)n+\log(pr)\right).$

Proof. Notice that $\left\|X_j^{(k)}\right\|_2^2$ is a χ^2 random variable with n degrees of freedom. According to [1], 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 $t=\left(\frac{\sqrt{3}-1}{2}\right)^2n$ and using the union bound, the result follows. □

G Proof of Theorem 3

We will actually prove a more general theorem, from which Theorem 3 would follow as a corollary. Let

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

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* $\left|\left\{j \in \textit{RowSupp}(\bar{B}) : \left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| = o(\lambda_s)\right\}\right| = (1-\tau)\alpha s$ *, then the result of Theorem 3 holds for* $\theta(n,s,p,\alpha) = \frac{n}{g(\kappa,\tau,\alpha)\, s\log(p-(2-\alpha)s)}$.

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 $g(\kappa,\tau,\alpha)$:

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

Proof of Theorem 3: The proof follows from Corollary 4 by setting $\tau = 0$.

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

Lemma 11. *For any* $j \in \textit{RowSupp}(\bar{B})$*, if* $\left|\bar{S}_j^{(k)}\right| < o\left(\lambda_s\right)$ *then* $\tilde{S}_j^{(k)} = 0$ *with probability* $1 - c_1\exp(-c_2 n)$.

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

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

Using the ℓ_∞ bound on the error, for some constant c, we have

$$\begin{aligned}
\mathbb{P}\left[\left|\tilde{S}_j^{(k)}\right| > 0\right] &\le \mathbb{P}\left[n\lambda_s < \frac{1}{c}\left|\bar{S}_j^{(k)}\right|\left\|X_j^{(k)}\right\|_2^2\right]\\
&= \mathbb{P}\left[\frac{c\lambda_s}{\left|\bar{S}_j^{(k)}\right|}n < \left\|X_j^{(k)}\right\|_2^2\right].
\end{aligned}$$

Notice that $\mathbb{E}\left[\left\|X_j^{(k)}\right\|_2^2\right] = n$. According to the concentration of χ^2 random variables concentration theorems (see [1]), this probability vanishes exponentially fast in n for $\left|\bar{S}_j^{(k)}\right| < c\lambda_s$.

□

G.1 Proof of Theorem 4

We will now provide the proofs of different parts separately.

$[4(a)]$ To prove the first part, recall the primal-dual pair $(\tilde{B}, \tilde{S}, \widetilde{Z})$ constructed in Appendix D. It suffices to show that the dual variable $\widetilde{Z}$ satisfies the conditions (C3) and (C4) of Lemma 6. By Lemma 12, 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 $D_{max} = 1$.

G.1.1 Proof of Theorem 4(b)

The proof of Theorem 4(b) follows from the next lemma.

Lemma 12. *Under assumptions of Theorem 3, the conditions (C3) and (C4) in Lemma 6 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

$$\left|\left(P_{U_s^c}(\widetilde{Z})\right)_j^{(k)}\right| = \begin{cases} \dfrac{\lambda_b - \lambda_s \|\tilde{S}_j\|_0}{\left|M_j^{\pm}(\tilde{B})\right| - \|\tilde{S}_j\|_0} & j \in \text{RowSupp}(\tilde{B}) \quad \& \quad (j,k) \notin \text{Supp}(\tilde{S}) \\ \left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^{r} \mathcal{U}_k^c \\ 0 & \text{ow.} \end{cases}.$$

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

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

By Lemma 10, if $n \geq \frac{2}{2-\sqrt{3}} \log(pK)$ then with high probability $\left\| X_j^{(k)} \right\|_2^2 \leq 2n$ and hence $\text{Var}\left(\mathcal{W}_j^{(k)}\right) \leq \frac{2\sigma^2}{n}$. 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 $\mathcal{W}_j^{(k)}$ and using the union bound, we get

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

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

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

According to the result of [4] on singular values of Gaussian matrices, for the matrix $X_{\mathcal{U}_k}^{(k)}$, we have

$$\mathbb{P}\left[\sigma_{min}\left(X_{\mathcal{U}_k}^{(k)} \right) \leq (1-\delta)\left(\sqrt{n} - \sqrt{s}\right) \right] \leq \exp\left(-\frac{\delta^2 \left(\sqrt{n} - \sqrt{s}\right)^2}{2} \right) \quad \forall \delta > 0,$$

