GEP-MSCRA for computing the group zero-norm regularized least squares estimator[*]

Shujun Bi[†] and Shaohua Pan[‡]

December 29, 2017

Abstract

This paper concerns with the group zero-norm regularized least squares estimator which, in terms of the variational characterization of the zero-norm, can be obtained from a mathematical program with equilibrium constraints (MPEC). By developing the global exact penalty for the MPEC, this estimator is shown to arise from an exact penalization problem that not only has a favorable bilinear structure but also implies a recipe to deliver equivalent DC estimators such as the SCAD and MCP estimators. We propose a multi-stage convex relaxation approach (GEP-MSCRA) for computing this estimator, and under a restricted strong convexity assumption on the design matrix, establish its theoretical guarantees which include the decreasing of the error bounds for the iterates to the true coefficient vector and the coincidence of the iterates after finite steps with the oracle estimator. Finally, we implement the GEP-MSCRA with the subproblems solved by a semismooth Newton augmented Lagrangian method (ALM) and compare its performance with that of SLEP and MALSAR, the solvers for the weighted $\ell_{2,1}$-norm regularized estimator, on synthetic group sparse regression problems and real multi-task learning problems. Numerical comparison indicates that the GEP-MSCRA has significant advantage in reducing error and achieving better sparsity than the SLEP and the MALSAR do.

Keywords: group sparse regression; group zero-norm; global exact penalty; GEP-MSCRA

1 Introduction

Many regression and learning problems aim at finding important explanatory factors in predicting the response variable, where each explanatory factor may be represented by a group of derived input variables (see, e.g., [9, 35, 14, 20, 25, 36, 12, 22]). The most common example is the multifactor analysis-of-variance problem, in which each factor

[*] Supported by the National Natural Science Foundation of China under project No.11571120 and No.11701186, and the Natural Science Foundation of Guangdong Province under project No.2015A030313214 and No.2017A030310418.

[†] bishj@scut.edu.cn. School of Mathematics, South China University of Technology, Guangzhou.

[‡] shhpan@scut.edu.cn. School of Mathematics, South China University of Technology, Guangzhou.

may have several levels and can be expressed through a group of dummy variables. Let $\mathcal{J}_1,\mathcal{J}_2,\ldots,\mathcal{J}_m$ be a collection of index sets to represent the group structure of explanatory factors, where $\mathcal{J}_i\cap\mathcal{J}_j=\emptyset$ for all $i\neq j\in\{1,\ldots,m\}$ and $\bigcup_{i=1}^m\mathcal{J}_i=\{1,2,\ldots,p\}$. This class of regression problems can be stated via the following observation model

$$b=\textstyle\sum_{i=1}^m A_{\mathcal{J}_i}\overline{x}_{\mathcal{J}_i}+\varepsilon, \tag{1}$$

where $\overline{x}\in\mathbb{R}^p$ is the true (unknown) coefficient vector, $A_{\mathcal{J}_i}$ $(i=1,\ldots,m)$ is an $n\times|\mathcal{J}_i|$ design matrix corresponding to the ith factor, and $\varepsilon\in\mathbb{R}^n$ is the noise vector. Clearly, when $\mathcal{J}_i=\{i\}$ for $i=1,2,\ldots,m$, (1) reduces to the common linear regression model.

Sparse estimation, using penalization or regularization technique to perform variable selection and estimation simultaneously, has become a mainstream approach especially for high-dimensional data [6]. In the past decade, some popular penalized estimators were proposed one after another, including the ℓ_1-type estimators such as the Lasso [31] and the Dantzig [7], and the nonconvex penalized estimators such as the SCAD [10] and the MCP [39]. For the model (1), one may embrace the $\ell_{2,1}$-norm regularized estimator due to the simplicity of computation (see, e.g., [3, 35]), but this estimator inherits the bias of the Lasso. The major reason for this dilemma is the significant difference between the ℓ_1-norm and the zero-norm (or cardinality function). To enhance the quality of the l_1-type selector, some researchers focused on the estimator induced by nonconvex surrogates for the zero-norm regularized problem, such as the bridge [13], the SCAD [10] and the MCP [39]. In particular, some algorithms were also developed for computing these nonconvex penalized estimators [10, 41, 19, 4, 5]; for example, the local quadratic approximation (LQA) algorithm [10] and the local linear approximation approximation (LLA) algorithm [41]. Recently, Fan, Xue and Zou [11] also provided a unified theory to show explicitly how to obtain the oracle solution via the LLA algorithm for the class of folded concave penalized estimators, which covers the SCAD and MCP as special cases.

Let $A:=[A_{\mathcal{J}_1}\ A_{\mathcal{J}_2}\cdots A_{\mathcal{J}_m}]\in\mathbb{R}^{n\times p}$ and $\mathcal{G}(x):=\big(\|x_{\mathcal{J}_1}\|,\|x_{\mathcal{J}_2}\|,\ldots,\|x_{\mathcal{J}_m}\|\big)^{\mathbb{T}}$ for $x\in\mathbb{R}^p$. In this work, we are interested in the following group zero-norm regularized estimator

$$\widehat{x}\in\mathop{\arg\min}_{x\in\Omega}\left\{\frac{\nu}{2n}\|Ax-b\|^2+\|\mathcal{G}(x)\|_0\right\}, \tag{2}$$

where $\nu>0$ is the regularization parameter, $\|\cdot\|_0$ denotes the zero-norm of a vector, and $\Omega:=\{x\in\mathbb{R}^p\colon\|x\|_\infty\leq R\}$ for some $R>0$. Here, the simple constraint $x\in\Omega$ is imposed to (2) in order to guarantee that the group zero-norm estimator $\widehat{x}$ is well-defined. In fact, similar simple constraints are also used for the ℓ_1-regularized models (see [1]). The estimator $\widehat{x}$ may be unacceptable for many statisticians since, compared with the convex $\ell_{2,1}$-regularized estimator and the nonconvex SCAD and MCP penalized estimators, it seems that $\widehat{x}$ is unapproachable due to the combinatorial property of the zero-norm. The main motivation for us to study such an estimator comes from the following facts:

- **Good group sparsity and unbiasedness of $\widehat{x}$.** By the definition of $\|\cdot\|_0$, clearly, $\widehat{x}$ can automatically set small estimated coeffcients to be zero, which reduces well the model complexity. In addition, by following the analysis in [10, Section 2], $\widehat{x}$ is nearly unbiased when A is orthonormal and the true coefficients are not too small.

- **The estimator $\widehat{x}$ is the restriction of the SCAD and MCP over the ball Ω.** The SCAD and MCP estimators were well studied in the past ten years, but there are few works to discuss their relation with the zero-norm regularized estimator except that they are effective nonconvex surrogates of the latter. In Section 2, we shall show that the SCAD and MCP functions arise from the global exact penalty for the equivalent MPEC (mathematical program with equilibrium constraints) of (2), and so $\widehat{x}$ is the restriction of the SCAD and MCP estimators over the ball Ω.

- **Approachability of the estimator $\widehat{x}$.** As will be shown in Section 3, with the global exact penalty for the MPEC of (2) which is actually a primal-dual equivalent model of (2), there is a large space to design efficient algorithms for computing $\widehat{x}$.

Specifically, by means of the variational characterization of the zero-norm, the group zero-norm regularized problem (2) can be rewritten as an MPEC. We show that the penalty problem, yielded by moving the equilibrium constraint into the objective, is a global exact penalty for the MPEC in the sense that it has the same global optimal solution set as the MPEC does. Consequently, one may approach the estimator $\widehat{x}$ by solving a global exact penalization problem. This result is significant since, on one hand, the global exact penalty is an Lipschitz continuous optimization problem whose objective function possesses a structure favorable to the design of effective algorithms; and on the other hand, it provides a recipe to deliver equivalent DC (difference of convex functions) penalized functions whose global minimizers provide an estimator with three desirable properties stated in [10]; for example, the popular SCAD and MCP penalized estimators.

By the biconvex structure of the global exact penalty, we solve it in an alternating way and develop a multi-stage convex relaxation approach called GEP-MSCRA for computing $\widehat{x}$ (see Section 3). The GEP-MSCRA consists of solving a sequence of weighted $\ell_{2,1}$-norm regularized subproblems. In this sense, it is similar to the LLA algorithm [41] for the nonconcave penalized likelihood model. However, it is worth emphasizing that the start-up of the LLA algorithm depends implicitly on an initial estimator x^0, while the start-up of the GEP-MSCRA depends explicitly on a dual variable w^0. In addition, the involving subproblems may be different since the subproblems of the LLA are obtained from the primal angle, and those of the GEP-MSCRA are yielded from the primal-dual angle. For the proposed GEP-MSCRA, under a restricted strong convex (RSC) assumption on A, we verify in Section 4 that the error bounds of the iterates to the true $\overline{x}$ is decreasing as the number of iterates increases, and if the smallest nonzero group element of $\overline{x}$ is not too small, the iterates after finite steps will coincide with the oracle solution, and hence the support of $\overline{x}$ is exactly identified. Since the RSC assumption holds with high probability by [23], the GEP-MSCRA has the theoretical guarantees in a statistical sense.

We implement the GEP-MSCRA by solving the weighted $\ell_{2,1}$-norm regularized subproblems with a semismooth Newton ALM. The semismooth Newton ALM is a dual method that can solve the weighted $\ell_{2,1}$-norm regularized problems more efficiently than the existing first-order methods by exploiting the second-order information of the objective function in an economic way. As illustrated in [15, Section 3.3], the dual structure of the weighted $\ell_{2,1}$-norm regularized problems implies a nonsingular generalized Hessian matrix, which is well suitable for the semismooth Newton method. We compare the performance of the GEP-MSCRA with that of the SLEP and the MALSAR in [16], the solvers to the unconstrained weighted $\ell_{2,1}$-norm regularized least squares problems on synthetic group sparse regression problems and real multi-task learning problems, respectively. Numerical comparisons demonstrates that the GEP-MSCRA has a remarkable advantage in reducing error and achieving the exact group sparsity although it requires a little more time than the SLEP and the MALSAR do; for example, for the synthetic group sparse regression problems, the GEP-MSCRA reduces the relative recovery error of the SLEP at least 60% for the design matrix of Gaussian or sub-Gaussian type (see the first four subfigures in Figure 3), and for the real (School data) multi-task learning, the GEP-MSCRA can reduce the prediction error of the MALSAR at least 20% when there are more than 50% examples are used as training samples (see Figure 8).

To close this section, we introduce some necessary notations. We denote $\|A\|$ by the spectral norm and $\|A\|_{2,\infty}$ by the maximum column norm of A, respectively. Let e and I denote the vector of all ones and the identity matrix, respectively, whose dimensions are known from the context. For a convex function $g\colon \mathbb{R}\to(-\infty,+\infty]$, g^* denotes the conjugate of g; for a given closed set $S\subseteq\mathbb{R}^n$, $\delta_S(\cdot)$ means the indicator function over the set S, i.e., $\delta_S(x)=0$ if $x\in S$ and otherwise $\delta_S(x)=+\infty$; and for a given index set $F\subseteq\{1,\ldots,m\}$, write $F^c=\{1,\ldots,m\}\backslash F$ and $\mathcal{J}_F:=\bigcup_{i\in F}\mathcal{J}_i$, and denote by $\mathbb{I}_F(\cdot)$ the characteristic function of F, i.e., $\mathbb{I}_F(i)=1$ if $i\in F$ and otherwise $\mathbb{I}_F(i)=0$.

2 A new perspective on the estimator $\widehat{x}$

We shall examine the estimator $\widehat{x}$ from an equivalent MPEC of (2) and a global exact penalty of this MPEC, and conclude that $\widehat{x}$ can be obtained by solving an exact penalty problem which is constructed by moving the complementarity (or equilibrium) constraint into the objective of the MPEC. For convenience, we write $f(x):=\frac{1}{2n}\|Ax-b\|^2$ and denote by L_f the Lipschitz constant of f relative to the set Ω. One will see that the results of this section are also applicable to a general continuous loss function.

Let Φ denote the family of closed proper convex functions $\phi\colon\mathbb{R}\to(-\infty,+\infty]$ satisfying $[0,1]\subseteq\mathrm{int}(\mathrm{dom}\phi)$, $\phi(1)=1$ and $\phi(t_\phi^*)=0$ where t_ϕ^* is the unique minimizer of ϕ over $[0,1]$. Let $\bar{t}_\phi$ be the minimum element in $[t_\phi^*,1)$ such that $\frac{1}{1-t_\phi^*}\in\partial\phi(\bar{t}_\phi)$, where $\partial\phi$ is the subdifferential mapping of ϕ. The existence of such $\bar{t}_\phi$ is guaranteed by Lemma 4.

