# Non-Asymptotic Analysis of Block-Regularized Regression Problem

Salar Fattahi and Somayeh Sojoudi

***Abstract*—Given a linear multivariate regression problem with block sparsity structure on the regression matrix, one popular approach for estimating its unknown parameter is block-regularization, where the sparsity of different blocks of the regression matrix is promoted by penalizing their $\ell_\infty$-norms. The main goal of this work is to characterize the properties of this estimator under high-dimensional scaling, where the growth rate of the dimension of the problem is comparable or even faster than that of the sample size. In particular, this work generalizes the existing non-asymptotic results on special instances of block-regularized estimators to the case where the unknown regression matrix has an arbitrary number of blocks each with a potentially different size. When the design matrix is deterministic, a sharp non-asymptotic rate is derived on the element-wise error of the proposed estimator. Furthermore, it is proven that the same error rate approximately holds for the block-regularized estimator when the design matrix is randomly generated, provided that the number of samples exceeds a lower bound. The accuracy of the proposed estimator is illustrated on several test cases.**

## I. Introduction

Since the turn of the millennium, big-data analysis has become an indispensable part of different real-world problems. In many applications, including neuroscience, transportation, and system identification, the dimensionality of the data is overwhelmingly large, often surpassing the number of available data samples [1]–[3]. Under such *large dimension-small sample size* regime, many classical consistency results in statistics face major breakdowns. For instance, given a number of independent and identically distributed samples drawn from a multivariate Gaussian distribution, none of the eigenvectors and eigenvalues of the sample covariance matrix may converge to their nominal values if the sample size does not grow sufficiently faster than the dimension of the random vector [4].

On the other hand, the data collected from real-world systems often possesses a simple, localized and sparse underlying model: the traffic flows are correlated only locally in transportation networks [2], brain networks have sparse functional connectivity [1], and state and input vectors have limited interactions in distributed large-scale dynamical systems [5]. A substantial body of work has been devoted to analyzing non-asymptotic behavior of regularized estimators by leveraging the underlying sparsity of the true model. In particular, special attention has been devoted to $\ell_1$-regularized estimators, namely Lasso [6] and its variants such as group Lasso, fused Lasso, and graphical Lasso [7]–[9].

Email: fattahi@berkeley.edu and sojoudi@berkeley.edu.

Salar Fattahi is with the Department of Industrial Engineering and Operations Research, University of California, Berkeley. Somayeh Sojoudi is with the Departments of Electrical Engineering and Computer Sciences and Mechanical Engineering as well as the Tsinghua-Berkeley Shenzhen Institute, University of California, Berkeley. This work was supported by the ONR grant N00014-17-1-2933, DARPA grant D16AP00002, and AFOSR grant FA9550-17-1-0163.

Driven by the existing non-asymptotic results on the classical Lasso problem, the main focus of this paper is on the block-regularized estimators for linear regression problems, where the goal is to impose sparsity constraints on different blocks of the regression parameter rather than on its individual elements. To this goal, the $\ell_\infty$-norms of the blocks are penalized instead of their $\ell_1$-norms. One motivation behind the analysis of the behavior of this type of estimator stems from topology extraction in consensus networks and system identification problems, especially in the multi-agent setting [10], [11]. In this problem, given a number of subsystems (agents) whose interactions are defined via an unknown sparse topology network, the objective is to estimate the state-space model governing the entire system based on a limited number of input-output sample trajectories. Since the subsystems have their own local state and input vectors with potentially different sizes, the parameters of the state-space model admit a block-sparse structure. Together with the recent results on the sample complexity of the system identification problem [3], [12], the findings of the present paper can pave the way toward a rigorous statistical analysis of different methods for sparse system identification problem.

While the traditional Lasso is heavily studied in the literature [6], [13], the high-dimensional behavior of the block-regularized estimator is less known when the dimensions of blocks are arbitrary. [14] analyzes the high-dimensional consistency of this estimator when each block of the regression parameter is a row vector. That work assumes that the regression parameter consists of one column of blocks. In the present paper, these results are significantly generalized to problems with an arbitrary number of blocks each with general sizes. In particular, it is proven that the elementwise error of the block-regularized estimator decreases at the rate $O(\sqrt{((pr)^2 + pr\log\bar{p})/d})$, where $d$ is the number of samples, $\bar{p}$ is the number of row blocks in the unknown regression matrix, and $p \times r$ is the size of its largest block. Furthermore, it is shown that, for random design matrices, $d = \Omega(k(pr)^2 + pr\log\bar{p})$ is enough to guarantee such estimation error rate with high probability, where $k$ is the maximum number of nonzero elements in the columns of the regression matrix. A number of important special cases of these results will be discussed and connections with the existing works will be drawn. Finally, the efficacy of the proposed estimator will be demonstrated on different case studies.

**Notations:** Given integer sets $I$ and $J$ together with a matrix $M$, the notation $M_{IJ}$ refers to the submatrix of $M$ whose rows and columns are indexed by $I$ and $J$, respectively. $M_{:,j}$ is used to denote the $j^{\text{th}}$ column of $M$. Given the sequences $f_1(n)$ and $f_2(n)$, the notations $f_1(n) = O(f_2(n))$ and $f_1(n) = \Omega(f_2(n))$ imply that there exist $c_1 < \infty$ and $c_2 > 0$ such that $f_1(n) \leq c_1 f_2(n)$ and $f_1(n) \geq c_2 f_2(n)$, re-

spectively. $f_1(n) = o(f_2(n))$ indicates that $f_1(n)/f_2(n) \to 0$ as $n \to \infty$. A zero-mean Gaussian distribution with covariance $\Sigma$ is referred to as $N(0, \Sigma)$. Given an event $\mathcal{T}$, let $\mathbb{P}(\mathcal{T})$ denote its probability. For a matrix $M$, the symbols $\|M\|_F$, $\|M\|_1$ and $\|M\|_\infty$ are used to denote its Frobenius, $\ell_1/\ell_1$ (summation of the absolute values of the elements), and $\ell_\infty/\ell_\infty$ (maximum of the absolute values of the elements) norms, respectively. $\lambda_{\min}(M)$ denotes the minimum eigenvalue of a symmetric matrix $M$.

## II. Problem Formulation

Consider the linear system

$$Y = X\Theta + W \tag{1}$$

where $Y \in \mathbb{R}^{d\times r}$ is the observation matrix, $X \in \mathbb{R}^{d\times p}$ is the design matrix, $\Theta \in \mathbb{R}^{p\times r}$ is the regression matrix, and $W \in \mathbb{R}^{d\times r}$ is the observation noise. Here, $d$ is the number of available data samples. We assume that the elements of $W$ are independently drawn from $N(0, \sigma_w^2)$. The goal is to estimate $\Theta$ based on $X$ and $Y$ under the assumption that $\Theta$ is sparse. Depending on the problem at hand, $X$ can be a known deterministic or a randomly generated matrix with a predefined distribution. Under the sparsity assumption on $\Theta$, one way to estimate $\Theta$ is by solving Lasso–a least-squares optimization problem augmented by an $\ell_1$ penalty on the elements of $\Theta$–which is defined as

$$\breve{\Theta} = \arg\min_{\Theta} \frac{1}{2d}\|Y - X\Theta\|_F^2 + \lambda_d\|\Theta\|_1 \tag{2}$$