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

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

According to Lemma 11, if $\left|\left|\Theta_j^{*(1)}\right| - \left|\Theta_j^{*(2)}\right|\right| = o(\lambda_s)$, 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

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

Similarly,

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

By concentration of Gaussian random variables, we have

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

Using these bounds, we get

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

This probability goes to 1 for $t_0 = \frac{\sqrt{(f_1(\kappa)+f_2(\kappa))ns}\lambda_s}{\sqrt{(f_1(\kappa)+f_2(\kappa))ns}\lambda_s+2\sigma(\sqrt{n}-\sqrt{s})}\lambda_s$ (the solution to $\frac{t_0^2\left(\sqrt{n}-\sqrt{s}\right)^2}{(f_1(\kappa)+f_2(\kappa))s\lambda_s^2} = \frac{(\lambda_s-t_0)^2 n}{4\sigma^2}$), if

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

provided that (substituting $r=2$),

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

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

$$n > \big(f_1(\kappa)+f_2(\kappa)\big)\,s\log\Big(p-(2-\alpha)s\Big). \tag{8}$$

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

$$\sum_{k=1}^r\left|\Big(P_{U_b^c}(\widetilde{Z})\Big)_j^{(k)}\right| = \begin{cases} \lambda_s\|\tilde{S}_j\|_0 & j \in \bigcup_{k=1}^r \mathcal{U}_k - \text{RowSupp}(\bar{B}) \\ \sum_{k=1}^r\left|\tilde{Z}_j^{(k)}\right| & j \in \bigcap_{k=1}^r \mathcal{U}_k^c \\ 0 & \text{ow} \end{cases}.$$

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

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

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,

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

Using the union bound and previous discussion, we get

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

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

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

ans consequently,

$$\begin{aligned}\mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\sum_{k=1}^{r}\left|\mathcal{Z}_j^{(k)}\right| \geq t\right] &= \mathbb{P}\left[\max_{j\in\bigcap_{k=1}^{r}\mathcal{U}_k^c}\max_{\mathbf{v}\in\{-1,+1\}^r}\sum_{k=1}^{r} v_k\mathcal{Z}_j^{(k)} \geq t\right] \\ &\leq 2\exp\left(-\frac{t^2\left(\sqrt{n}-\sqrt{s}\right)^2}{\frac{1}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\lambda_b^2} + r\log(2) + \log\left(p-(2-\alpha)s\right)\right) \qquad \forall t \geq 0.\end{aligned}$$

Finally, we have

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

This probability goes to 1 for $t_0 = \frac{\sqrt{\frac{1}{\kappa^2}(f_1(\kappa)+f_2(\kappa))ns}\lambda_b}{\sqrt{\frac{1}{\kappa^2}(f_1(\kappa)+f_2(\kappa))ns}\lambda_b + 2\sigma(\sqrt{n}-\sqrt{s})}\lambda_b$ (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

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

provided that (substituting $r = 2$),

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

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

$$n > \frac{2}{\kappa^2}\left(f_1(\kappa)+f_2(\kappa)\right)s\log\Big(p-(2-\alpha)s\Big).$$

Combining this result with (8), the lemma follows. □

G.1.2 Proof of Theorem 4(c)

We prove this result by contradiction. Suppose there exist a solution to (2), say $(\hat{B}, \hat{S})$ such that sign $\left(\text{Supp}(\hat{B}+\hat{S})\right) = \text{sign}\left(\text{Supp}(B^*+S^*)\right)$. By Lemma 4, this is equivalent to having sign $\left(\text{Supp}(\hat{B})\right) = \text{sign}\left(\text{Supp}(B^*)\right)$ and sign $\left(\text{Supp}(\hat{S})\right) = \text{sign}\left(\text{Supp}(S^*)\right)$ and $\frac{\lambda_s}{\lambda_b} = \kappa$.

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

Case (i).

We will show that 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 5.