Now we show that with an arbitrary $\phi\in\Phi$, the problem (2) can be rewritten as an MPEC. Fix an arbitrary $z\in\mathbb{R}^m$ and $\phi\in\Phi$. By the definition of Φ, one may check that

$$\|z\|_0=\min_{w\in\mathbb{R}^m}\Big\{\textstyle\sum_{i=1}^m\phi(w_i)\colon \|z\|_1-\langle w,|z|\rangle=0,\ 0\le w\le e\Big\}. \tag{3}$$

This variational characterization of $\|\cdot\|_0$ means that the problem (2) is equivalent to

$$\min_{x\in\Omega,w\in\mathbb{R}^m}\Big\{\nu f(x)+\sum_{i=1}^m\phi(w_i)\colon\ 0\le w\le e,\ \langle e-w,\mathcal{G}(x)\rangle=0\Big\} \tag{4}$$

in the following sense: if x^* is globally optimal to (2), then $(x^*, \max(\mathrm{sign}(\|\mathcal{G}(x^*)\|), t_\phi^* e))$ is a global optimal solution of (4); and conversely, if (x^*, w^*) is a global optimal solution of (4), then x^* is globally optimal to (2) with $\|\mathcal{G}(x^*)\|_0 = \sum_{i=1}^m \phi(w_i^*)$. This means that the difficulty to compute the estimator $\widehat{x}$ comes from the following equilibrium condition

$$e - w \geq 0,\ \mathcal{G}(x) \geq 0 \ \text{ and } \ \langle e{-}w, \mathcal{G}(x)\rangle = 0. \tag{5}$$

Also, it is the equilibrium constraint to bring the bothersome nonconvexity of (2). Since the constraint set of (4) involves the equilibrium constraint (5), we call it an MPEC.

It is well known that the MPEC is a class of very difficult problems in optimization. In the past two decades, there was active research on its theory and algorithms especially for the one over the polyhedral cone, and the interested readers may refer to [18, 34]. We notice that most of existing algorithms are generic and inappropriate for solving (4). To handle the tough equilibrium constraint, we here consider its penalized version

$$\min_{x\in\Omega, w\in[0,e]} \Big\{\nu f(x) + \sum_{i=1}^m \phi(w_i) + \rho\langle e{-}w, \mathcal{G}(x)\rangle\Big\} \tag{6}$$

where $\rho > 0$ is the penalty factor. The following theorem states that (6) is a global exact penalty for (4) in the sense that it has the same global optimal solution set as (4) does.

Theorem 2.1 *Let $\phi \in \Phi$. Then, for every $\rho > \overline{\rho} = \nu L_f \frac{(1-t_\phi^*)\phi'_-(1)}{1-\bar{t}_\phi}$, we have $\mathcal{S}_\rho^* = \mathcal{S}^*$, where $\mathcal{S}_\rho^*$ is the global optimal solution set of (6) associated to ρ, and $\mathcal{S}^*$ is that of (4).*

The proof of Theorem 2.1 is included in Appendix B. From the proof, we see that the constraint $x \in \Omega$ in (2) is also necessary to establish the exact penalty for the MPEC. By Theorem 2.1 and the equivalence between (4) and (2), the estimator $\widehat{x}$ can be computed by solving a single penalty problem (6) associated to $\rho > \overline{\rho}$. Since $[0,1] \subseteq \mathrm{int}(\mathrm{dom}\phi)$, the function $\sum_{i=1}^m \phi(w_i)$ is Lipschitzian relative to $[0,e]$ by [28, Theorem 10.4], and so is the objective function of (6) relative to its feasible set $\Omega \times [0,e]$. Thus, compared with the discontinuous nonconvex problem (2), the problem (6) is at least an Lipschitz-type one, for which the Clarke generalized differential [8] can be used for its analysis.

Interestingly, the equivalence between (2) and (6) also implies a mechanism to yield equivalent DC surrogates for the group zero-norm $\|\mathcal{G}(x)\|_0$, and the popular SCAD function [10] and MCP function [39] are one of the products. Next we demonstrate this fact. For each $\phi \in \Phi$, let $\psi\colon \mathbb{R} \to (-\infty, +\infty]$ be the associated closed proper convex function:

$$\psi(t) := \begin{cases} \phi(t) & \text{if } t \in [0,1], \\ +\infty & \text{otherwise.} \end{cases} \tag{7}$$

By using the function ψ, the problem (6) can be rewritten in the following compact form

$$\min_{x\in\Omega, w\in\mathbb{R}^m} \Big\{\nu f(x) + \rho\|x\|_{2,1} + \sum_{i=1}^m \big[\psi(w_i) - \rho w_i \|x_{J_i}\|\big]\Big\}.$$

Together with the definition of conjugate functions and the above discussion, we have

$$\widehat{x} \in \mathop{\arg\min}_{x\in\Omega} \Big\{ f(x) + \frac{1}{\nu}\big[\rho\|x\|_{2,1} - \sum_{i=1}^{m} \psi^*(\rho\|x_{\mathcal{J}_i}\|)\big]\Big\} \quad \text{for } \rho > \overline{\rho}. \tag{8}$$

This means that the following function provides an equivalent DC surrogate for $\frac{1}{\nu}\|\mathcal{G}(x)\|_0$:

$$\Theta(x) := \frac{1}{\nu}\textstyle\sum_{i=1}^{m}\theta(\rho\|x_{\mathcal{J}_i}\|) \ \text{ with } \ \theta(s) := s - \psi^*(s). \tag{9}$$

In particular, when ϕ is chosen as the one in Example 2.1, it becomes the SCAD function. Indeed, from the expression of ψ^* in Example 2.1 below, it follows that

$$\varphi(1)\theta(\tau/\varphi(1)) = \begin{cases} \tau & \text{if } |\tau| \le 1, \\ \frac{-\tau^2+2a\tau-1}{2(a-1)} & \text{if } 1 < |\tau| \le a, \\ \frac{a+1}{2} & \text{if } |\tau| > a \end{cases}$$

which, by setting $s := \tau/\lambda$ for some constant $\lambda > 0$, implies that

$$\lambda^2\varphi(1)\theta(s/\lambda\varphi(1)) = \begin{cases} s\lambda & \text{if } |s| \le \lambda, \\ \frac{-s^2+2as\lambda-\lambda^2}{2(a-1)} & \text{if } \lambda < |s| \le a\lambda, \\ (a+1)\lambda^2/2 & \text{if } |s| > a\lambda. \end{cases}$$

Thus, when $\mathcal{J}_i = \{i\}$, by taking $\nu = \frac{1}{\lambda^2\varphi(1)}$ and $\rho = \frac{1}{\lambda\varphi(1)}$, the function Θ in (9) reduces to the SCAD function in [10]. Similarly, when ϕ is chosen as the one in Example 2.2, by taking $\nu = \frac{2}{\lambda^2 a}$ and $\rho = \frac{1}{\lambda}$, the function Θ in (9) becomes the MCP function in [39].

Now we give four examples for $\phi \in \Phi$. In the sequel we shall call ϕ_1-ϕ_4 as the function in Example 2.1-2.4, respectively, and ψ_1-ψ_4 as the corresponding ψ defined by (7).

Example 2.1 *Take $\phi(t) := \frac{\varphi(t)}{\varphi(1)}$ with $\varphi(t) := \frac{a-1}{2}t^2 + t$ for $t \in \mathbb{R}$, where $a > 1$ is a constant. Clearly, $\phi \in \Phi$ with $t_\phi^* = 0$ and $\overline{t}_\phi = \frac{1}{2}$. After a simple computation,*

$$\psi^*(s) = \begin{cases} 0 & \textit{if } |s| \le \frac{1}{\varphi(1)}, \\ \frac{(\varphi(1)|s|-1)^2}{2(a-1)\varphi(1)} & \textit{if } \frac{1}{\varphi(1)} < |s| \le \frac{a}{\varphi(1)}, \\ |s| - \frac{a+1}{2\varphi(1)} & \textit{if } |s| > \frac{a}{\varphi(1)}. \end{cases}$$

Example 2.2 *Let $\phi(t) := \frac{\varphi(t)}{\varphi(1)}$ with $\varphi(t) := \frac{a^2}{4}t^2 - \frac{a^2}{2}t + at + \frac{(a-2)_+^2}{4}$ where $a > 0$ is a constant and $(\cdot)_+ = \max(0, \cdot)$. Clearly, $\phi \in \Phi$ with $t_\phi^* = \frac{(a-2)_+}{a}$ and $\overline{t}_\phi = \max(\frac{a-1}{a}, \frac{1}{2})$. An elementary calculation yields that ψ^* takes the following form*

$$\psi^*(s) = \begin{cases} \frac{(a-2)_+^2}{4} & \textit{if } |s| \le \frac{a-a^2/2}{\varphi(1)}, \\ \frac{1}{a^2\varphi(1)}\big(\frac{a^2-2a}{2} + s\varphi(1)\big)^2 - \frac{(a-2)_+^2}{4\varphi(1)} & \textit{if } \frac{a-a^2/2}{\varphi(1)} < |s| \le \frac{a}{\varphi(1)}, \\ |s| - 1 & \textit{if } |s| > \frac{a}{\varphi(1)}. \end{cases}$$

Example 2.3 *Take $\phi(t) := t$ for $t \in \mathbb{R}$. Clearly, $\phi \in \Phi$ with $t_\phi^* = 0$ and $\overline{t}_\phi = 0$. Also,*

$$\psi^*(s) = \begin{cases} s - 1 & \text{if } s > 1, \\ 0 & \text{if } s \le 1. \end{cases}$$

In this case, the function Θ in (9) is exactly the capped l_1-surrogate of $\|\mathcal{G}(x)\|_0$ in [12].

Example 2.4 *Let $\phi(t) := \frac{\varphi(t)}{\varphi(1)}$ with $\varphi(t) := -t - \frac{q-1}{q}(1-t+\epsilon)^{\frac{q}{q-1}} + \epsilon + \frac{q-1}{q}$ $(0 < q < 1)$ for $t \in (-\infty, 1+\epsilon]$, where $\epsilon \in (0, 0.1)$ is a small constant. It is not hard to check that $\phi \in \Phi$ with $t_\phi^* = \epsilon$ and $\overline{t}_\phi = 1 + \epsilon - (\frac{1-\epsilon}{1-\epsilon+\varphi(1)})^{1-q}$. Also, $\psi^*(s) = \frac{h(\varphi(1)s)}{\varphi(1)}$ with*

$$h(t) := \begin{cases} t + \frac{q-1}{q}\epsilon^{\frac{q}{q-1}} - \epsilon - \frac{q-1}{q} & \text{if } t > \epsilon^{\frac{1}{q-1}} - 1, \\ (1+\epsilon)t - \frac{1}{q}(t+1)^q + \frac{1}{q} & \text{if } (1+\epsilon)^{\frac{1}{q-1}} - 1 < t \le \epsilon^{\frac{1}{q-1}} - 1, \\ \frac{q-1}{q}(1+\epsilon)^{\frac{q}{q-1}} - \epsilon - \frac{q-1}{q} & \text{if } t \le (1+\epsilon)^{\frac{1}{q-1}} - 1. \end{cases}$$

To close this section, we take a look at the local optimality relation of (6) and (2).

Theorem 2.2 *If $(\overline{x}, \overline{w})$ with $\langle e - \overline{w}, \mathcal{G}(\overline{x})\rangle = 0$ is a local optimal solution of (6) associated to $\rho > 0$, then $(\overline{x}, \overline{w})$ is locally optimal to (4) and so is $\overline{x}$ to (2). If $\overline{x}$ is locally optimal to (2), then $(\overline{x}, \overline{w})$ with $\overline{w} = \max(\text{sign}(\|\mathcal{G}(\overline{x})\|), t_\phi^* e)$ is locally optimal to (6) for any $\rho > 0$.*

Proof: Since $(\overline{x}, \overline{w})$ is a local optimal solution of (6) associated to $\rho > 0$, there exists $\delta' > 0$ such that for all $(x, w) \in \Omega \times [0, e]$ with $\|(x, w) - (\overline{x}, \overline{w})\| \le \delta'$,

$$\nu f(\overline{x}) + \textstyle\sum_{i=1}^m \phi(\overline{w}_i) + \rho\langle e - \overline{w}, \mathcal{G}(\overline{x})\rangle \le \nu f(x) + \sum_{i=1}^m \phi(w_i) + \rho\langle e - w, \mathcal{G}(x)\rangle.$$

Fix an arbitrary $(x, w) \in \mathcal{F}$ with $\|(x, w) - (\overline{x}, \overline{w})\| \le \delta'/2$ where $\mathcal{F}$ denotes the feasible set of (4). Then, from the last inequality, it immediately follows that

$$\begin{aligned} \nu f(\overline{x}) + \textstyle\sum_{i=1}^m \phi(\overline{w}_i) &= \nu f(\overline{x}) + \textstyle\sum_{i=1}^m \phi(\overline{w}_i) + \rho\langle e - \overline{w}, \mathcal{G}(\overline{x})\rangle \\ &\le \nu f(x) + \textstyle\sum_{i=1}^m \phi(w_i) + \rho\langle e - w, \mathcal{G}(x)\rangle = \nu f(x) + \textstyle\sum_{i=1}^m \phi(w_i). \end{aligned}$$

This shows that $(\overline{x}, \overline{w})$ is a locally optimal solution of the problem (4).