where $\lambda_d$ is a user-defined regularization coefficient. The second term in (2) has the role of promoting sparsity in $\breve{\Theta}$. Suppose that $\Theta$ can be partitioned as $\Theta = [\Theta^{(i,j)}]$ where $(i,j) \in \{1,...,\bar{p}\} \times \{1,...,\bar{r}\}$ and $\Theta^{(i,j)}$ is the $(i,j)^{\text{th}}$ block of $\Theta$ with size $p_i \times r_j$, and $\bar{p}$ and $\bar{r}$ are the numbers of block rows and columns in $\Theta$. Note that $\sum_{i=1}^{\bar{p}} p_i = p$ and $\sum_{i=1}^{\bar{r}} r_i = r$. Suppose that it is known *a priori* that all elements in $\Theta^{(i,j)}$ are simultaneously zero or nonzero. This implies that, as long as one element in $\Theta^{(i,j)}$ is nonzero, there is no reason to promote sparsity in the remaining elements of $\Theta^{(i,j)}$. Clearly, this kind of block sparsity constraint is not correctly reflected in (2). Instead, one can resort to an $\ell_1/\ell_\infty$ variant of the Lasso problem–known as block-regularized problem:

$$\hat{\Theta} = \arg\min_{\Theta} \frac{1}{2d}\|Y - X\Theta\|_F^2 + \|\Theta\|_{\text{block}} \tag{3}$$

where $\|\Theta\|_{\text{block}}$ is defined as the summation of $\|\Theta^{(i,j)}\|_\infty$ over $(i,j) \in \{1,...,\bar{p}\} \times \{1,...,\bar{r}\}$. Let the true regression matrix be denoted by $\Theta^*$. Given (3), consider the following problems:

**Problem 1.** *For a deterministic design matrix $X$, what are the conditions under which $\hat{\Theta}$ recovers the nonzero blocks of $\Theta^*$ and has a small estimation error with high probability?*

**Problem 2.** *For a randomly generated design matrix $X$, where each row is drawn independently from $N(0,\Sigma)$, what are the conditions under which $\hat{\Theta}$ recovers the nonzero blocks of $\Theta^*$ and has a small estimation error with high probability?*

Answering these questions is at the core of this paper. Before delving into the detailed description of results of this work, note that Problem 2 is significantly more challenging than Problem 1. In particular, the main difficulty of this problem lies in the random nature of the design matrix. Roughly speaking, the proposed bounds should not directly depend on the randomly generated $X$, but rather be in terms of the characteristics of its distribution. In fact, it will be proven that a lower bound on the sample size is required in order to *control* the behavior of the design matrix, whereas such assumption is not needed in the proposed solution for Problem 1.

## III. Main Results

In this section, the main results of this work will be presented. Define $\mathcal{A}_j(\Theta) = \{i : \Theta^{(i,j)} \neq 0\}$. Unless stated otherwise, $\mathcal{A}_j$ is used to refer to $\mathcal{A}_j(\Theta^*)$. Define $\mathcal{A}_j^c$ as the complement of $\mathcal{A}_j$. For $\mathcal{T} \subseteq \{1,...,\bar{p}\}$, denote $I(\mathcal{T})$ as the index set of rows in $\Theta^*$ corresponding to the blocks $\{\Theta^{*(i,:)} : i \in \mathcal{T}\}$. For an index set $\mathcal{U}$, define $X_\mathcal{U}$ as a $d \times |\mathcal{U}|$ submatrix of $X$ after removing the columns with indices not belonging to $\mathcal{U}$. With a slight abuse of notation, $X_{(i)}$, $X_{\mathcal{A}_j}$, and $X_{\mathcal{A}_j^c}$ are used to denote $X_{I(\{i\})}$, $X_{I(\mathcal{A}_j)}$, and $X_{I(\mathcal{A}_j^c)}$ when there is no ambiguity. Similarly, $\Sigma_{(i),\mathcal{A}_j}$ and $\Sigma_{\mathcal{A}_j,\mathcal{A}_j}$ are used instead of $\Sigma_{I(\{i\}),I(\mathcal{A}_j)}$ and $\Sigma_{I(\mathcal{A}_j),I(\mathcal{A}_j)}$, respectively. Denote $k_j$ as the maximum number of nonzero elements in different columns of $\Theta^{*(:,j)}$, i.e., the $j^{\text{th}}$ column of blocks. Finally, define

$$\begin{aligned} &p_{\max} = \max_{1\le i\le \bar{p}} p_i, \quad r_{\min} = \min_{1\le j\le \bar{r}} r_j, \quad r_{\max} = \max_{1\le j\le \bar{r}} r_j \\ &k_{\min} = \min_{1\le j\le \bar{r}} k_j \quad k_{\max} = \max_{1\le j\le \bar{r}} k_j, \quad \sigma^2_{\max} = \max_{1\le i\le p} \Sigma_{ii} \end{aligned} \tag{4}$$

Depending on the deterministic or random nature of the design matrix, a number of assumptions should be made to guarantee the consistency of the block-regularized estimator.

**Assumption 1** (Deterministic Design)**.** For a deterministic design matrix $X$, assume that

A1. (Mutual Incoherence Property): There exists a number $\gamma_d \in (0,1]$ such that

$$\max_{j=1,...,\bar{r}} \left\{ \max_{i=1,...,|\mathcal{A}_j^c|} \left\| X_{(i)}^T X_{\mathcal{A}_j} (X_{\mathcal{A}_j}^T X_{\mathcal{A}_j})^{-1} \right\|_1 \right\} \le 1 - \gamma_d \tag{5}$$

A2. (Bounded Minimum Eigenvalue): There exists a number $\Lambda_{\min}^d > 0$ such that

$$\Lambda_{\min}^d \le \min_{j=1,...,\bar{r}} \lambda_{\min}\left(\frac{1}{d} X_{\mathcal{A}_j}^T X_{\mathcal{A}_j}\right) \tag{6}$$

A3. (Bounded Inverse): There exists a number $q_d < \infty$ such that

$$q_d \ge \max_{j=1,...,\bar{r}} \left\| \left(\frac{1}{d} X_{\mathcal{A}_j}^T X_{\mathcal{A}_j}\right)^{-1} \right\|_\infty \tag{7}$$

**Assumption 2** (Random Design)**.** For a randomly generated design matrix $X$ where each row is independently drawn from $N(0,\Sigma)$, assume that

A1. (Mutual Incoherence Property): There exists a number $\gamma_r \in (0,1]$ such that

$$\max_{j=1,...,\bar{r}} \left\{ \max_{i=1,...,|\mathcal{A}_j^c|} \left\| \Sigma_{(i),\mathcal{A}_j} (\Sigma_{\mathcal{A}_j,\mathcal{A}_j})^{-1} \right\|_1 \right\} \le 1 - \gamma_r \tag{8}$$

A2. (Bounded Minimum Eigenvalue): There exists a number $\Lambda^r_{\min} > 0$ such that

$$\Lambda^r_{\min} \leq \min_{j=1,\dots,\bar{r}} \lambda_{\min}(\Sigma_{\mathcal{A}_j,\mathcal{A}_j}) \tag{9}$$

A3. (Bounded Inverse): There exists a number $q_r < \infty$ such that

$$q_r \geq \max_{j=1,\dots,\bar{r}} \left\|(\Sigma_{\mathcal{A}_j,\mathcal{A}_j})^{-1}\right\|_\infty \tag{10}$$