Using (6) 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

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

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

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

The second inequality holds with probability at least $1 - c_1 \exp\left(-c_2\left(\sqrt{n}+\sqrt{s}\right)^2\right)$ as a result of [4] on the eigenvalues of Gaussian matrices. The third inequality holds with probability at least $1 - c_3\exp(-c_4 n)$ as a result of [1] on the magnitude of χ^2 random variables. 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

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

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

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

Given these lower bounds on the variance, by results on Gaussian maxima (see [4]), for any $\delta > 0$, with high probability,

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

This in turn can be bound as

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

Consider two cases:

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

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

   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 $n < (1-\nu)\left(f(\kappa)\right)s\log(p-(2-\alpha)s)$.

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

   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 5.

Case (ii). We need to show that with high probability, there exist 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 5.

Following the same proof technique, notice that $\sum_{k=1}^{r} \tilde{Z}_j^{(k)}$ is a zero-mean Gaussian random variable with $\text{Var}\left(\sum_{k=1}^{r} \tilde{Z}_j^{(k)}\right) \geq r(\widetilde{\sigma}_1^2 + \widetilde{\sigma}_2^2)$. 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 Theorem 3(b). □

H Synthetic and Real World Data used for Performance Analysis

In this section we present the structure of synthetic and real datasets we used and how we fit them into our model.

H.1 Synthetic Dataset

Data Generating: We explain how we generated the data for our simulation here. We pick three different values of $p = 128, 256, 512$ and let $s = \lfloor 0.1p \rfloor$. For different values of α, we let n swipe from $0.Es\log(p-(2-\alpha)s)$ to $Ds\log(p-(2-\alpha)s)$. We generate 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. Then, we multiply 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 matrices $X^{(1)}$, $X^{(2)}$ and W and use one of them for training and the other one for cross validation (test), 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 $X^{(k)}$, $Theta^*$ 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 want to recover matrices $\hat{B}$ and $\hat{S}$ that satisfy (2). We use the coordinate descent algorithm to numerically solve the problem (see Appendix I). The algorithm inputs 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 $(\hat{B}, \hat{S})$. The inputs $(\underline{B}, \underline{S})$ are initial guess and can be set to zero. However, when we search for optimal penalty regularizer coefficients, we can use the result for previous set of coefficients (λ_b, λ_s) as a good initial guess for the next coefficients $(\lambda_b + \xi, \lambda_s + \zeta)$. The parameter ϵ captures the stopping criterion threshold of the algorithm. We iterate inside the algorithm until the relative update change of the objective function is less than ϵ. Since we do not run the algorithm completely (until $\epsilon = 0$ works), we need to filter the small magnitude values in the solution $(\hat{B}, \hat{S})$ and set them to be zero.

Choosing penalty regularizer coefficients: Dictated by optimality conditions, we have $1 > \frac{\lambda_s}{\lambda_b} > \frac{1}{2}$. Thus, searching range for one of the coefficients is bounded and known. We set $\lambda_b = c\sqrt{\frac{rlog(p)}{n}}$ and search for $c \in [0.01, 100]$, where this interval is partitioned logarithmic. For any pair (λ_b, λ_s) we compute the objective function of Y_{Ts} and $X_{\text{Ts}}^{(k)}$ for $k = 1, 2$ using the filtered $(\hat{B}, \hat{S})$ from the coordinate descent algorithm. Then across all choices of (λ_b, λ_s), we pick the one with minimum objective function on the test data. Finally we let $\hat{\Theta} = \text{Filter}(\hat{B} + \hat{S})$ for $(\hat{B}, \hat{S})$ corresponding to the optimal (λ_b, λ_s).

H.2 Handwritten Digits Dataset

Structure of the Dataset: In this dataset, there are 200 instances of handwritten digits 0-9 (totally 2000 digits). Each instance of each digit is scanned to an image of the size 30×48 pixels. This image is NOT provided by the dataset.

	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 1: 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.

Using the full resolution image of each digit, the dataset provides six different classes of features. A total of 649 features are provided for each instance of each digit. The information about each class of features is provided in Table 1. The combined handwriting images of the record number 100 is shown in Fig 1 (ten images are concatenated together with a spacer between each two).