We next argue that $\overline{x}$ is locally optimal to (2). Since $(\overline{x}, \overline{w})$ is a locally optimal solution of (4), there exists $\widehat{\delta} > 0$ such that for all $(x, w) \in \mathcal{F}$ with $\|(x, w) - (\overline{x}, \overline{w})\| \le \widehat{\delta}$,

$$\nu f(\overline{x}) + \textstyle\sum_{i=1}^m \phi(\overline{w}_i) \le \nu f(x) + \sum_{i=1}^m \phi(w_i). \tag{10}$$

Let $\delta := \min(\frac{1}{2\nu L_f}, \frac{c}{4}, \widehat{\delta})$ where c is the minimal nonzero component of $\mathcal{G}(\overline{x})$. Fix an arbitrary $x \in \Omega \cap \{z \in \mathbb{R}^p : \|z - \overline{x}\| \le \delta\}$. We shall establish the following inequality

$$\nu f(\overline{x}) + \|\mathcal{G}(\overline{x})\|_0 \le \nu f(x) + \|\mathcal{G}(x)\|_0, \tag{11}$$

and consequently $\overline{x}$ is a local optimal solution to (2). If $\|\mathcal{G}(x)\|_0 \ge \|\mathcal{G}(\overline{x})\|_0 + 1$,

$$\|\mathcal{G}(x)\|_0 - \|\mathcal{G}(\overline{x})\|_0 \ge 1 > \nu L_f \delta \ge \nu L_f \|x - \overline{x}\| \ge \nu f(\overline{x}) - \nu f(x).$$

This implies that inequality (11) holds. If $\|\mathcal{G}(x)\|_0 \le \|\mathcal{G}(\overline{x})\|_0$, by using $\|x-\overline{x}\| \le \frac{c}{4}$, we have $\text{supp}(\mathcal{G}(x)) = \text{supp}(\mathcal{G}(\overline{x}))$, which implies that $\langle e-\overline{w}, \mathcal{G}(x)\rangle = 0$. Thus, $(x,\overline{w}) \in \mathcal{F}$ and $\|(x,\overline{w})-(\overline{x},\overline{w})\| \le \widehat{\delta}$. From (10), we obtain $\nu f(\overline{x})+\sum_{i=1}^m \phi(\overline{w}_i) \le \nu f(x)+\sum_{i=1}^m \phi(\overline{w}_i)$ or $\nu f(\overline{x}) \le \nu f(x)$. Along with $\text{supp}(\mathcal{G}(x)) = \text{supp}(\mathcal{G}(\overline{x}))$, the inequality (11) follows.

Since $\overline{x}$ is locally optimal to the problem (2), there exists $\overline{\delta} > 0$ such that for all $x \in \Omega$ with $\|x-\overline{x}\| \le \overline{\delta}$, it holds that $\nu f(\overline{x}) + \|\mathcal{G}(\overline{x})\|_0 \le \nu f(x) + \|\mathcal{G}(x)\|_0$. Fix an arbitrary $(x,w) \in \Omega \times [0,e]$ with $\|(x,w)-(\overline{x},\overline{w})\| \le \overline{\delta}$. Then, for any $\rho > 0$,

$$\begin{aligned}\nu f(\overline{x}) + \textstyle\sum_{i=1}^m \phi(\overline{w}_i) + \rho\langle e-\overline{w}, \mathcal{G}(\overline{x})\rangle &= \nu f(\overline{x}) + \|\mathcal{G}(\overline{x})\|_0 \le \nu f(x) + \|\mathcal{G}(x)\|_0 \\ &\le \nu f(x) + \textstyle\sum_{i=1}^m \phi(w_i) \\ &\le \nu f(x) + \textstyle\sum_{i=1}^m \phi(w_i) + \rho\langle e-w, \mathcal{G}(x)\rangle.\end{aligned}$$

where the second inequality is due to (3). Thus, $(\overline{x},\overline{w})$ is locally optimal to (6). □

3 GEP-MSCRA for computing the estimator $\widehat{x}$

From the last section, to compute the estimator $\widehat{x}$, one only needs to solve the penalty problem (6) associated to $\rho > \overline{\rho}$ with $\phi \in \Phi$, where the threshold $\overline{\rho} > 0$ is easily estimated once $\nu > 0$ is given since $L_f = \max_{x\in\Omega} \frac{1}{n}\|A^{\mathbb{T}}(Ax-b)\| \le \frac{R\sqrt{p}}{n}\|A\|^2 + \frac{1}{n}\|A^{\mathbb{T}}b\|$. For a given $\rho > \overline{\rho}$, although the problem (6) is nonconvex due to the coupled term $\sum_{i=1}^m w_i\|x_{\mathcal{J}_i}\|$, its special structure makes it much easier to cope with than do the problem (2). Specifically, when the variable w is fixed, the problem (6) reduces to a convex minimization in x; and when the variable x is fixed, it reduces to a convex minimization in w which, as will be shown below, has a closed-form solution. Motivated by this, we propose a multi-stage convex relaxation approach for computing $\widehat{x}$ by solving (6) in an alternating way.

Algorithm 3.1 *(**GEP-MSCRA for computing** $\widehat{x}$)*

(S.0) *Choose $\phi\in\Phi$, $\nu>0$ and an initial $w^0\in[0,\overline{t}_\phi e]$. Set $\lambda^0=\nu^{-1}$ and $k := 1$.*

(S.1) *Compute the following minimization problem*

$$x^k \in \mathop{\arg\min}_{x\in\Omega} \left\{ \frac{1}{2n}\|Ax-b\|^2 + \lambda^{k-1}\sum_{i=1}^m (1-w_i^{k-1})\|x_{\mathcal{J}_i}\| \right\}. \tag{12}$$

If $k=1$, by the information of x^1 select a suitable $\rho > \overline{\rho}$ and set $\lambda^k = \rho\nu^{-1}$.

(S.2) *Seek an optimal solution w_i^k $(i=1,2,\ldots,m)$ to the minimization problem*

$$w_i^k \in \mathop{\arg\min}_{w_i\in[0,1]} \left\{ \phi(w_i) - \rho w_i\|x_{\mathcal{J}_i}^k\| \right\}. \tag{13}$$

(S.3) *Set $k \leftarrow k+1$, and then go to Step (S.1).*

Remark 3.1 **(a)** *By the definition of ψ, clearly, w_i^k is an optimal solution to* (13) *if and only if $w_i^k \in \partial\psi^*(\rho\|x^k_{\mathcal{J}_i}\|)$. Since ψ^* is a convex function in $\mathbb{R}$, the subdifferential $\partial\psi^*(\rho\|x^k_{\mathcal{J}_i}\|)$ is easily characterized by [28]; for example, for the function ϕ_1, it holds that*

$$\partial\psi_1^*(\rho\|x^k_{\mathcal{J}_i}\|) = \Big\{\min\Big(1, \max\Big(\frac{\varphi_1(1)\rho\|x^k_{\mathcal{J}_i}\| - 1}{a-1}, 0\Big)\Big)\Big\}.$$

Thus, the main computation work in each iterate of the GEP-MSCRA is to solve (12).

(b) *When $k \geq 2$, since $w_i^k \in \partial\psi^*(\rho\|x^k_{\mathcal{J}_i}\|)$ for $i = 1, \ldots, m$, we have $1 - w_i^k \in \partial\theta(\rho\|x^k_{\mathcal{J}_i}\|)$ for all i, which means that the subproblem* (12) *for $k \geq 2$ has a similar form to the one yielded by applying the linear approximation technique in [41] to $\frac{1}{2n}\|Ax - b\|^2 + \lambda\Theta(x)$. Together with part (a), the GEP-MSCRA is analogous to the LLA algorithm in [41] for nonconvex penalized LS problems except the start-up and the weights. We see that the initial subproblem of the GEP-MSCRA depends explicitly on the dual variable w^0, while the initial subproblem of the LLA algorithm depends implicitly on an initial estimator x^0. This means that the start-up of the GEP-MSCRA is more easily controlled.*

(c) *By following the first part of proofs for Theorem 2.2, when an iterate x^k satisfies $\langle e - w^{k-1}, \mathcal{G}(x^k)\rangle = 0$, x^k is a local optimal solution of* (2), *and then $(x^k, \overline{w}^k)$ with $\overline{w}^k = \max(\mathrm{sign}(\|\mathcal{G}(x^k)\|), t_\phi^* e)$ is locally optimal to* (6) *for any $\rho > 0$ by Theorem 2.2 .*

4 Statistical guarantees

For convenience, throughout this section, we denote by $\overline{S}$ the group support of the true $\overline{x}$, i.e., $\overline{S} := \{i : \overline{x}_{\mathcal{J}_i} \neq 0\}$, and write $\overline{r} = |\overline{S}|$. With $\overline{S}$ and an integer $l > 0$, we define

$$\mathcal{C}(\overline{S}, l) := \Big\{z \in \mathbb{R}^p : \exists S \supset \overline{S} \text{ with } |S| \leq l \text{ such that } \sum_{i \in S^c} \|z_{\mathcal{J}_i}\| \leq \frac{2}{1 - \overline{t}_\phi} \sum_{i \in S} \|z_{\mathcal{J}_i}\|\Big\}.$$

Recall that the matrix A is said to satisfy the RSC of constant $\kappa > 0$ in a set $\mathcal{C}$ if

$$\frac{1}{2n}\|Ax\|^2 \geq \kappa\|x\|^2 \quad \forall x \in \mathcal{C}.$$

In this section, under an RSC assumption on A over $\mathcal{C}(\overline{S}, 1.5\overline{r})$, we shall establish an error bound for the iterate x^k to $\overline{x}$ and verify that the error sequence is strictly decreasing as k increases, and if in addition the nonzero group vectors of $\overline{x}$ are not too small, the iterate x^k of the GEP-MSCRA after finite steps satisfies $\mathrm{supp}(x^k) = \mathrm{supp}(\overline{x})$. Throughout the analysis, we assume that the components of the noise vector ε are independent (not necessarily identically distributed) sub-Gaussians, i.e., the following assumption holds.

Assumption 1 *Assume that ε_i $(i = 1, \ldots, m)$ are independent (but not necessarily identically distributed) sub-Gaussians, i.e., there exists $\sigma > 0$ such that for all i and $t \in \mathbb{R}$*

$$\mathbb{E}[\exp(t\varepsilon_i)] \leq \exp(\sigma^2 t^2/2).$$

The proofs of the main results of this section are all included in Appendix C.

4.1 Theoretical performance bounds

First of all, we characterize the error bound for every iterate x^k to the true vector $\overline{x}$.