The mutual incoherence property is a commonly known assumption for the exact recovery of unknown parameters in compressive sensing and classical Lasso regression problems [13], [15], [16]. This assumption entails that the effect of the submatrices of $\frac{1}{d}X^TX$ or $\Sigma$ (depending on the deterministic or random nature of the design matrix) corresponding to zero elements of $\Theta^*$ on the remaining entries of $\frac{1}{d}X^TX$ or $\Sigma$ should not be large. Roughly speaking, this condition guarantees that the unknown regression matrix is *recoverable* in the noiseless scenario, i.e., when $W = 0$. If the recovery cannot be guaranteed in the noise-free setting, then there is little hope for the block-regularized estimator to recover the true structure of $\Theta^*$. Another implication of Assumptions 1 and 2 is that the minimum eigenvalue and the maximum absolute value of the elements in the inverse of $\frac{1}{d}X^T_{\mathcal{A}_j}X_{\mathcal{A}_j}$ and $\Sigma_{\mathcal{A}_j,\mathcal{A}_j}$ do not scale with the dimension of the problem. Again, these are standard and easily satisfiable assumptions in the context of sparse estimation theory [6], [14], [16]

### A. Deterministic Design Matrix

The objective of this subsection is to provide an answer to Problem 1. To this goal, non-asymptotic properties of the block-regularized estimator is presented for a deterministic design matrix. Without loss of generality, it is assumed that the 2-norm of each column of $X$ is upper bounded by $\sqrt{2d}$ (the choice of the coefficient is arbitrary). Indeed, this can be achieved by an appropriate scaling of the design matrix. First, we consider a special case where $\bar{r} = 1$, implying that $\Theta$ consists of a single column of the blocks with arbitrary sizes. The following theorem provides an answer to Problem 1 in this special case.

**Theorem 1** (Deterministic $X$, single block column)**.** *Given arbitrary constants $c_1, c_2 > 1$, suppose that $\bar{r} = 1$, $X$ is deterministic, and $\lambda_d$ is chosen such that*

$$\lambda_d \geq \sqrt{\frac{4c_1\sigma_w^2}{\gamma_d^2} \cdot \frac{(r_1 p_{\max})^2 + r_1 p_{\max}\log\bar{p}}{d}} \tag{11}$$

*Then, with probability of at least*

$$\begin{aligned} &1 - 2\exp\big(-(c_1-1)(r_1 p_{\max} + \log\bar{p})\big) \\ &\quad - 2\exp\big(-2(c_2-1)\log(k_1 r_1 p_{\max})\big) \to 1 \end{aligned} \tag{12}$$

*the following statements hold:*

1. *$\hat{\Theta}$ is unique and, in addition,*

$$\|\hat{\Theta} - \Theta^*\|_\infty \leq \sqrt{\frac{4c_2\sigma_w^2\log(k_1 r_1 p_{\max})}{\Lambda^d_{\min} d}} + \lambda_d q_d = g_1 \tag{13}$$

2. *The set of nonzero blocks of $\hat{\Theta}$ excludes the zero blocks of $\Theta^*$. Furthermore, the nonzero blocks of $\hat{\Theta}$ and $\Theta^*$ are equivalent if $\min_{i\in\mathcal{A}_1}\|\Theta^{(i,1)}\|_\infty > g_1$.*

The following corollary is a result of Theorem 1.

**Corollary 1.** *Assuming that $p_{\max} = O(\bar{p})$, the inequality*

$$\|\hat{\Theta} - \Theta^*\|_\infty = O\left(\sqrt{\frac{(r_1 p_{\max})^2 + r_1 p_{\max}\log\bar{p}}{d}}\right) \tag{14}$$

*holds with probability of at least 12.*

It is worthwhile to mention that Theorem 1 is based on a generalization of the primal-dual witness method introduced in [14], where it is assumed that $p_i = 1$ for every $i = 1, \dots, \bar{p}$. Under such circumstances, $p_{\max} = 1$ and the rate 14 coincides with the one in [14]. Furthermore, assuming that $r_1 p_{\max} = O(1)$, Corollary 1 gives rise to the rate of $O(\sqrt{\log p/d})$; this is a well-known rate for the estimation error of the Classical Lasso problem [6]. Next, it is shown that Theorem 1 is the building block of our main result for the block-regularized estimator with general values of $\bar{r}$.

**Assumption 3.** *There exist $\alpha, \delta_{\min} > 0$ such that*

$$\bar{r} = \mathcal{O}(\bar{p}^\alpha) \tag{15}$$

$$k_{\min} r_{\min} p_{\max} = \Omega(\bar{p}^{\delta_{\min}}) \tag{16}$$

Note that this assumption adds two mild restrictions: the rate of increase in the number of block columns of $\Theta^*$ can be at most a polynomial function of the number of its block rows. Furthermore, the growth rate of $k_{\min} r_{\min} p_{\max}$ should be at least a polynomial function of the number of its block rows. As it will be shown later, these assumptions are helpful to guarantee the consistency of the block-regularized estimator in a high-dimensional setting.

**Theorem 2** (Deterministic $X$, multiple block columns)**.** *Given arbitrary constants $c_1 > \alpha + 1$ and $c_2 > \frac{\alpha}{2\delta_{\min}} + 1$, suppose that $X$ is deterministic and $\lambda_d$ is chosen such that*

$$\lambda_d \geq \sqrt{\frac{4c_1\sigma_w^2}{\gamma_d^2} \cdot \frac{(r_{\max} p_{\max})^2 + r_{\max} p_{\max}\log\bar{p}}{d}} \tag{17}$$

*Then, with probability of at least*

$$1 - O\Big(\frac{1}{\bar{p}^{c_1-\alpha-1}} + \frac{1}{\bar{p}^{2\delta_{\min}(c_2-1)-\alpha}}\Big) \to 1 \tag{18}$$

*the following statements hold:*

1. *$\hat{\Theta}$ is unique and, in addition,*

$$\|\hat{\Theta} - \Theta^*\|_\infty \leq \sqrt{\frac{4c_2\sigma_w^2\log(k_{\max} r_{\max} p_{\max})}{\Lambda^d_{\min} d}} + \lambda_d q_d = g_2 \tag{19}$$

2. *The set of nonzero blocks of $\hat{\Theta}$ excludes the zero blocks of $\Theta^*$. Furthermore, the nonzero blocks of $\hat{\Theta}$ and $\Theta^*$ are equivalent if $\min_{1\leq j\leq\bar{r}}\min_{i\in\mathcal{A}_j}\|\Theta^{(i,j)}\|_\infty > g_2$.*

Note that the new lower bound on $c$ in Theorem 2 is to ensure the consistency of the estimator, i.e., the equivalence of the estimated and true regression matrices with a probability converging to 1. One major strength of Theorem 2 over similar

results in [6], [14] is that it allows the number of rows and columns within each block of $\Theta^*$ to grow with $\bar{p}$.

### B. Random Design Matrix

In this subsection, the goal is to address Problem 2. Assume that the rows of $X$ are independently drawn from a common Gaussian distribution $N(0,\Sigma)$. The random nature of $X$ makes the analysis of the estimator more difficult than the deterministic case since additional steps should be taken toward controlling the behavior of the random matrix $X$. Upon controlling $X$, one can make arguments analogous to the deterministic case to prove the high-dimensional consistency of the block-regularized estimator. Similar to the deterministic case, we first focus on the case $\bar{r}=1$.