Fitting the dataset to our model: Regardless of the nature of the features, 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. We divide each feature by an appropriate number (perhaps larger than the maximum of that feature in the dataset) to make sure that the dynamic range of all features is a (not too small) subset of $[-1, 1]$. Notice that in this division process, we don't care about the minimum and maximum of the training set. We just divide each feature by a fixed and predetermined number we provided as maximum in Table 1. For example, we divide the Pixel Shape feature by 6, Karhunen-Loeve coefficients by 17 or the last morphological feature by 18000 and so on. We do not shift the data; we only scale it.

Out of 200 samples provided for each digit, we take $n \leq 200$ samples for training. Let $X^{(k)} = X \in \mathbb{R}^{10n \times 649}$ for all $0 \leq k \leq 9$ 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. Consequently, we set the vector $y^{(k)} \in \{0, 1\}^{10n}$ to be the vector such that $y_j^{(k)} = 1$ if and only if the j^{th} row of the feature matrix X corresponds to the digit k. This setup is called binary classification setup.

We want to find a block-sparse matrix $\hat{B} \in \mathbb{R}^{649 \times 10}$ and a sparse matrix $\hat{S} \in \mathbb{R}^{649 \times 10}$, so that for a given feature vector $\mathbf{x} \in \mathbb{R}^{649}$ extracted from the image of a handwritten digit $0 \leq k^* \leq 9$, we ideally have $k^* = \arg\max_{0 \leq k \leq 9} \mathbf{x} \left(\hat{B} + \hat{S} \right)$.

To find such matrices $\hat{B}$ and $\hat{S}$, we solve (2). We tune the parameters λ_b and λ_s in order to get the best result by cross validation. 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, empirically $c \in [0.01, 10]$ is a constant to be searched.

I 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 pseudocode is in Algorithm 1 to Algorithm 3.

Algorithm 1 Dirty Model Solver

Input: X, Y, λ_b, λ_s, B, S and ε
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 2 UpdateB

Input: c, d, λ_b, B and S
Output: B
Update B using the cyclic coordinate descent algorithm for ℓ_1/ℓ_∞ while keeping S unchanged.

for $j = 1 : p$ **do**
  for $k = 1 : r$ **do**
    $\alpha_j^{(k)} \leftarrow c_j^{(k)} - \sum_{i \neq j} (b_i^{(k)} + s_i^{(k)}) d_{i,j}^{(k)} - s_i^{(k)} d_{j,j}^{(k)}$
    if $\sum_{k=1}^{r} |\alpha_j^{(k)}| \leq \lambda_b$ **then**
      $b_j \leftarrow 0$
    else
      Sort α 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

Figure 1: An instance of images of the ten digits extracted from the dataset

I.1 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.

For updating B, the subproblem is updating B_j

$$\hat{B}_j = \arg\min_{B_j} \quad \frac{1}{2}\sum_{k=1}^{r}\left\|R_j^{(k)} - B_j^{(k)}X_j^{(k)}\right\|_2^2 + \lambda_b\|B_j\|_\infty. \tag{9}$$

Algorithm 3 Update-S

Input: c, d, λ_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**
    $\alpha_j^{(k)} \leftarrow c_j^{(k)} - \sum_{i \neq j}(b_i^{(k)} + s_i^{(k)})d_{i,j}^{(k)} - s_i^{(k)} d_{j,j}^{(k)}$
    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

If we take the partial residual vector $R_j^{(k)} = y^{(k)} - \sum_{l \neq j}(B_l^{(k)} X_l^{(k)}) - \sum_l (S_l^{(k)} X_l^{(k)})$, the correctness of this algorithm will directly follow from the correctness of coordinate descent algorithm of ℓ_1/ℓ_{inf} in [2]. With the same argument, the correctness of the Algorithm 3 can be proven.

References

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

[2] 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.

[3] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[4] Davidson K. R. and Szarek S. J. Local operator theory, random matrices and banach spaces. *Handbook of Banach Spaces*, 1:317–336, 2001.

[5] 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.