Theorem 4.1 *Let $\widehat{\varepsilon} := \frac{1}{n}A^{\mathbb{T}}\varepsilon$. Suppose that A has the RSC of constant κ over $\mathcal{C}(\overline{S}, 1.5\overline{r})$, and $\frac{\sqrt{3}(4-2\overline{t}_\phi)(1-t^*_\phi)}{(3-\overline{t}_\phi)\kappa} \le \nu \le \frac{1-\overline{t}_\phi}{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}$. If $\frac{\nu(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi} \le \rho \le \sqrt{\frac{(3-\overline{t}_\phi)\kappa\nu}{\sqrt{3}(4-2\overline{t}_\phi)(1-t^*_\phi)}}$, then*

$$\|x^k - \overline{x}\| \le \begin{cases} \frac{\nu^{-1}(4-2\overline{t}_\phi)}{\kappa(3-\overline{t}_\phi)}\sqrt{1.5\overline{r}} & \text{if } k=1;\\ \frac{\rho\nu^{-1}(4-2\overline{t}_\phi)}{\kappa(3-\overline{t}_\phi)}\sqrt{1.5\overline{r}} & \text{if } k=2,3,\ldots,. \end{cases} \tag{14}$$

Remark 4.1 **(a)** *When $k=1$, the subproblem (12) reduces to the $\ell_{2,1}$-regularized least squares problem, and the bound in (14) has the same order as the one in [23, Corollary 4] except that the coefficient 2 there is improved to be $\frac{\sqrt{1.5}(4-2\overline{t}_\phi)}{3-\overline{t}_\phi}$. From the choice interval of ν, the worst bound of x^1 is $\frac{\sqrt{0.5\overline{r}}}{1-t^*_\phi}$ and that of x^k for $k \ge 2$ is $\frac{\rho\sqrt{0.5\overline{r}}}{1-t^*_\phi}$.*

(b) *The restriction on ν and ρ implies that $\lambda^{k-1} \ge \frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$. Such a restriction on λ^{k-1} is also required in the analysis of the $\ell_{2,1}$-regularized LS estimator [23, 17]. The choice interval of ν depends on the RSC property of A in $\mathcal{C}(\overline{S}, 1.5\overline{r})$ and the noise level. Clearly, for those problems in which A has a better RSC property over $\mathcal{C}(\overline{S}, 1.5\overline{r})$ or the noise $\|\mathcal{G}(\widehat{\varepsilon})\|_\infty$ is smaller, there is a larger choice interval for the parameter ν. In addition, those ϕ with larger t^*_ϕ and smaller $\overline{t}_\phi$ can deliver a larger choice interval of ν.*

(c) *If $\frac{(3-\overline{t}_\phi)\nu\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi} \ge \nu L_f\frac{(1-t^*_\phi)\phi'_-(1)}{1-\overline{t}_\phi}$ or equivalently $L_f \le \frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{(1-t^*_\phi)\phi'_-(1)}$, the choice interval of ρ in Theorem 4.1 is included in $[\overline{\rho}, +\infty)$. In this case, each subproblem (12) is a convex approximation of the exact penalty problem (6) in a low dimensional space.*

Theorem 4.1 provides an error bound for every iterate of the GEP-MSCRA, but it is unclear whether the error bound for the current iterate x^k is better than that of the previous iterate x^{k-1}, i.e., the error bound sequence is decreasing or not. We resolve this problem by bounding $(1-w_i^{k-1})^2$ for $i \in \overline{S}$ with $\mathbb{I}_\Delta(i)$ where $\Delta := \big\{i : \|\overline{x}_{\mathcal{J}_i}\| \le 2\phi'_-(1)/\rho\big\}$. The following theorem states this main result, and its proof involves the index sets

$$F^0 := \overline{S} \;\; \text{and} \;\; F^k := \left\{i : \big|\|x^k_{\mathcal{J}_i}\| - \|\overline{x}_{\mathcal{J}_i}\|\big| \ge \frac{1}{(1-t^*_\phi)\rho}\right\} \;\; \text{for } k = 1,2,\ldots. \tag{15}$$

Theorem 4.2 *Suppose that A has the RSC of constant κ over the set $\mathcal{C}(\overline{S}, 1.5\overline{r})$. If the parameter ν and ρ are chosen in the same way as in Theorem 4.1, then for each $k \in \mathbb{N}$,*

$$\|x^k - \overline{x}\| \le \frac{\sqrt{3}\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\|}{\kappa(\sqrt{3}-1)} + \frac{\sqrt{3}\rho}{\kappa(\sqrt{3}-1)\nu}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)} + \Big(\frac{1}{\sqrt{3}}\Big)^{k-1}\|x^1 - \overline{x}\|. \tag{16}$$

The error bound in (16) consists of three terms: the first term is the statistical error induced by the noise, the second one is the identification error related to the choice of ρ and ν, and the third one is the computation error. As will be shown in Subsection 4.2, the identification error will become zero if the parameters ρ and ν are appropriately chosen. Thus, inequality (16) implies that as k increases the error bound sequence is decreasing, and it will decrease to the statistical error $\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\|$ if the parameters ρ and ν are appropriately chosen, and otherwise it will decrease to the sum of the statistical error and the identification error. From (16), we also see that a smaller error bound of x^1 is beneficial to reduce the error bounds of x^k for $k \geq 2$. In practice, since $\|\mathcal{G}(\widehat{\varepsilon})\|_\infty$ is unknown, one may replace $\|\mathcal{G}(\widehat{\varepsilon})\|_\infty$ with $\|\mathcal{G}(Ax^1 - b)\|_\infty$ to estimate the choice interval of ρ. This means that the error bound of x^1 is important to the choice of ρ.

From [27] or [23, Page 549], we know that for a design matrix $Z \in \mathbb{R}^{n\times p}$ from the Σ-Gaussian ensemble (i.e., Z is formed by independently sampling each row $Z^i \sim N(0,\Sigma)$), there exists a constant $\kappa > 0$ (depending on the positive definite matrix Σ) such that Z has the RSC over $\mathcal{C}(\overline{S}, l)$ with probability greater than $1 - c_1 \exp(-c_2 n)$ as long as $n > c\sum_{i\in\overline{S}} |\mathcal{J}_i| \log p$, where c, c_1 and c_2 are absolutely positive constants. Together with Theorem 4.2 and Lemma 1 in Appendix A, we immediately get the following result.

Corollary 4.1 *Suppose Assumption 1 holds and* $\frac{\sqrt{3}(4-2\overline{t}_\phi)(1-t_\phi^*)}{(3-\overline{t}_\phi)\kappa} \leq \nu \leq \frac{1-\overline{t}_\phi}{(3-\overline{t}_\phi)K}$, *where* $K = \frac{\sigma}{n}\sqrt{2\log\frac{2p}{\eta}\|\mathcal{J}\|_\infty}\|A\|_{2,\infty}$ *for some* $\eta \in (0,1)$. *If* $\frac{\nu(3-\overline{t}_\phi)K}{1-\overline{t}_\phi} \leq \rho \leq \sqrt{\frac{(3-\overline{t}_\phi)\kappa\nu}{\sqrt{3}(4-2\overline{t}_\phi)(1-t_\phi^*)}}$, *then as long as* $n > \mathcal{O}(\sum_{i\in\overline{S}} |\mathcal{J}_i| \log p)$, *for each* $k \in \mathbb{N}$ *the following inequality*

$$\|x^k - \overline{x}\| \leq \frac{\sqrt{3}}{\kappa(\sqrt{3}-1)}\Big(K\sqrt{\overline{r}} + \frac{\rho}{\nu}\sqrt{\textstyle\sum_{i\in\overline{S}} \mathbb{I}_\Delta(i)}\Big) + \Big(\frac{1}{\sqrt{3}}\Big)^{k-1}\|x^1 - \overline{x}\|$$

holds with probability at least $1-\eta$.

4.2 Group selection consistency

In this part, we shall show that in finite steps the GEP-MSCRA can deliver an output x^l satisfying $\mathrm{supp}(\mathcal{G}(x^l)) = \mathrm{supp}(\mathcal{G}(\overline{x}))$ if the nonzero group vectors of $\overline{x}$ is not too small. To this end, we need to assume that the following least squares solution belongs to Ω:

$$x^{\mathrm{LS}} \in \mathop{\arg\min}_{x\in\mathbb{R}^p}\left\{\frac{1}{2n}\|Ax-b\|^2 : \ \mathrm{supp}(\mathcal{G}(x)) \subseteq \mathcal{J}_{\overline{S}}\right\}. \tag{17}$$

For the solution x^{LS}, one may establish the following ℓ_∞-norm error bound result.

Lemma 4.1 *Suppose that* A *has the RSC of constant* κ *over the set* $\mathcal{C}(\overline{S}, 1.5\overline{r})$. *Then,*

$$\|\mathcal{G}(x^{\mathrm{LS}} - \overline{x})\|_\infty \leq \|\mathcal{G}(\widehat{\varepsilon}^\dagger)\|_\infty \quad \text{with } \widehat{\varepsilon}^\dagger := (A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}} A_{\mathcal{J}_{\overline{S}}})^{-1} A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}\varepsilon. \tag{18}$$

Proof: When the matrix A satisfies the RSC over the set $\mathcal{C}(\overline{S}, l)$ for some $l > \overline{r}$, we have

$$\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})/\sqrt{n} \geq \sqrt{2\kappa} \tag{19}$$

and hence $A_{\mathcal{J}_{\overline{S}}}$ has full column rank and $\widehat{\varepsilon}^{\dagger}$ is well defined. Here $\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})$ is the smallest singular value of the matrix $A_{\mathcal{J}_{\overline{S}}}$. Indeed, for any $x\in\mathbb{R}^p$ with $\mathrm{supp}(\mathcal{G}(x))\subseteq\overline{S}$, we have $\frac{1}{2n}\|Ax\|^2=\frac{1}{2n}\|A_{\mathcal{J}_{\overline{S}}}x_{\mathcal{J}_{\overline{S}}}\|^2\geq\frac{1}{2n}\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})^2\|x_{\mathcal{J}_{\overline{S}}}\|^2=\frac{1}{2n}\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})^2\|x\|^2$, which along with $x\in\mathcal{C}(\overline{S},l)$ implies that $\kappa\leq\frac{1}{2n}[\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})]^2$, i.e., the inequality (19) holds. Now by the optimality of $x^{\rm LS}$ to the problem (17), we have $A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}(Ax^{\rm LS}-b)=0$. For $j\in\mathcal{J}_{\overline{S}}$,

$$\begin{aligned}
\big|x_j^{\rm LS}-\overline{x}_j\big| &= |e_j^{\mathbb{T}}(A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}A_{\mathcal{J}_{\overline{S}}})^{-1}A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}(A_{\mathcal{J}_{\overline{S}}}\overline{x}_{\mathcal{J}_{\overline{S}}}-A_{\mathcal{J}_{\overline{S}}}x_{\mathcal{J}_{\overline{S}}}^{\rm LS})|\\
&= \big|e_j^{\mathbb{T}}(A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}A_{\mathcal{J}_{\overline{S}}})^{-1}A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}(A\overline{x}-b+b-Ax^{\rm LS})\big|\\
&= \big|e_j^{\mathbb{T}}(A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}A_{\mathcal{J}_{\overline{S}}})^{-1}A_{\mathcal{J}_{\overline{S}}}^{\mathbb{T}}\varepsilon\big|.
\end{aligned}\tag{20}$$

Along with $x_j^{\rm LS}=0$ and $\overline{x}_j=0$ for $j\notin\mathcal{J}_{\overline{S}}$, we immediately obtain (18). □

Now we are ready to state the group selection consistency of the GEP-MSCRA.

Theorem 4.3 *Suppose that the matrix A has the RSC of constant κ over the set $\mathcal{C}(\overline{S},1.5\overline{r})$ and $\frac{(1-t_\phi^*)(1+3\sqrt{5})}{4\kappa}\leq\nu\leq\frac{1-\overline{t}_\phi}{2\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty}$ with $\varepsilon^{\rm LS}:=\frac{1}{n}A^{\mathbb{T}}(Ax^{\rm LS}-b)$. If ρ is chosen such that $\max\big(\frac{2\phi'_-(1)}{\min_{i\in\overline{S}}\|\overline{x}_{\mathcal{J}_i}\|},\frac{2\nu\max(\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty,2\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty)}{1-\overline{t}_\phi}\big)<\rho\leq\sqrt{\frac{4\kappa\nu}{(1-t_\phi^*)(1+3\sqrt{5})}}$, then for each $k\in\mathbb{N}$*

$$\begin{aligned}
\|x^k-x^{\rm LS}\| &\leq \frac{\max(1,\sqrt{5}\rho/2)}{\nu\kappa}\sqrt{|F^{k-1}|},\\
\sqrt{|F^k|} &\leq \frac{\max(1,\rho)\rho(1-t_\phi^*)(1+3\sqrt{5})}{6\nu\kappa}\sqrt{|F^{k-1}|}.
\end{aligned}\tag{21}$$

Also, $x^k=x^{\rm LS}$ and $\mathrm{supp}(\mathcal{G}(x^k))=\overline{S}$ for $k\geq\overline{k}:=\lceil\frac{0.5\ln(\overline{r})}{\ln(6\nu\kappa)-\ln[(\max(1,\rho)\rho(1-t_\phi^*)(1+3\sqrt{5}))]}\rceil+1$.

Remark 4.2 **(a)** *Theorem 4.3 shows that if the parameters ν and ρ are appropriately chosen, then the iterate x^k with $k>\overline{k}$ coincides with the oracle solution $x^{\rm LS}$ and its group support coincides with $\overline{S}$. Similar to Remark 4.1(b), for those problems in which A has a better RSC property in $\mathcal{C}(\overline{S},1.5\overline{r})$ and the noise $\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty$ is smaller, the choice interval of ν is larger. If the smallest nonzero group vector of $\overline{x}$ is suitable large, say $\min_{i\in\overline{S}}\|\overline{x}_{\mathcal{J}_i}\|\geq\frac{\phi'_-(1)(1-t_\phi^*)}{\nu\max(\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty,2\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty)}$, then the choice of ρ depends only on the noise. It is not hard to observe that those ϕ with smaller $\overline{t}_\phi$ and larger t_ϕ^* lead to a larger choice interval of ν and ρ and a smaller $\overline{k}$. Together with Remark 4.1(b), the GEP-MSCRA with such ϕ is better in terms of the error bound and the group consistency.*

(b) *By Lemma 2, we have $\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty\leq K$ w.p. at least $1-\eta$ for $\eta\in(0,1)$. We next show that $\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty\leq K$ w.p. no less than $1-\eta$ for $\eta\in(0,1)$. Indeed, by Lemma 3,*

$$\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty\leq\frac{\kappa Kn}{\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})\|A\|_{2,\infty}}\leq\frac{K\sqrt{\kappa n}}{\sqrt{2}\|A\|_{2,\infty}}.$$

In addition, since for any $e_j\in\mathbb{R}^p$ with $j=1,2,\ldots,p$, we have $e_j\in\mathcal{C}(\overline{S},1.5\overline{r})$ which, together with $\frac{1}{2n}\|Ae_j\|^2=\frac{1}{2n}\|A_j\|^2\|e_j\|^2$, implies that $\sqrt{\kappa}\leq\frac{1}{\sqrt{2n}}\|A\|_{2,\infty}$. Substituting *this relation into the last inequality yields that $\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty \le K$. Thus, $\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty$ and $\kappa\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty$ have the upper bound of the same order in a high probability.*

Using Lemma 2-3, Remark 4.2(b) and Theorem 4.3, we obtain the following result.