**Theorem 3** (Random $X$, single block column)**.** *Given arbitrary constants $c_1,c_2,c_3>1$, suppose that $\bar{r}=1$, each row of $X$ is drawn independently from $N(0,\Sigma)$, and that $\lambda_d$ and $d$ are chosen such that*

$$\lambda_d \geq \sqrt{\frac{32c_1\sigma_w^2\sigma_{\max}^2}{\gamma_r^2}\cdot\frac{(r_1p_{\max})^2+r_1p_{\max}\log\bar{p}}{d}} \tag{20}$$

$$d \geq \frac{72c_2\sigma_{\max}^2k_1r_1p_{\max}}{\gamma_r^2\Lambda_{\min}^r}\cdot(r_1p_{\max}+\log\bar{p}) \tag{21}$$

*Then, with probability of at least*

$$\begin{aligned}&1-3\exp\big(-(c_1-1)(r_1p_{\max}+\log\bar{p})\big)\\&-2\exp\big(-(c_2-1)(r_1p_{\max}+\log\bar{p})\big)\\&-2\exp\big(-2(c_3-1)(\log(k_1r_1p_{\max}))\big)-6\exp\left(-\frac{|\mathcal{A}_1|}{2}\right)\to 1\end{aligned} \tag{22}$$

*Then, the following statements hold:*

1. *$\hat{\Theta}$ is unique and, in addition,*

$$\begin{aligned}\|\hat{\Theta}-\Theta^*\|_\infty \leq&\sqrt{\frac{36c_3\sigma_w^2\log(k_1r_1p_{\max})}{\Lambda_{\min}^r d}}\\&+\lambda_d\left(\frac{8k_1}{\Lambda_{\min}^r}\sqrt{\frac{r_1p_{\max}}{d}}+q_r\right)=g_3\end{aligned} \tag{23}$$

2. *The set of nonzero blocks of $\hat{\Theta}$ excludes the zero blocks of $\Theta^*$. Furthermore, the nonzero blocks of $\hat{\Theta}$ and $\Theta^*$ are equivalent if $\min_{i\in\mathcal{A}_1}\|\Theta^{(i,1)}\|_\infty>g_3$.*

Despite of being more sophisticated, Theorem 3 is similar to Theorem 1 in nature. In what follows, the similarities and the differences between these two theorems will be elaborated. First, note that Theorem 3 introduces a lower bound on the sample size. This is to ensure that the random matrix $(\frac{1}{d}X_{\mathcal{A}_1}^TX_{\mathcal{A}_1})^{-1}$ does not deviate excessively from $(\Sigma_{\mathcal{A}_1,\mathcal{A}_1})^{-1}$. Furthermore, notice that similar lower bounds (modulo constant factor) are required on the regularization coefficient $\lambda_d$ for both deterministic and random design matrices. Next, consider the probability (22) and notice that this is more conservative than its counterpart in Theorem 1. Again, this is due to the random nature of $X$. Finally, consider the upper bound on the estimation error (23). Observe that this upper bound has an additional term $\lambda_d\cdot\frac{8k_1}{\Lambda_{\min}^r}\sqrt{\frac{r_1p_{\max}}{d}}$ compared to (13), which accounts for the deviation of different norms of $X$ from their mean values. The following corollary follows from Theorem 3.

**Corollary 2.** *Assuming that $p_{\max}=O(\bar{p})$ and $d=\Omega\left(k_1r_1p_{\max}(k_1+r_1p_{\max}+\log\bar{p})\right)$, the inequality*

$$\|\hat{\Theta}-\Theta^*\|_\infty \leq O\left(\sqrt{\frac{(r_1p_{\max})^2+r_1p_{\max}\log\bar{p}}{d}}\right) \tag{24}$$

*holds with probability of at least* (22)*.*

Corollary 2 clearly shows the effect of a random design matrix on the estimation error bound. In particular, recall that when the design matrix is deterministic, the error rate 14 holds independent of the number of samples $d$. However, the randomness in the design matrix gives rise to a lower bound on the sample size to achieve the same error rate. Finally, the counterpart of Theorem 2 is presented for problems with random design matrices and general values of $\bar{r}$. To this goal, a slight modification is needed in Assumption 3.

**Assumption 4.** *There exist $\alpha,\delta_{\min}>0$ and $\beta>1$ such that*

$$\bar{r}=\mathcal{O}(\bar{p}^\alpha) \tag{25}$$

$$k_{\min}r_{\min}p_{\max}=\Omega(\bar{p}^{\delta_{\min}}) \tag{26}$$

$$\min_{j=1,\ldots,\bar{r}}|\mathcal{A}_j|=\Omega\left((\log\bar{p})^\beta\right) \tag{27}$$

Compared to Assumption 3, an additional lower bound is added to the minimum number of nonzero blocks in the block columns of $\Theta^*$.

**Theorem 4** (Random $X$, multiple block columns)**.** *Given arbitrary constants $c_1,c_2>\alpha+1$ and $c_3>\frac{\alpha}{2\delta_{\min}}+1$, suppose that each row of $X$ is drawn independently from $N(0,\Sigma)$, and that $\lambda_d$ and $d$ are chosen such that*

$$\lambda_d \geq \sqrt{\frac{32c_1\sigma_w^2\sigma_{\max}^2}{\gamma_r^2}\cdot\frac{(r_{\max}p_{\max})^2+r_{\max}p_{\max}\log\bar{p}}{d}} \tag{28}$$

$$d \geq \frac{72c_2\sigma_{\max}^2k_{\max}r_{\max}p_{\max}}{\gamma_r^2\Lambda_{\min}^r}\cdot(r_{\max}p_{\max}+\log\bar{p}) \tag{29}$$

*Then, with probability of at least*

$$1-O\left(\frac{1}{\bar{p}^{c_1-\alpha-1}}+\frac{1}{\bar{p}^{c_2-\alpha-1}}+\frac{1}{\bar{p}^{2\delta_{\min}(c_3-1)-\alpha}}\right)\to 1 \tag{30}$$

*the following statements hold:*

1. *$\hat{\Theta}$ is unique and, in addition,*

$$\begin{aligned}\|\hat{\Theta}-\Theta^*\|_\infty \leq&\sqrt{\frac{36c_3\sigma_w^2\log(k_{\max}r_{\max}p_{\max})}{\Lambda_{\min}^r d}}\\&+\lambda_d\left(\frac{8k_{\max}}{\Lambda_{\min}^r}\sqrt{\frac{r_{\max}p_{\max}}{d}}+q_r\right)=g_4\end{aligned} \tag{31}$$

2. *The set of nonzero blocks of $\hat{\Theta}$ excludes the zero blocks of $\Theta^*$. Furthermore, the nonzero blocks of $\hat{\Theta}$ and $\Theta^*$ are equivalent if $\min_{1\leq j\leq\bar{r}}\min_{i\in\mathcal{A}_j}\|\Theta^{(i,j)}\|_\infty>g_4$.*

Similar to Theorem 2, the lower bound on $c$ is to ensure that the probability of success converges to 1 as $\bar{p}$ increases. Furthermore, it is easy to verify that, due to (27), the decay rate of the last term in (22) is faster than the remaining terms and, hence, it vanishes in the big-$O$ analysis of Theorem 4.

## IV. Numerical Results

In this section, we demonstrate the accuracy of the block-regularized estimator on synthetically generated case studies and under different scenarios. The reported results are for a serial implementation in MATLAB. The `PQN` package from [17] is used to solve the block-regularization problem. Define *mismatch error* as the total number of false positives and false negatives in the blocks of the block-regularized estimator, i.e., the total number of blocks that are incorrectly identified as zero or nonzero in $\hat{\Theta}$. Furthermore, define *estimation error* as the 2-norm of $\hat{\Theta} - \Theta^*$, *relative number of samples* (RNS) as the sample size $d$ normalized by the dimension $p$, and *relative mismatch error* (RME) as the mismatch error normalized by total number of blocks in $\Theta$.