Corollary 4.2 *Suppose that Assumption 1 holds and $\frac{(1-t_\phi^*)(1+3\sqrt{5})}{4\kappa} < \nu \le \frac{1-\bar{t}_\phi}{2K}$. If $\max\Big(\frac{2\phi'_-(1)}{\min_{i\in\overline{S}}\|\overline{x}_{\mathcal{J}_i}\|}, \frac{4K\nu}{1-\bar{t}_\phi}\Big) < \rho \le \sqrt{\frac{4\kappa\nu}{(1-t_\phi^*)(1+3\sqrt{5})}}$, then as long as $n > \mathcal{O}(\sum_{i\in\overline{S}}|\mathcal{J}_i|\log p)$, we have $x^k = x^{\rm LS}$ and $\mathrm{supp}(\mathcal{G}(x^k)) = \overline{S}$ for $k \ge \overline{k}$ w.p. at least $1-2\eta$ for $\eta\in(0,1)$.*

Corollary 4.1 and 4.2 provide the theoretical guarantees in statistical sense. We need to point out, when a similar column normalization condition is imposed to the design matrix A, one may follow the analysis in [23] to improve the probability bound results.

5 Numerical experiments for the GEP-MSCRA

The GEP-MSCRA consists in solving a sequence of weighted $\ell_{2,1}$-norm regularized problems. The key to its implementation is to develop an effective solver to (12) or equivalently

$$\min_{x,u\in\mathbb{R}^p,z\in\mathbb{R}^n}\Big\{\frac{1}{2}\|z\|^2 + \textstyle\sum_{i=1}^m \omega_i\|x_{\mathcal{J}_i}\| + \delta_\Omega(u):\ Ax-z=b,\ x-u=0\Big\},\tag{22}$$

where $\omega_i = n\lambda(1-w_i^k)$ for $i=1,\ldots,m$ are nonnegative weights. There are some solvers developed for the unconstrained counterpart of (12); for example, the LARS-type algorithm in [33], the R-package **gglasso** developed by Yang and Zou [32] with the groupwise-majorization-descent algorithm, the Matlab package **SLEP** developed by Liu and Ye [16] with the accelerated proximal gradient method [24], and the semismooth Newton ALM developed by Li, Sun and Toh [15]. The first three solvers are solving (12) with $\Omega=\mathbb{R}^p$, while the last one is solving its dual problem. These solvers can not be applied directly to the problem (22) since it involves an additional nonsmooth term $\delta_\Omega(u)$.

5.1 Implementation of the GEP-MSCRA

Motivated by the good performance of the semismooth Newton ALM (see [15, 29]), we shall develop it for solving the dual of (22) which takes the following form

$$\min_{\xi\in\mathbb{R}^n,\eta,\zeta\in\mathbb{R}^p}\Big\{\frac{1}{2}\|\xi\|^2 + \langle b,\xi\rangle + R\|\eta\|_1 + \delta_\Lambda(\zeta):\ A^{\mathbb{T}}\xi+\eta-\zeta=0\Big\},\tag{23}$$

where $\Lambda=\Lambda_1\times\Lambda_2\times\cdots\times\Lambda_m$ with $\Lambda_i:=\{z\in\mathbb{R}^{|\mathcal{J}_i|}\mid \|z\|\le\omega_i\}$ for $i=1,2,\ldots,m$. For a given $\sigma>0$, the augmented Lagrangian function of problem (23) is defined as

$$L_\sigma(\eta,\xi,\zeta;x) := \frac{1}{2}\|\xi\|^2 + \langle b,\xi\rangle + R\|\eta\|_1 + \delta_\Lambda(\zeta) + \langle x, A^{\mathbb{T}}\xi+\eta-\zeta\rangle + \frac{\sigma}{2}\|A^{\mathbb{T}}\xi+\eta-\zeta\|^2.$$

The iteration steps of the augmented Lagrangian method for (23) is described as follows.

```
Algorithm 1  An inexact ALM for the dual problem (23)
Initialization: Choose $\sigma_0 > 0$ and a starting point $(\eta^0, \xi^0, \zeta^0, x^0)$. Set $j = 0$.
while the stopping conditions are not satisfied do
    1. Solve the following nonsmooth convex minimization problem inexactly
           $(\eta^{j+1}, \xi^{j+1}, \zeta^{j+1}) \approx \operatorname*{arg\,min}_{\xi \in \mathbb{R}^n, \eta, \zeta \in \mathbb{R}^p} L_{\sigma_j}(\eta, \xi, \zeta; x^j).$   (24)
    2. Update the multiplier by the formula $x^{j+1} = x^j + \sigma_j(A^{\mathbb{T}}\xi^{j+1} + \eta^{j+1} - \zeta^{j+1})$.
    3. Update $\sigma_{j+1} \uparrow \sigma_\infty \leq \infty$. Set $j \leftarrow j + 1$, and then go to Step 1.
end while
```

Observe that the augmented Lagrangian subproblem (24) is a two-block nonsmooth convex program. We use the accelerated block coordinate descent (ABCD) method to seek $(\eta^{j+1}, \xi^{j+1}, \zeta^{j+1})$ in (24). The iterations of the ABCD method are described below.

```
Algorithm 2  An ABCD for solving the Lagrangian subproblem (24)
Initialization: Choose the initial point $(\widetilde{\xi}^1, \widetilde{\zeta}^1) = (\xi^j, \zeta^j)$ and let $t_1 = 1$. Set $k := 1$.
while the stopping conditions are not satisfied do
    1. Compute the following minimization problems
           $\begin{cases} \eta^{k,j} = \operatorname*{arg\,min}_{\eta \in \mathbb{R}^p} L_{\sigma_j}(\eta, \widetilde{\xi}^k, \widetilde{\zeta}^k; x^j), & (25a) \\ (\xi^{k,j}, \zeta^{k,j}) = \operatorname*{arg\,min}_{\xi \in \mathbb{R}^n, \zeta \in \mathbb{R}^p} L_{\sigma_j}(\eta^{k,j}, \xi, \zeta; x^j). & (25b) \end{cases}$
    2. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\beta_k = \frac{t_k - 1}{t_{k+1}}$, and then compute
           $\widetilde{\xi}^{k+1} = \xi^{k,j} + \beta_k(\xi^{k,j} - \xi^{k-1,j})$ and $\widetilde{\zeta}^{k+1} = \zeta^{k,j} + \beta_k(\zeta^{k,j} - \zeta^{k-1,j})$.
    3. Let $k \leftarrow k + 1$, and go to Step 1.
end while
```

Let $\mathrm{prox}_{\ell_1,\gamma} : \mathbb{R}^p \to \mathbb{R}^p$ denote the proximal mapping of ℓ_1-norm of parameter γ, i.e.,

$$\mathrm{prox}_{\ell_1,\gamma}(z) := \min_{x' \in \mathbb{R}^p} \left\{ \frac{1}{2}\|x' - z\|^2 + \gamma\|x'\|_1 \right\}.$$

From the definition of the augmented Lagrangian function, the solution $\eta^{k,j}$ has the form

$$\eta^{k,j} = \mathrm{prox}_{\ell_1, R/\sigma_j}\big(\widetilde{\zeta}^k - A^{\mathbb{T}}\widetilde{\xi}^k - x^j/\sigma_j\big).$$

Let $\Phi_{k,j}(\xi) := \min_{\zeta\in\mathbb{R}^p} L_{\sigma_j}(\eta^{k,j},\xi,\zeta;x^j)$ for $\xi\in\mathbb{R}^n$. It is not difficult to verify that

$$\xi^{k,j} = \mathop{\arg\min}_{\xi\in\mathbb{R}^n} \Phi_{k,j}(\xi) \;\text{ and }\; \zeta^{k,j} = \Pi_\Lambda\big(A^{\mathbb{T}}\xi^{k,j}+\eta^{k,j}+x^j/\sigma_j\big).$$

After an elementary calculation, one may obtain the expression of $\Phi_{k,j}$ as follows

$$\Phi_{k,j}(\xi) = \frac{\sigma_j}{2}\Big\|\Pi_\Lambda\Big(A^{\mathbb{T}}\xi+\eta^{k,j}+\frac{x^j}{\sigma_j}\Big)-\Big(A^{\mathbb{T}}\xi+\eta^{k,j}+\frac{x^j}{\sigma_j}\Big)\Big\|^2+\frac{1}{2}\|\xi\|^2+\langle b,\xi\rangle + R\|\eta^{k,j}\|_1.$$

By the strong convexity of $\Phi_{k,j}$, $\xi^{k,j} = \arg\min_{\xi\in\mathbb{R}^n}\Phi_{k,j}(\xi)$ iff $\xi^{k,j}$ satisfies the system

$$\nabla\Phi_{k,j}(\xi) = b+\xi+\sigma_j A\left[\Big(A^{\mathbb{T}}\xi+\eta^{k,j}+\frac{x^j}{\sigma_j}\Big)-\Pi_\Lambda\Big(A^{\mathbb{T}}\xi+\eta^{k,j}+\frac{x^j}{\sigma_j}\Big)\right]=0. \tag{26}$$

The system (26) is strongly semismooth (see [21, 26, 30] for the related discussion), and we apply the semismooth Newton method for solving it. Write $y := A^{\mathbb{T}}\xi+\eta^{k,j}+\frac{x^j}{\sigma_j}$. By [8, Proposition 2.3.3 & Theorem 2.6.6], the Clarke Jacobian $\partial\nabla\Phi_{k,j}$ of $\Phi_{k,j}$ satisfies

$$\partial(\nabla\Phi_{k,j})(\xi)\subseteq\widehat{\partial}^2\Phi_{k,j}(\xi) := I+\sigma_j A(I-\partial\Pi_\Lambda(y))A^{\mathbb{T}} \tag{27}$$

where $\widehat{\partial}^2\Phi_{k,j}$ is the generalized Hessian of $\Phi_{k,j}$ at ξ. Since the exact characterization of $\partial\nabla\Phi_{k,j}$ is difficult to obtain, we replace $\partial\nabla\Phi_{k,j}$ with $\widehat{\partial}^2\Phi_{k,j}$ in the solution of (26). Let $W\in\partial\Pi_\Lambda(y)$. By [8, Theorem 2.6.6], $W = W_{\mathcal{J}_1}\times\cdots\times W_{\mathcal{J}_m}$ with $W_{\mathcal{J}_i}\in\partial\Pi_{\Lambda_i}(y_{\mathcal{J}_i})$ where

$$\partial\Pi_{\Lambda_i}(y_{\mathcal{J}_i}) = \begin{cases} \{I\} & \text{if } \|y_{\mathcal{J}_i}\|<\omega_i,\\ \mathrm{conv}\big(I, I-\frac{1}{\omega_i^2}y_{\mathcal{J}_i}y_{\mathcal{J}_i}^{\mathbb{T}}\big) & \text{if } \|y_{\mathcal{J}_i}\|=\omega_i,\\ \Big\{\omega_i\Big(\frac{1}{\|y_{\mathcal{J}_i}\|}I-\frac{1}{\|y_{\mathcal{J}_i}\|^3}y_{\mathcal{J}_i}y_{\mathcal{J}_i}^{\mathbb{T}}\Big)\Big\} & \text{if } \|y_{\mathcal{J}_i}\|>\omega_i \end{cases} \tag{28}$$

where $\mathrm{conv}\big(I, I-\frac{1}{\omega_i^2}y_{\mathcal{J}_i}y_{\mathcal{J}_i}^{\mathbb{T}}\big)$ means the convex combination of I and $I-\frac{1}{\omega_i^2}y_{\mathcal{J}_i}y_{\mathcal{J}_i}^{\mathbb{T}}$. From (27) and (28), each element $I+\sigma_j A(I-W)A^{\mathbb{T}}$ in $\widehat{\partial}^2\Phi_{k,j}(\xi)$ is positive definite, which by [26] implies that the following semismooth Newton method has a fast convergence rate.

During the implementation of the semismooth Newton ALM for (23), we terminated the algorithm once $\max\{\varepsilon^j_{\rm pinf},\varepsilon^j_{\rm dinf},\varepsilon^j_{\rm gap}\}\le\epsilon^j$, where $\varepsilon^j_{\rm gap}$ is the primal-dual gap, i.e., the sum of the objective values of (22) and (23) at (η^j,ξ^j,ζ^j), and $\varepsilon^j_{\rm pinf}$ and $\varepsilon^j_{\rm dinf}$ are the primal and dual infeasibility measure at (η^j,ξ^j,ζ^j), respectively, defined as follows

$$\varepsilon^j_{\rm pinf} := \frac{\sigma_{j-1}\|(\zeta^{k,j}-\widetilde{\zeta}^k)+A^{\mathbb{T}}(\widetilde{\xi}^k-\xi^{k,j})\|}{1+\|b\|} \;\text{ and }\; \varepsilon^j_{\rm dinf} := \frac{\|x^j-x^{j-1}\|}{\sigma_{j-1}}.$$

Now we return to the choice of parameters in the GEP-MSCRA. Taking into account the choice of ρ in the first stage may not be the best, we use a dynamic adjustment for ρ during the test. Specifically, we choose $\rho^1 = \frac{2}{\|\mathcal{G}(x^1)\|_\infty}$ and increase it by the rule

```
Algorithm 3  A semismooth Newton-CG (SNCG) algorithm for (26)
Initialization: Choose $\overline{\theta}\in(0,1), \tau\in(0,1), \delta\in(0,1), \mu\in(0,\frac{1}{2})$ and $\xi^0\in\mathbb{R}^n$. Set $l=0$.
while the stopping conditions are not satisfied do
    1. Choose a matrix $V^l\in\widehat{\partial}^2\Phi_{k,j}(\xi^l)$. Solve the following linear system
           $V^l d = -\nabla\Phi_{k,j}(\xi^l)$
       with the conjugate gradient (CG) algorithm to find $d^l$ such that
           $\|V^l d^l + \nabla\Phi_{k,j}(\xi^l)\| \le \min(\overline{\theta}, \|\nabla\Phi_{k,j}(\xi^l)\|^{1+\tau})$
    2. Set $\alpha_l=\delta^{m_l}$, where $m_l$ is the first nonnegative integer $m$ for which
           $\Phi_{k,j}(\xi^l+\delta^m d^l) \le \Phi_{k,j}(\xi^l) + \mu\delta^m\langle\nabla\Phi_{k,j}(\xi^l), d^l\rangle.$
    3. Set $\xi^{l+1}=\xi^l+\alpha_l d^l$ and $l\leftarrow l+1$, and then go to Step 1.
end while
```

$\rho^k=\min(2\rho^{k-1}, 10^8/\|\mathcal{G}(x^k)\|_\infty)$ for $k\ge 2$. The choice of ν is specified in the experiments. By Remark 3.1(c), we terminate the GEP-MSCRA at the iterate x^k whenever it satisfies

$$\big\langle e-w^{k-1}, \mathcal{G}(x^k)\big\rangle \le \epsilon_{\rm gap} \quad\text{or}\quad \frac{|f(x^k)-f(x^{k-1})|}{\max(1,f(x^k))} \le \epsilon_{\rm loss},\ \big|\|x^k\|_{a,0}-\|x^{k-1}\|_{a,0}\big| \le 1$$

where $\|z\|_{a,0} := \sum_{i=1}^m \mathbb{I}_{\{i:\,\|z_{\mathcal{J}_i}\|>10^{-6}\}}(z)$ means the approximate group zero-norm of z. During the testing, we choose $\epsilon_{\rm gap}=10^{-6}$ and $\epsilon_{\rm loss}=10^{-2}$, and solve the subproblem (12) by Algorithm 1 with the tolerance $\epsilon^j=\max(10^{-5}, 0.8\epsilon^{j-1})$ and $\epsilon^0=0.1\epsilon_{\rm loss}$. All numerical results of this section are obtained from a laptop running on 64-bit Windows Operating System with an Intel(R) Core(TM) i7-7700 CPU 2.8GHz and 16 GB memory.

5.2 Numerical experiments for group sparse regressions

We shall evaluate the performance of the GEP-MSCRA in the group sparse regression setting by using the simulated data. We generate the simulation data with the sample size n, the dimension of variables p, the number of groups m, and the dimension of each group $d=\lceil p/m\rceil$. The matrix A is generated randomly by one of the following ways:

(I) $A=\textbf{randn}(n,p)$;

(II) $A=\textbf{sign}(\textbf{rand}([n,p])-0.5)$; $\text{ind}=\text{find}(A=0)$; $A(\text{ind})=\text{ones}(\text{size}(\text{ind}))$;

(III) $A=\textbf{hadamard}(n)$; $\text{picks}=\text{randperm}(n)$; $\text{picks}=\text{sort}(\text{picks}(1{:}n))$; $A=A(\text{picks},:)$.

We select $\overline{r}$ groups randomly from m groups, say $\{m_1,\ldots,m_{\overline{r}}\}$, as the support of $\overline{x}$, and generate the entries of $\overline{x}_{\mathcal{J}_i}$ for $i\in\{m_1,\ldots,m_{\overline{r}}\}$ in one of the following seven ways:

(i) $\overline{x}_{\mathcal{J}_i} = \alpha\,\mathbf{randn}(|\mathcal{J}_i|, 1)$ for $i \in \{m_1, \ldots, m_{\overline{r}}\}$ with $\alpha = 2$ or 10^5;

(ii) $\overline{x}_{\mathcal{J}_i} = \alpha\,\mathbf{rand}(|\mathcal{J}_i|, 1) - 0.5$ for $i \in \{m_1, \ldots, m_{\overline{r}}\}$ with $\alpha = 2$ or 10^5;

(iii) $\overline{x}_{\mathcal{J}_i} = \alpha\,\mathbf{sign}(\mathbf{randn}(|\mathcal{J}_i|, 1))$ for $i \in \{m_1, \ldots, m_{\overline{r}}\}$ with $\alpha = 1$ or 10^5;

(iv) $\overline{x}_{\mathcal{J}_i} = -\frac{10^5}{\sqrt{i}}e$ for $i \in \{m_1, \ldots, m_{\overline{r}/2}\}$ and $\overline{x}_{\mathcal{J}_i} = \frac{10^5}{\sqrt{i}}e$ for $i \in \{m_{(\overline{r}+1)/2}, \ldots, m_{\overline{r}}\}$.

Then, we set $b = A(\overline{x} + \vartheta_1 \frac{\widetilde{\varepsilon}}{\|\widetilde{\varepsilon}\|}) + \vartheta_2 \frac{\varepsilon}{\|\varepsilon\|}$ where $\widetilde{\varepsilon} = \mathbf{randn}(p, 1)$, $\varepsilon = \mathbf{randn}(p, 1)$ and ϑ_1 and ϑ_2 are the nonnegative constants representing the scale of the noise vectors ε and $\widetilde{\varepsilon}$. Since the true $\overline{x}$ is known for these synthetic problems, we take $R = 1000\|\overline{x}\|_\infty$ for the set Ω. We find from experiments that Algorithm 1 is not sensitive to the value of R.

5.2.1 Performance of the GEP-MSCRA with different ϕ

This part aims to evaluate the performance of the GEP-MSCRA with $\phi \in \{\phi_1, \phi_2, \phi_3, \phi_4\}$ where $a = 3.7$ and $a = 3$ are used for ϕ_1 and ϕ_2 respectively, and $\epsilon = 10^{-2}$ is used for ϕ_4. With the design matrix $A \in \mathbb{R}^{n\times p}$ of type I for $(p, m, \overline{r}) = (2^{12}, 256, 10)$, we generate 10 test problems randomly as above for every type of $\overline{x}$ with $(\vartheta_1, \vartheta_2) = (0.1, 0.1)$, and apply the GEP-MSCRA for solving the test problems with $\lambda = (0.1/n)\|A^{\mathbb{T}}b\|_\infty$. Figure 1 plots the average relative prediction error curve and the average computing time curve, respectively, yielded by the GEP-MSCRA with each ϕ under the sample size $n = \lfloor \frac{p}{\beta} \rfloor$ for $\beta \in \{5, 6, \ldots, 17\}$. Here, for each sample size, the average relative error and computing time is the average of the total relative prediction error and computing time of the 70 test problems. The relative error is defined by $\mathbf{relerr} := \frac{\|x^{\rm out} - \overline{x}\|}{\|\overline{x}\|}$ where $x^{\rm out}$ is the output.

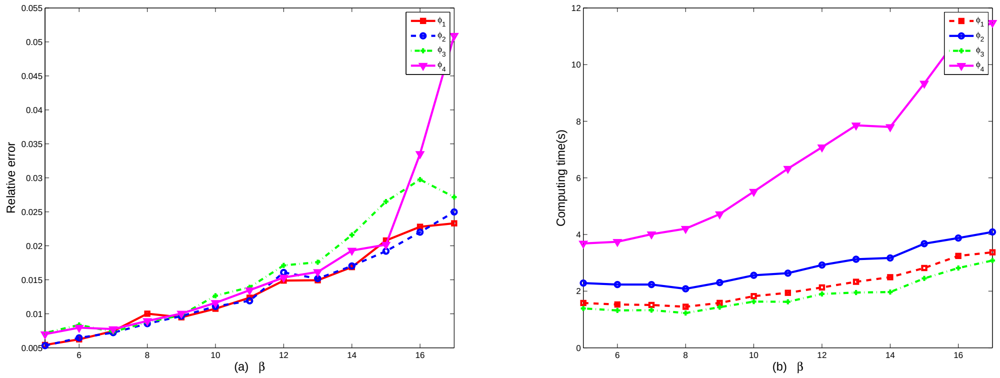


Figure 1: Performance of the GEP-MSCRA with ϕ_1-ϕ_4 under different sample size

Figure 1 shows that the relative errors yielded with ϕ_1-ϕ_4 are comparable, but those yielded with ϕ_3 and ϕ_4 have a little bigger fluctuation. In addition, the GEP-MSCRA with ϕ_2 and ϕ_4 requires more computing time than the GEP-MSCRA with ϕ_1 and ϕ_3 does. By this, we choose the GEP-MSCRA with ϕ_1 for the subsequent experiments.

5.2.2 Numerical comparison with the SLEP

The SLEP is a solver to the subproblem (12) without the constraint $x \in \Omega$ but with positive weights. So, we first compare the performance of Algorithm 1 for solving the subproblem (12) for $k = 1$ and $w^0 = 0$ with that of the SLEP for solving its counterpart without the constraint $x \in \Omega$ under different λ. Unless otherwise stated, all parameters involved in the SELP are set to be the default one. We generate 10 test problems randomly as above for every type of $\overline{x}$ with $(\vartheta_1, \vartheta_2) = (0.1, 0.1)$ and the design matrix $A \in \mathbb{R}^{n\times p}$ of type I for $(p, m, \kappa) = (2^{12}, 256, 15)$ and $n = \lfloor p/10 \rfloor$. Figure 2 plots the average relative error and computing time curves of Algorithm 1 and the SLEP for solving the 70 problems with $\lambda = (\beta/n)\|A^{\mathbb{T}}b\|_\infty$. We see that the relative error yielded by Algorithm 1 has less variation than the one yielded by the SLEP when $\lambda \in [0.03/n, 0.3/n]\|A^{\mathbb{T}}b\|_\infty$, which means that it is easier to choose an appropriate λ for Algorithm 1. Since the problem (12) is more difficult than its unconstrained counterpart, Algorithm 1 requires more time than the SLEP does, but its computing time decreases as λ increases, and when $\lambda \geq (0.2/n)\|A^{\mathbb{T}}b\|_\infty$ its time is less than three times that of the SLEP.