Given the integer numbers $p$, $r$, $w$, and $n$, the true regression matrix $\Theta^* \in \mathbb{R}^{p\times r}$ is constructed as follows: the blocks of $\Theta^*$ are defined by partitioning rows and columns of $\Theta^*$ into $\bar{p} = p/w$ and $\bar{r} = r/w$ equally-sized groups, respectively (for simplicity, it is assumed that $p$ and $r$ are divisable by $r$). Based on this construction, each block of $\Theta^*$ is a square submatrix of size $w$. At each block column of $\Theta^*$, the elements within $n$ randomly chosen blocks are set to $0.3$ or $-0.3$, with equal probability and the remaining blocks are set to zero. Furthermore, each row of the design matrix is randomly drawn from a standard normal distribution. Clearly, $\Sigma = I$ and it satisfies Assumption 2. To verify the presented theoretical results, $\lambda_d$ is set to

$$2.2\sqrt{\frac{w^4 + w^2\log(\bar{p})}{d}} \tag{32}$$

in all of the experiments. Note that this choice of $\lambda_d$ is at most a constant factor away from (28).

In the first set of experiments, the performance of the block-regularized estimator is showcased for different values of $p$ and $n$. In particular, suppose that $r = 25$, $w = 5$, and that the tuple $(p, n)$ is chosen from the set $\{(500, 3), (1000, 5), (1500, 8), (2000, 10)\}$. Figure 1a depicts the mismatch error with respect to the sample size for different values of $(p, n)$. It can be observed that, as the dimension of the system increases, a higher number of samples is required to obtain a small mismatch error. Conversely, the required value of RNS to achieve a small RME reduces as the dimension of the problem grows. More precisely, RNS should be at least $0.50$, $0.20$, $0.17$, and $0.15$ to guarantee RME $\leq 0.2\%$, when $(p, n)$ is equal to $(500, 3)$, $(500, 5)$, $(500, 8)$, and $(500, 10)$, respectively. Figure 1b shows the estimation error with respect to the sample size for different configurations of $(p, n)$. Not surprisingly, the overall estimation error decreases as the number of samples grows. Furthermore, as $(p, n)$ increases, a significantly higher number of samples is needed to achieve the same estimation error.

In the second set of experiments, the performance of the proposed block-regularized estimator is compared to the well-known least-squares estimator, defined as

$$\bar{\Theta} = \arg\min_{\Theta} \|Y - X\Theta\|_F^2 \tag{33}$$

It is straightforward to verify that 33 has the closed-form solution $\bar{\Theta} = (X^T X)^{-1} X^T Y$ and is not defined for $d < p$. Upon fixing $p$ and $r$ to the respective numbers $900$ and $30$, the accuracy of the block-regularized and least-squares estimators are depicted in Figure 1c for $w = 3, 5, 6$. As can be noticed in this figure, the block-regularized estimator significantly outperforms the least-squares one in terms of the estimation error. Furthermore, the least-squares estimator always results in a fully dense matrix and is undefined for $d < 900$.

## V. Proofs

In this section, we only present the proof of Theorem 3 due to space restrictions. The proof of Theorem 1 is a simplified version of the arguments made in the sequel. Furthermore, the proofs of Theorems 2 and 4 follow from union bounds applied to the results of Theorems 1 and 3, respectively. A number of preliminary definitions and lemmas are required to present the proof of Theorem 3. The proofs of lemmas are omitted for brevity.

**Definition 1** (sub-Gaussian random variable)**.** A zero-mean random variable $x$ is *sub-Gaussian* with parameter $\sigma^2$ if there exists a constant number $c < \infty$ such that

$$\mathbb{P}(|x| > t) \leq c \cdot \exp(-\frac{t^2}{2\sigma^2}) \tag{34}$$

**Lemma 1.** *Given a set of zero-mean sub-Gaussian variables $x_i$ with parameters $\sigma_i$ for $i = 1, 2, ..., m$, the inequality*

$$\mathbb{P}(\max_i |x_i| > t) \leq c \cdot \exp(-\frac{t^2}{2\min_i \sigma_i^2} + \log m) \tag{35}$$

*holds for some constant $c < \infty$.*

Define $I_d$ as the $d \times d$ identity matrix. The next lemma is borrowed from [14].

**Lemma 2.** *Given a set of random vectors $X_i \sim N(0, \sigma_i^2 I_d)$ for $i = 1, 2, ..., m$ and $d > 2\log m$, the inequality*

$$\mathbb{P}(\max_i \|X_i\|_2^2 \geq 4\sigma^2 d) \leq \exp(-\frac{d}{2} + \log m) \tag{36}$$

*holds, where $\sigma = \max_i \sigma_i$.*

The following lemma is well-known and can be found in [6].

**Lemma 3.** *Consider a matrix $X \in \mathbb{R}^{m\times n}$ whose rows are drawn from $N(0, \Sigma)$. Assuming that $n \leq m$, we have*

$$\mathbb{P}\left(\|(\frac{1}{d}X^T X)^{-1} - \Sigma^{-1}\|_2 \geq \frac{8}{\Lambda_{\min}^r}\sqrt{\frac{n}{m}}\right) \leq 2\exp\left(-\frac{n}{2}\right) \tag{37}$$

The following basic inequalities will be used frequently in our subsequent arguments:

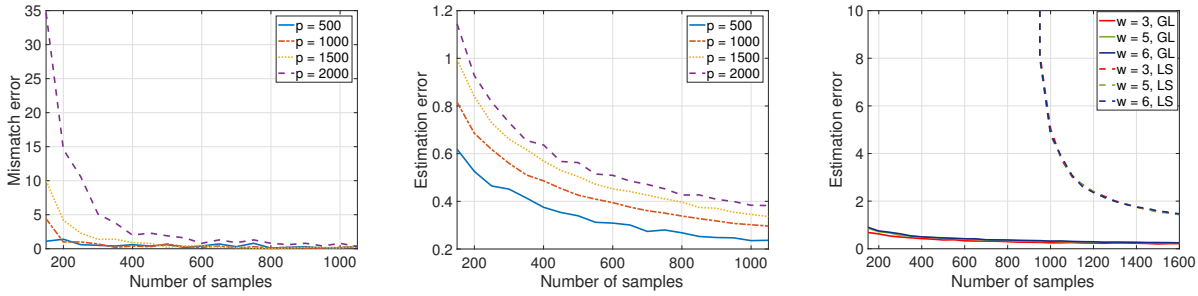
**Lemma 4.** *The following statements hold true:*

- *Given a number of (not necessarily independent) events $\mathcal{T}_i$ for $i = 1, 2, ..., n$, the following inequality is satisfied:*

$$\sum_{i=1}^{n} \mathbb{P}(\mathcal{T}_i) - (n-1) \leq \mathbb{P}(\mathcal{T}_1 \cap \mathcal{T}_2 \cap ... \cap \mathcal{T}_n) \tag{38}$$

- *Given events $\mathcal{B}$ and $\mathcal{C}$ together with the complement of $\mathcal{C}$, denoted as $\mathcal{C}^c$, the following inequality holds:*

$$\mathbb{P}(\mathcal{B}) \leq \mathbb{P}(\mathcal{B}|\mathcal{C}) + \mathbb{P}(\mathcal{C}^c) \tag{39}$$




(a) Mismatch error with varying $(p, n)$ (b) Estimation error with varying $(p, n)$ (c) Estimation error with varying $w$

Fig. 1: (a) The mismatch error with respect to the sample size for different values of $(p, n)$, (b) the estimation error with respect to the sample size for different values of $(p, n)$, (c) the estimation error of block-regularized estimator compared to its least-squares counterpart with respect to the sample size and for different values of $w$.