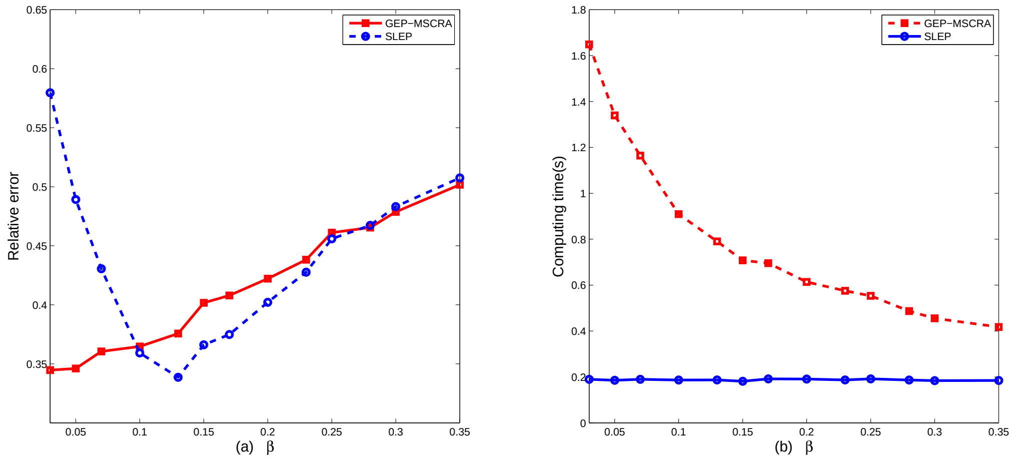
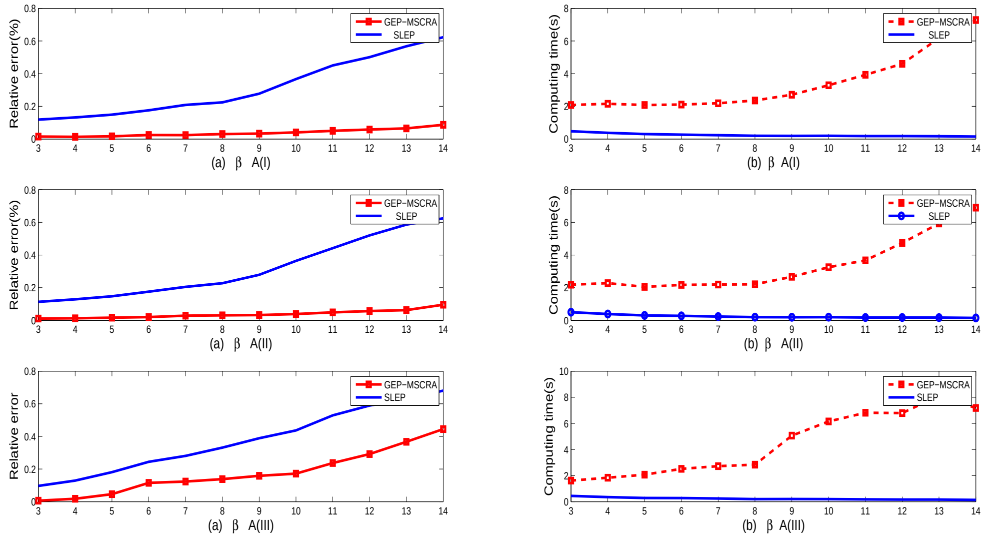


Figure 2: Performance of Algorithm 1 and the SLEP under different $\lambda = (\beta/n)\|A^{\mathbb{T}}b\|_\infty$

Next we compare the performance of the GEP-MSCRA for computing $\widehat{x}$ with that of the SLEP for computing the $\ell_{2,1}$-norm regularized LS estimator, i.e., the one defined by the subproblem (12) with $k = 1$ and $w^0 = 0$ but without the constraint $x \in \Omega$. To this end, for each type of A with $(p, m, \kappa) = (2^{12}, 256, 15)$, we generate 10 test problems randomly for every type of $\overline{x}$ with $(\vartheta_1, \vartheta_2) = (0.1, 0.3)$, and then apply the GEP-MSCRA and the SLEP, respectively, for solving the corresponding test problems. By Figure 2, we choose $\lambda = (0.1/n)\|A^{\mathbb{T}}b\|_\infty$ for the GEP-MSCRA and $\lambda = (0.13/n)\|A^{\mathbb{T}}b\|_\infty$ for the SLEP. Figure 3 plots the average relative error and computing time curves under different sample size $n = \lfloor \frac{p}{\beta} \rfloor$ for $\beta \in \{3, 4, \ldots, 15\}$. From Figure 3, we see that although the SLEP is faster than the GEP-MSCRA, for the matrix A of type I and II, the relative error of its output is about **six** or **seven** times higher than that of the GEP-MSCRA, and for the matrix A of type III, the relative error of its output is about **one and half** times higher than that of the GEP-MSCRA. In addition, Figure 4 shows under each sample size, the group sparsity of the output yielded by the SLEP is much higher than that of $\overline{x}$ when the sample size becomes less, but that of the output yielded by the GEP-MSCRA is close to that of the true $\overline{x}$. This means that the estimator yielded by the GEP-MSCRA is much better than the one yielded by the SLEP in terms of the relative error and the group sparsity. Notice that the matrix A of type I and II satisfies the RSC condition in a high probability. Thus, the numerical performance matches the theoretical analysis well.



(a) β A(I)

(b) β A(I)

(a) β A(II)

(b) β A(II)

(a) β A(III)

(b) β A(III)

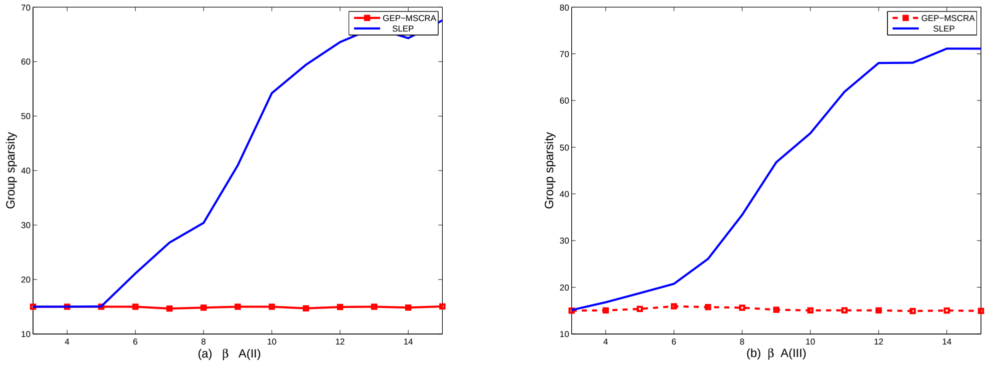
Figure 3: Relative error of the output yielded by the GEP-MSCRA and the SLEP

5.3 Numerical experiments for multi-task learning

In multi-task learning (see [2, 25, 36]), we are given a training set of m tasks $\{(a_i^k, y_i^k)\}_{i=1}^{n_k}$ from the linear models $h_k(a) = \langle w_{\mathcal{J}_k}, a\rangle$ for $k = 1, \ldots, m$, where $w_{\mathcal{J}_k} \in \mathbb{R}^{|\mathcal{J}_k|}$ is the weight vector for the kth task, $a_i^k \in \mathbb{R}^{|\mathcal{J}_k|}$ is the ith training sample for the kth task, y_i^k is the corresponding output, and n_k is the number of training samples for the kth task. Write $y_k = [y_1^k, \ldots, y_{n_k}^k]^{\mathbb{T}} \in \mathbb{R}^{n_k}$ and $b = [y_1^{\mathbb{T}}, \ldots, y_m^{\mathbb{T}}]^{\mathbb{T}} \in \mathbb{R}^n$ with $n = \sum_{j=1}^m n_j$. Let $A_{\mathcal{J}_k} = [a_1^k, \ldots, a_{n_k}^k]^{\mathbb{T}} \in \mathbb{R}^{n_k \times |\mathcal{J}_k|}$ denote the data matrix for the kth task. Clearly, the model (1) is also applicable to the multi-task learning by replacing $\overline{x}$ with $w = [w_{\mathcal{J}_1}^{\mathbb{T}}, \ldots, w_{\mathcal{J}_m}^{\mathbb{T}}]^{\mathbb{T}}$.

This part focuses on the comparison of the GEP-MSCRA and the MALSAR[1] for a

[1] `http://yelab.net/software/MALSRA(version1.1)/`



(a) β A(II)

(b) β A(III)

Figure 4: Group sparsity of the output yielded by the GEP-MSCRA and the SLEP

real data set (School data) from the Inner London Education Authority[2]. Among others, the MALSAR is a solver for the unconstrained $\ell_{2,1}$-regularized LS model, and since the true $\overline{x}$ is unknown for this real problem, we take $R = 2000$ for the set Ω. This data set has been used in previous works on multi-task learning (see [9]). It consists of examination scores of 15362 students from 139 secondary schools in London during the years 1985, 1986 and 1987. There are 139 tasks, corresponding to predicting student performance in each school. The input consists of the year of the examination (YR), 4 school-specific and 3 student-specific attributes, and each sample contains 28 attributes.

We first test the prediction performance of the GEP-MSCRA and the MALSAR with different $\lambda = \nu^{-1}$. We generate the training and test sets by 10 random splits of the data, so that **75**% of the examples from each school (task) belong to the training set and **25**% to the test set. The subfigures in the first line of Figure 5 plot the prediction error and time curves of two solvers with $\lambda = (0.001\beta/n)\|A^{\mathbb{T}}b\|_{\infty}$, where the solvers use the solution associated to the current λ as the initial point for solving the problem associated to the next λ, and the subfigures in the second line are plotted by the solutions yielded by the GEP-MSCRA with the initial $x^0 = 0$ and the MALSAR with the default one.

Figure 5 shows that the performance of the GEP-MASCRA does not depend on the initial point, but that of the MALSAR improves much if the solution corresponding to the current λ is used as the starting point for solving the problem associated to the next λ. The prediction error of the GEP-MASCRA is at least **20**% lower than that of the MALSAR when the latter does not use the solution corresponding to the current λ as the starting point, and is comparable even superior to that of the MALSAR even if it uses the solution associated to the current λ as the starting point. Also, from the left subfigure in Figure 6, the GEP-MSCRA yields the solution with better group sparsity than the MALSAR does; and from the right subfigure, the MALSAR does not yield a group sparse solution without using the solution associated to the current λ as the next

[2] Available at http://www.mlwin.com/intro/datasets.html

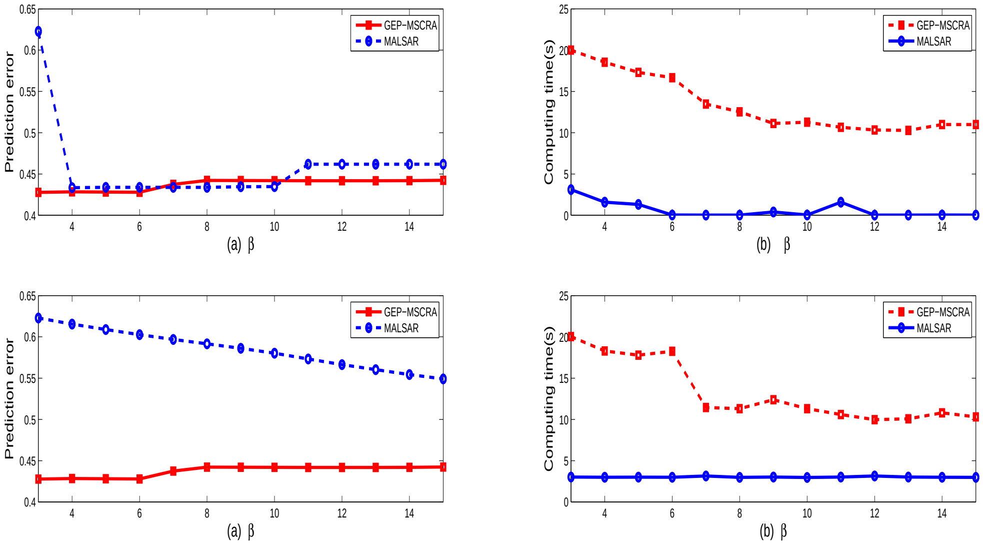


Figure 5: Prediction errors yielded by the GEP-MSCRA and the SLEP under different λ

starting point, but the GEP-MSCRA yields the solution with desirable group sparsity.

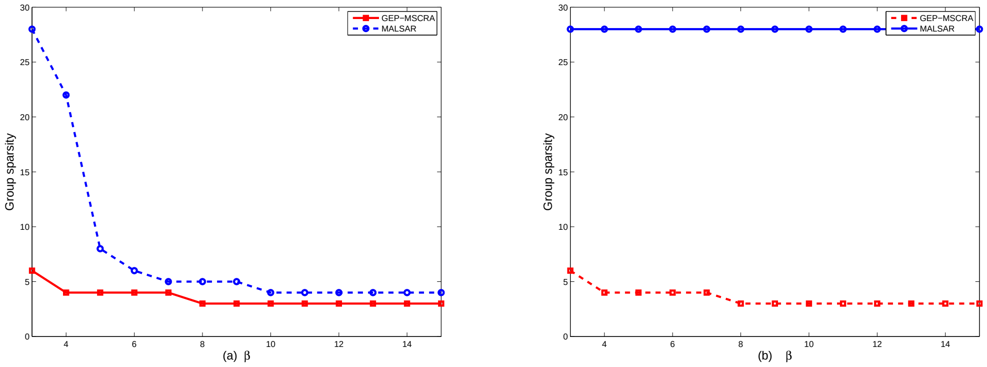


Figure 6: Group sparsity yielded by the GEP-MSCRA and the SLEP under different λ

Next we test the prediction performance of the GEP-MSCRA and the MALSAR with different numbers of training samples. We generate the training and test sets by 10 random splits of the data so that $100\beta\%$ of the examples from each school (task) belong to the training set and $100(1–\beta)\%$ to the test set. The subfigures in the first line of Figure 7 plots the prediction error curves and the computing time curves with $\lambda = (0.005/n)\|A^{\mathbb{T}}b\|_\infty$, and the subfigures in the second line are plotted with $\lambda = (0.013/n)\|A^{\mathbb{T}}b\|_\infty$. We see that the prediction error of the GEP-MSCRA is decreasing as the number of training samples increases, but that of the MALSAR does not improve even increases as the number of training samples increases. Moreover, the prediction error of the GEP-MSCRA is at least lower than **20**% that of the MALSAR when 50% of the examples are used as the training set, and the prediction error of the former is lower than **5**% that of the latter when only 35% of the examples are used as the training set. From Figure 8, we see that under each kind of training samples, the GEP-MSCRA yields the group sparsity less than **5**, but the MALSAR does not yield group sparsity under the two λ.