The next lemma characterizes the first-order optimality conditions for 3.

**Lemma 5** (KKT conditions)**.** *$\tilde{\Theta}$ is an optimal solution for* (3) *if and only if it satisfies*

$$\frac{1}{d}X^T X(\tilde{\Theta} - \Theta^*) - \frac{1}{d}X^T W + \lambda_d \tilde{S} = 0 \tag{40}$$

*for some $\tilde{S} \in \mathbb{R}^{p\times r} \in \partial\|\tilde{\Theta}\|_{\text{block}}$, where $\partial\|\tilde{\Theta}\|_{\text{block}}$ denotes the sub-differential of $\|\cdot\|_{\text{block}}$ at $\tilde{\Theta}$.*

Since $\tilde{S}$ and $W$ have the same dimension as $\tilde{\Theta}$, $\tilde{S}^{(i)}$ and $W^{(i)}$ are used to denote the blocks of $\tilde{S}$ and $W$ corresponding to $\Theta^{(i)}$, respectively. Note that in Theorem 3, it is assumed that $\bar{r} = 1$. Therefore, in order to streamline the presentation, $\Theta^{(i,1)}$ and $\mathcal{A}_1$ will be referred to as $\Theta^{(i)}$ and $\mathcal{A}$, respectively.

**Lemma 6.** *$Q \in \partial\|\tilde{\Theta}\|_{\text{block}}$ if and only if the following conditions are satisfied for every $i = 1, 2, ..., \bar{p}$:*

- *If $\|\tilde{\Theta}^{(i)}\|_\infty \neq 0$, define $M^{(i)} = \{(k,l) : \tilde{\Theta}^{(i)}_{kl} = \|\tilde{\Theta}^{(i)}\|_\infty\}$. Then, $Q^{(i)}_{kl} = \eta_{kl} \cdot \text{sign}(\tilde{\Theta}^{(i)}_{kl})$, where $\sum_{(k,l)\in M^{(i)}} \eta_{kl} = 1$ and $\eta_{kl} = 0$ if $(k,l) \notin M^{(i)}$.*
- *If $\|\tilde{\Theta}^{(i)}\|_\infty = 0$, then $\|Q^{(i)}\|_1 \leq 1$.*

The proof for Theorem 2 is based on the primal-dual witness approach introduced in [6], [14], which is defined as follows:

**Primal-dual witness approach( [6], [14]):**

- *Step 1:* Define the restricted regularized problem as

$$\tilde{\Theta} = \arg\min_{\Theta\in\mathbb{R}^{p\times r}} \quad \frac{1}{2d}\|Y - X\Theta\|_F^2 + \lambda_d\|\Theta\|_{\text{block}} \tag{41a}$$

$$\text{s.t.} \quad \Theta^{(i)} = 0 \quad \forall i \in \mathcal{A}^c \tag{41b}$$

  whose solution is unique if $X^T_{\mathcal{A}} X_{\mathcal{A}}$ is invertible.
- *Step 2:* With a slight abuse of notation, $\tilde{\Theta}$ can be written as $(\tilde{\Theta}_{\mathcal{A}}, 0)$. Choose $\tilde{S}_{\mathcal{A}}$ as an element of the sub-differential $\partial\|\tilde{\Theta}_{\mathcal{A}}\|_{\text{block}}$.
- *Step 3:* Find $\tilde{S}^c_{\mathcal{A}}$ by solving the KKT equations (40), given $\tilde{\Theta}$ and $\tilde{S}_{\mathcal{A}}$. Then, verify

$$\|\tilde{S}^{(i)}\|_1 < 1 \quad \forall i \in \mathcal{A}^c \tag{42}$$

If (42) can be verified in the last step, it is said that the primal-dual witness approach (PDW) *succeeds*. Next lemma unveils a close relationship between the block-regularized estimator, the PDW approach, and the true regression parameter $\Theta^*$.

**Lemma 7.** *The following statements hold:*

- *If the PDW approach succeeds, then $\tilde{\Theta}$ is the unique optimal solution of* (3)*, i.e. $\hat{\Theta} = \tilde{\Theta}$.*
- *Conversely, suppose that $\hat{\Theta}$ is the optimal solution of* (3) *such that $\hat{\Theta}^{(i)} = 0$ for every $i \in \mathcal{A}^c$. Then, the PDW approach succeeds.*

Lemma 7 is the building block of our subsequent arguments. Based on this lemma, it is enough to verify that the PDW approach succeeds with high probability in order to show the consistency of the proposed estimator. Upon proving this statement, our focus can be devoted to the optimal solution of the restricted problem (41) and bounding its difference from the true parameter.

**Lemma 8.** *Define $\tilde{\Theta} - \Theta = E$. The following equalities hold:*

$$E_{\mathcal{A}^c} = 0 \tag{43}$$

$$E_{\mathcal{A}} = (\frac{1}{d}X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}\frac{1}{d}X^T_{\mathcal{A}}W - (\frac{1}{d}X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}\lambda_d\tilde{S}_{\mathcal{A}} \tag{44}$$

$$\tilde{S}_{\mathcal{A}^c} = \frac{1}{d\lambda_d}\left(X^T_{\mathcal{A}^c} - (X^T_{\mathcal{A}^c}X_{\mathcal{A}})(X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}X^T_{\mathcal{A}}\right)W + \frac{1}{d}X^T_{\mathcal{A}^c}X_{\mathcal{A}}(\frac{1}{d}X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}} \tag{45}$$

*Proof of Theorem 2:* First, we show that the PDW succeeds with high probability. This immediately implies that the nonzero blocks of $\hat{\Theta}$ belong to $\mathcal{A}$.

**1. Success of PDW approach:** It is enough to show that $\max_{i\in\mathcal{A}^c}\|\tilde{S}^{(i)}\|_1 < 1$ with high probability. Lemma 8 yields that

$$\|\tilde{S}^{(i)}\|_1 \leq \underbrace{\left\|\frac{1}{d\lambda_d}\left({X_{(i)}}^T - ({X_{(i)}}^T X_{\mathcal{A}})(X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}X^T_{\mathcal{A}}\right)W^{(i)}\right\|_1}_{Z_1^{(i)}} + \underbrace{\left\|\frac{1}{d}{X_{(i)}}^T X_{\mathcal{A}}(\frac{1}{d}X^T_{\mathcal{A}}X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}\right\|_1}_{Z_2^{(i)}} \tag{46}$$

where $\tilde{S}_{\mathcal{A}}$ is obtained by removing the blocks of $\tilde{S}$ with indices not belonging to $\mathcal{A}$. We will show that $\max_{i\in\mathcal{A}^c} Z_1^{(i)} < \gamma_r/2$ and $\max_{i\in\mathcal{A}^c} Z_2^{(i)} < 1-\gamma_r/2$ with high probability. First, consider $\max_{i\in\mathcal{A}^c} Z_1^{(i)}$. We have

$$Z_1^{(i)} = \sum_{(k,l)\in\Theta^{(i)}} \Big| \underbrace{\frac{1}{d\lambda_d}(X_{(i)})_{:,k}{}^T (I - X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T) W_{:,l}}_{R_{kl}^{(i)}} \Big| \tag{47}$$

Given $X$, $R_{kl}^{(i)}$ is Gaussian with variance

$$\frac{\sigma_w^2}{d^2\lambda_d^2}\Big((X_{(i)})_{:,k}{}^T (I - X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T)^2 (X_{(i)})_{:,k}\Big) \tag{48}$$

Note that $X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T$ is an orthogonal projection onto the range of $X_A$. Therefore,

$$\begin{aligned}
&\frac{\sigma_w^2}{d^2\lambda_d^2}\Big((X_{(i)})_{:,k}{}^T (I - X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T)^2 (X_{(i)})_{:,k}\Big) \\
&= \frac{\sigma_w^2}{d^2\lambda_d^2}\Big((X_{(i)})_{:,k}{}^T (I - X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T)(X_{(i)})_{:,k}\Big) \\
&\le \frac{\sigma_w^2}{d^2\lambda_d^2}\|(X_{(i)})_{:,k}\|_2^2
\end{aligned} \tag{49}$$

Due to Lemma 2, the last inequality is upper bounded by $4\sigma_w^2\sigma_{\max}^2/d\lambda_d^2$ for every $k$ with probability of at least $1-\exp(-d/2+\log p)$ for $d > 2\log p$. Conditioned on this event, one can write

$$Z_1^{(i)} = \max_{\epsilon\in\{-1,+1\}^{p_i\times r_1}} \sum_{(k,l)\in\Theta^{(i)}} \epsilon_{kl} R_{kl}^{(i)} \tag{50}$$

which means that $\sum_{(k,l)\in\Theta^{(i)}} \epsilon_{kl} R_{kl}^{(i)}$ is sub-Gaussian with parameter $4p_i r\sigma_w^2\sigma_{\max}^2/d\lambda_d^2$. This yields that

$$\begin{aligned}
\mathbb{P}(\max_{i\in\mathcal{A}^c} Z_1^{(i)} \ge \zeta) &= \mathbb{P}(\max_{i\in\mathcal{A}^c} \max_{\epsilon\in\{-1,+1\}^{p_i\times r}} \sum_{(k,l)\in\Theta^{(i)}} \epsilon_{kl} R_{kl}^{(i)} \ge \zeta) \\
&\le 2\exp(-\frac{d\lambda_d^2\zeta^2}{8p_{\max} r\sigma_w^2\sigma_{\max}^2} + p_{\max} r_1 + \log\bar{p}) \\
&\quad + \exp(-d/2+\log\bar{p})
\end{aligned} \tag{51}$$

where we have used Lemma 1, the second statement of Lemma 4 and the fact that $p_i \le p_{\max}$ in the last inequality. Now, setting $\zeta = \gamma_r/2$ and

$$\lambda_d \ge \sqrt{\frac{32c_1\sigma_w^2\sigma_{\max}^2}{\gamma_r^2}\cdot\frac{(p_{\max}r_1)^2 + p_{\max}r_1\log\bar{p}}{d}} \tag{52}$$

for some arbitrary constant $c_1 > 1$ yields that

$$\begin{aligned}
\mathbb{P}(\max_{i\in\mathcal{A}^c} Z_1^{(i)} < \gamma_r/2) &\ge 1 - 2\exp(-(c_1-1)(p_{\max}r_1+\log\bar{p})) \\
&\quad - \exp(-d/2+\log\bar{p}) \\
&\ge 1 - 3\exp(-(c_1-1)(p_{\max}r_1+\log\bar{p}))
\end{aligned} \tag{53}$$

where the last inequality is due to the lower bound (21) on $d$. Next, an upper bound on $\max_{i\in\mathcal{A}^c} Z_2^{(i)}$ will be derived. Since each row of $X$ is drawn from $N(0,\Sigma)$, one can write the distribution of $X_{\mathcal{A}^c}^T$, conditioned on $X_{\mathcal{A}}$ as

$$N\left(\Sigma_{\mathcal{A}^c,\mathcal{A}}(\Sigma_{\mathcal{A},\mathcal{A}})^{-1} X_{\mathcal{A}}^T, \Sigma_{\mathcal{A}^c,\mathcal{A}^c} - \Sigma_{\mathcal{A}^c,\mathcal{A}}(\Sigma_{\mathcal{A},\mathcal{A}})^{-1}\Sigma_{\mathcal{A},\mathcal{A}^c}\right) \tag{54}$$

Based on 54, one can verify that $\frac{1}{d}X_{\mathcal{A}^c}^T X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}$ has the same distribution as

$$\Sigma_{\mathcal{A}^c,\mathcal{A}}(\Sigma_{\mathcal{A},\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}} + \frac{1}{d}V^T X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}} \tag{55}$$

where $V$ is a random matrix with the same size as $X_{\mathcal{A}}$ and independent of it, whose elements are sub-Gaussian with parameters of at most $\sigma_{\max}^2$. This implies that

$$\begin{aligned}
\max_{i\in\mathcal{A}^c} Z_2^{(i)} &\le \max_{i\in\mathcal{A}^c} \left\|\Sigma_{i,\mathcal{A}}(\Sigma_{\mathcal{A},\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}\right\|_1 \\
&\quad + \max_{i\in\mathcal{A}^c} \left\|\frac{1}{d}V_{(i)}{}^T X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}\right\|_1 \\
&\le 1-\gamma_r + \max_{i\in\mathcal{A}^c} \underbrace{\left\|\frac{1}{d}V_{(i)}{}^T X_{\mathcal{A}}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\tilde{S}_{\mathcal{A}}\right\|_1}_{Z_3^{(i)}}
\end{aligned} \tag{56}$$

where we have used the mutual incoherence property and the fact that $\|\tilde{S}_{\mathcal{A}}\|_\infty \le 1$. Now, it remains to show that $\max_{i\in\mathcal{A}^c} Z_3^{(i)} < \gamma_r/2$ with high probability. First, Lemma 3 is used to bound the 2-norm of $\frac{1}{d}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}$:

$$\begin{aligned}
\left\|\frac{1}{d}(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\right\|_2 &\le \frac{1}{d}\cdot\frac{8}{\Lambda_{\min}^r}\sqrt{\frac{|I(\mathcal{A})|}{d}} + \frac{1}{d}\cdot\left\|\Sigma_{\mathcal{A},\mathcal{A}}^{-1}\right\|_2 \\
&\le \frac{1}{d}\cdot\frac{8}{\Lambda_{\min}^r}\sqrt{\frac{|I(\mathcal{A})|}{d}} + \frac{1}{d}\cdot\frac{1}{\Lambda_{\min}^r} \\
&\le \frac{9}{\Lambda_{\min}^r d}
\end{aligned} \tag{57}$$

with probability of at least $1-2\exp(-|I(\mathcal{A})|/2)$. Similar to the arguments made for bounding $\max_{i\in\mathcal{A}^c} Z_1^{(i)}$, one can verify that

$$\begin{aligned}
\mathbb{P}(\max_{i\in\mathcal{A}^c} Z_3^{(i)} < \gamma_r/2) \ge 1 - 2\exp\Big(-\frac{\Lambda_{\min}^r d\gamma_r^2}{72\sigma_{\max}^2 k_{\max}p_{\max}r_1} + p_{\max}r_1 \\
+\log\bar{p}\Big) - 2\exp\Big(-\frac{|I(\mathcal{A})|}{2}\Big)
\end{aligned} \tag{58}$$

Now, choosing

$$d \ge \frac{72c_2\sigma_{\max}^2 k_{\max}p_{\max}r_1}{\Lambda_{\min}^r\gamma_r^2}\cdot(p_{\max}r_1+\log\bar{p}) \tag{59}$$

for some arbitrary constant $c_2 > 1$ results in

$$\begin{aligned}
\mathbb{P}(\max_{i\in\mathcal{A}^c} Z_3^{(i)} < \gamma_r/2) &\ge 1 - 2\exp(-(c_2-1)(p_{\max}r_1+\log\bar{p})) \\
&\quad - 2\exp\Big(-\frac{|I(\mathcal{A})|}{2}\Big)
\end{aligned} \tag{60}$$

Therefore, it is shown that $\max_{i\in\mathcal{A}^c}\|\tilde{S}^{(i)}\|_1 < 1$ and, hence, PDW succeeds with probability that is lower bounded by (18).

**2. Bounding the error:** In order to bound the estimation error, an upper bound on $\|E\|_\infty$ will be derived, conditioning on the success of the PDW approach. Note that $E_{\mathcal{A}^c} = 0$ according to Lemma 8 and, hence, it suffices to bound $\|E_{\mathcal{A}}\|_\infty$. Again, due to Lemma 8, one can write

$$\begin{aligned}
\max_{k=1,\dots,r}\|(E_{\mathcal{A}})_{:,k}\|_\infty &\le \max_{k=1,\dots,r} \underbrace{\left\|(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\frac{1}{d}X_{\mathcal{A}}^T W_{:,k}\right\|_\infty}_{Z_4^k} \\
&\quad + \max_{k=1,\dots,r} \underbrace{\left\|(\frac{1}{d}X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}\lambda_d(\tilde{S}_{\mathcal{A}})_{:,k}\right\|_\infty}_{Z_5^k}
\end{aligned} \tag{61}$$

for $k = 1, 2, ..., r$. For bounding $Z_5^k$, it can be argued similar to (57) that

$$\begin{aligned}\max_{k=1,\dots,r} Z_5^k \leq & \max_{k=1,\dots,r} \left\| \left( (\frac{1}{d} X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} - \Sigma_{\mathcal{A},\mathcal{A}}^{-1} \right) \lambda_d (\tilde{S}_{\mathcal{A}})_{:,k} \right\|_\infty \\ & + \max_{k=1,\dots,r} \left\| \Sigma_{\mathcal{A},\mathcal{A}}^{-1} \lambda_d (\tilde{S}_{\mathcal{A}})_{:,k} \right\|_\infty \\ \leq & \left\| (\frac{1}{d} X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} - \Sigma_{\mathcal{A},\mathcal{A}}^{-1} \right\|_2 \lambda_d \sqrt{k_{\max}} + q_r \lambda_d \\ \leq & \lambda_d \left( \frac{8 k_{\max}}{\Lambda_{\min}^r} \sqrt{\frac{p_{\max} r_1}{d}} + q_r \right) \end{aligned} \tag{62}$$

with probability of at least $1 - 2\exp(-|I(\mathcal{A})|/2)$, where we have used the matrix norm properties and the fact that $|I(\mathcal{A})| \leq p_{\max} k_{\max} r_1$. Now, it remains to bound $\max_{k=1,\dots,r} Z_4^k$. This can be done similar to the previous arguments, i.e., by making use of (57) and obtaining a sub-Gaussian parameter for $(\frac{1}{d} X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} \frac{1}{d} X_{\mathcal{A}}^T W_{:,k}$. For brevity, only the final key inequality is stated below:

$$\begin{aligned}\mathbb{P}\big( \max_{k=1,\dots,r} Z_4^k \geq \zeta \big) \leq & \, 2\exp\Big( -\frac{d\Lambda_{\min}^r \zeta^2}{18\sigma_w^2} + \log r_1 \\ & + \log(p_{\max} k_{\max} r_1) \Big) + 2\exp\Big( -\frac{|I(\mathcal{A})|}{2} \Big) \end{aligned} \tag{63}$$

Now, setting

$$\zeta = \sqrt{\frac{36 c_3 \sigma_w^2 \log(p_{\max} k_{\max} r_1)}{d\Lambda_{\min}^r}} \tag{64}$$

for an arbitrary constant $c_3 > 1$, together with the inequality $\log r_1 \leq \log(p_{\max} k_{\max} r_1)$ leads to

$$\max_{k=1,\dots,r} Z_4^k \leq \sqrt{\frac{36 c_3 \sigma_w^2 \log(p_{\max} k_{\max} r_1)}{d\Lambda_{\min}^r}} \tag{65}$$

with probability of at least

$$1 - 2\exp\left(-2(c_3 - 1)\log(p_{\max} k_{\max} r_1)\right) - 2\exp\Big( -\frac{|I(\mathcal{A})|}{2} \Big) \tag{66}$$

Combining this inequality with (62) results in the elementwise error bound 19 with probability of at least 18. This concludes the proof. □

## VI. Conclusion

This paper is concerned with the linear regression problem, where the unknown regression matrix possesses a block sparsity structure. Given a limited number of observed samples, the objective is to obtain an estimate of the regression matrix by taking into account its block-sparse structure. To this goal, a block-regularized estimator is considered and its high-dimensional properties are analyzed in this work. In particular, existing results on this estimator are generalized to the case where the regression matrix has an arbitrary number of blocks, each with an arbitrary size. For problems with deterministic design matrices, a sharp upper bound on the elementwise error of the proposed estimator is derived that depends on the sample size and the dimension of the problem, as well as the sparsity level of the true regression matrix and the maximum size of its blocks. Furthermore, it is proven that the proposed estimator benefits from similar error rates for problems with randomly generated design matrices, provided that the sample size is higher than a threshold. The developed theoretical results are demonstrated on different test cases with various parameters.

## References


[1] P. E. Vértes, A. F. Alexander-Bloch, N. Gogtay, J. N. Giedd, J. L. Rapoport, and E. T. Bullmore, "Simple models of human brain functional networks," *Proceedings of the National Academy of Sciences*, vol. 109, no. 15, pp. 5868–5873, 2012.

[2] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *Journal of Transportation Engineering*, vol. 138, no. 11, pp. 1358–1367, 2012.

[3] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688*, 2017.

[4] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of statistics*, pp. 295–327, 2001.

[5] N. Kalouptsidis, G. Mileounis, B. Babadi, and V. Tarokh, "Adaptive algorithms for sparse system identification," *Signal Processing*, vol. 91, no. 8, pp. 1910–1919, 2011.

[6] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[7] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, no. Jun, pp. 1179–1225, 2008.

[8] A. Rinaldo *et al.*, "Properties and refinements of the fused lasso," *The Annals of Statistics*, vol. 37, no. 5B, pp. 2922–2952, 2009.

[9] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, "High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[10] K. Chernyshov, "Towards the knowledge-based multi-agent system identification," in *IEEE 10th Conference on Industrial Electronics and Applications*, 2015, pp. 399–404.

[11] S. Hassan-Moghaddam, N. K. Dhingra, and M. R. Jovanović, "Topology identification of undirected consensus networks via sparse inverse covariance estimation," in *IEEE 55th Conference on Decision and Control*, 2016, pp. 4624–4629.

[12] S. Tu, R. Boczar, A. Packard, and B. Recht, "Non-asymptotic analysis of robust control from coarse-grained identification," *arXiv preprint arXiv:1707.04791*, 2017.

[13] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine learning research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[14] S. N. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block $\ell_1/\ell_\infty$-regularization," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3841–3863, 2011.

[15] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, pp. 1436–1462, 2006.

[16] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[17] M. Schmidt, E. van den Berg, M. P. Friedlander, and K. Murphy, "Pqn: Optimizing costly functions with simple constraints," 2009. [Online]. Available: https://www.cs.ubc.ca/~schmidtm/Software/PQN.html