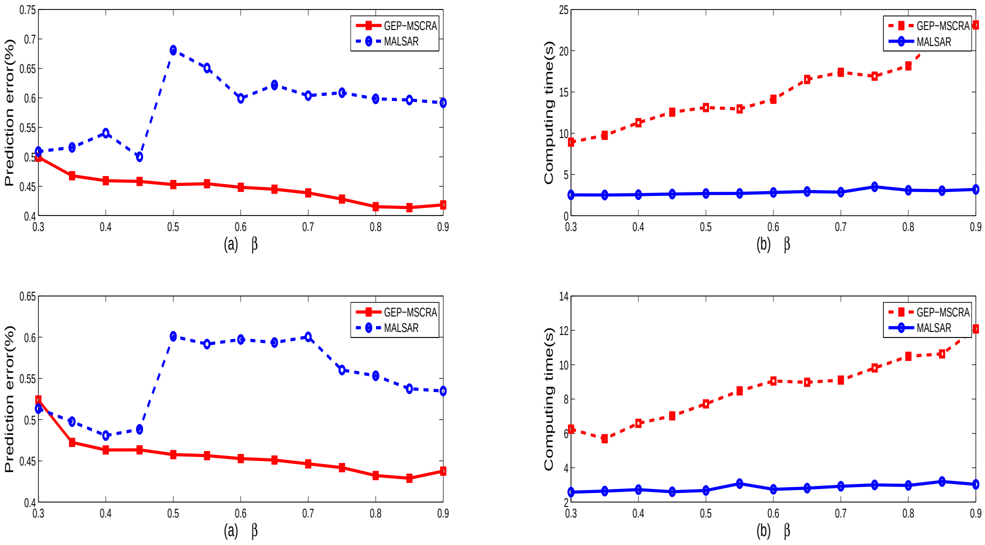


Figure 7: Prediction errors of the GEP-MSCRA and the SLEP under different train samples

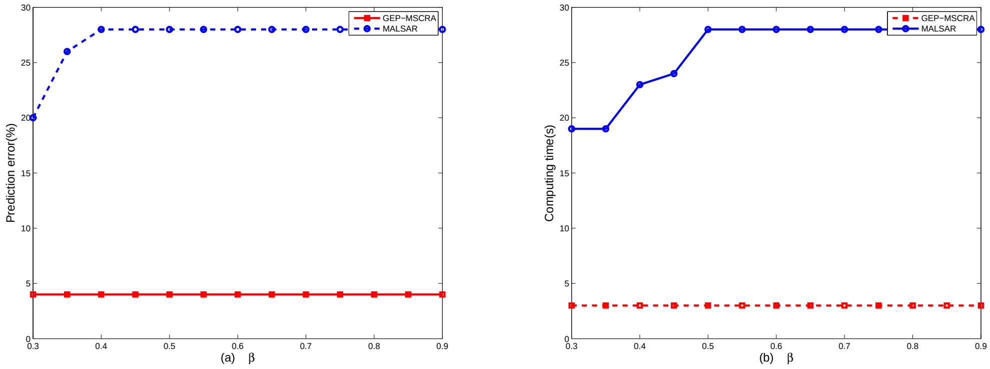


Figure 8: Group sparsity of the GEP-MSCRA and the SLEP under different train samples

6 Conclusions

In this paper we showed that the group zero-norm regularized least squares estimator can be obtained from an exact penalization problem by using the equivalent MPEC of (2) and developing the global exact penalty for the MPEC, and found that the popular SCAD and MCP penalized estimators also arise from the global exact penalty framework. Based on the structure of the exact penalty problem, we proposed a primal-dual convex relaxation approach for computing this estimator. For the proposed GEP-MSCRA, we provided its statistical guarantees and confirmed its efficiency by making comparison with the SLEP and the MALSAR on synthetic group sparse regression problems and real multi-task learning problems. In our future work, we shall further study the global exact penalty results for the MPEC from statistical angle, and develop global exact penalty results for other statistical problems with a certain combinatorial property.

References

[1] A. A. Agarwal, S. Negahban, M. Wainwright, *Fast global convergence of gradient methods for high-dimensional statistical recovery*, The Annals of Statistics, 40(2012): 2452-2482.

[2] A. A. Argyriou, T. Evgeniou and M. Pontil, *Convex multi-task feature learning*, Machine Learning, 73(2008): 243-272.

[3] F. R. Bach, *Consistency of the group Lasso and multiple kernel learning*, Journal of Machine Learning Research, 9(2008): 1179-1225.

[4] P. Breheny and J. Huang, *Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection*, Annals of Applied Statistics, 5(2011): 232-253.

[5] P. Breheny and J. Huang, *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors*, Statistics and Computing, 25(2015): 173-187.

[6] P. Bühlmann and V. D. G. Sara, *Statistics for high-dimensional data: methods, theory and applications*, Springer, 2011.

[7] E. J. Candès and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, Annals of Statistics, 35(2007): 2313-2351.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, New York, 1983.

[9] T. Evgeniou, C. A. Micchelli and M. Pontil, *Learning multiple tasks with kernel methods*, Journal of Machine Learning Research, 6(2005): 615-637.

[10] J. Q. Fan and R. Z. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of American Statistics Association, 96(2001): 1348-1360.

[11] J. Q. Fan, L. Z. Xue and H. Zou, *Strong oracle optimality of folded concave penalized estimation*, Annals of Statistics, 42(2014): 819-849.

[12] P. Gong, J. Ye and C. Zhang, *Multi-stage multi-task feature learning*, Journal of Machine Learning Research, 14(2013): 2979-3010.

[13] J. Huang, J. L. Horowitz and S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, Annals of Statistics, 36(2008): 587-613.

[14] J. Huang and T. Zhang, *The benefit of group sparsity*, The Annals of Statistics, 38(2010): 1978-2004.

[15] X. D. Li, D. F. Sun and K. C. Toh, *A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems*, arXiv:1607.05428v3.

[16] J. Liu, S. Ji and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University. URL: http://www.public.asu.edu/jye02/Software/SLEP (2009)

[17] K. Lounici, M. Pontil, A. B. Tsybakov and S. van de Geer, *Oracle inequalities and optimal inference under group sparsity*, The Annals of Statistics, 39(2011): 2164-2204.

[18] Z. Q. Luo, J. S. Pang and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, 1996.

[19] R. Mazumder, J. H. Friedman and T. Hastie, *SparseNet: Coordinate descent with nonconvex penalties*, Journal of the American Statistical Association, 106 (495): 1125-1138, 2011.

[20] L. Meier, S. Van De Geer, and P. Bühlmann, *The group Lasso for logistic regression*, Journal of the Royal Statistical Society, series B, 70(2008): 53-71.

[21] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM Journal on Control and Optimization, 15(1977): 959-972.

[22] S. Nandy, C. Y. Lim and T. Maiti, *Additive model building for spatial regression*, Journal of the Royal Statistical Society, series B, 79(2017): 779-800.

[23] S. Negahban, P. Ravikumar, M. Wainwright and B. Yu, *A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers*, Statistical Science, 27(2012): 538-557.

[24] Y. Nesterov, *Gradient methods for minimizing composite objective function*, Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL) (2007).

[25] G. Obozinski, B. Taskar and M. I. Jordan, *Joint covariate selection for grouped classification*, Statistics and Computing, 20(2010): 231-252.

[26] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Mathematical Programming, 58(1993): 353-367.

[27] G. Raskutti, M. J. Wainwright and B. Yu, Restricted eigenvalue properties for correlated Gaussian designs, Journal of Machine Learning Research, 11(2010): 2241-2259.

[28] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[29] D. F. Sun, L. Q. Yang and K. C. Toh, *An efficient inexact ABCD method for least squares semidefinite programming*, SIAM Journal on Optimization, 26(2016):1072-1100.

[30] D. F. Sun and J. Sun, *Semismooth matrix-valued functions*, Mathematics of Operations Research, 27(2002): 150-169.

[31] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of Royal Statistical Society B, 58(1996): 267-288.

[32] Y. Yang and H. Zou, *A fast unified algorithm for solving group-lasso penalize learning problems*, Statistical Computation, 15(2015): 1129-1141.

[33] C. Y. Yau and T. S. Hui, *LARS-type algorithm for group lasso*, Statistics & Computing, 4(2017): 1041-1048.

[34] J. J. Ye, D. L. Zhu and Q. J. Zhu, *Exact penalization and necessary optimality conditions for generalized bilevel programming problems*, SIAM Journal on Optimization, 7(1997): 481-507.

[35] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of The Royal Statistical Society, series B, 68(2006): 49-67.

[36] J. Zhang, Z. Ghahramani and Y. Yang, *Flexible latent variable models for multi-task learning*, Machine Learning, 73(2008): 221-242.

[37] T. Zhang, *Multi-stage convex relaxation for feature selection*, Bernoulli, 19(2011): 2277-2293.

[38] T. Zhang, *Analysis of multi-stage convex relaxation for sparse regularization*, Journal of Machine Learning Research, 11(2010): 1081-1107.

[39] C. H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, Annals of Statistics, 38(2010): 894-942.

[40] C. H. Zhang and T. Zhang, *A general theory of concave regularization for high-dimensional sparse estimation problems*, Statistical Science, 27(2012): 576-593, .

[41] H. Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association, 101(2006): 1418-1429.

Appendix A

The following several lemmas provide some upper estimations for the noise vector ε. Among others, Lemma 1 follows directly by using the same arguments as [38, Lemma 5], and Lemma 2 and 3 follow from the same arguments as those for [37, Lemma 3].

Lemma 1 *Let* $\|\mathcal{J}\|_\infty := \max_{1\le i\le m}\{|\mathcal{J}_i|\}$. *Then, under Assumption 1, for any given* $\eta\in(0,1)$ *the following inequality holds with probability at least* $1-\eta$*:*

$$\|\mathcal{G}(\widehat{\varepsilon})\|_\infty \le \frac{\sigma}{n}\sqrt{2\|\mathcal{J}\|_\infty\log(2p/\eta)}\,\|A\|_{2,\infty}.$$

Lemma 2 *Suppose that* $A_{\mathcal{J}_{\overline{S}}}$ *has full column rank. Then, under Assumption 1, for any given* $\eta\in(0,1)$ *the following inequality holds with probability (w.p.) at least* $1-\eta$*:*

$$\|\mathcal{G}(\widehat{\varepsilon}^{\dagger})\|_\infty \le \frac{\sigma}{\sigma_{\min}(A_{\mathcal{J}_{\overline{S}}})}\sqrt{2\|\mathcal{J}\|_\infty\log(2p/\eta)}.$$

Lemma 3 *Define* $\varepsilon^{\rm LS} := \frac{1}{n}A^{\mathbb{T}}(Ax^{\rm LS}-b)$ *where* $x^{\rm LS}$ *is the solution defined by (17). Under Assumption 1, for any given* $\eta\in(0,1)$ *the following inequality holds w.p. at least* $1-\eta$*:*

$$\varepsilon^{\rm LS}_{\mathcal{J}_i}=0\ \ \text{for } i\in\overline{S}\ \text{and}\ \|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty \le \frac{\sigma\sqrt{2\|\mathcal{J}\|_\infty\log(2p/\eta)}}{n}\|A\|_{2,\infty}\ \text{for } i\notin\overline{S}.$$

Appendix B

Next we shall provide the proof of Theorem 2.1. This requires three technical lemmas. The first two characterize some important properties of the function family Φ.

Lemma 4 *Let* $\phi\in\Phi$. *Then, the set* $(\partial\phi)^{-1}(\frac{1}{1-t_\phi^*})\cap[t_\phi^*,1)$ *is nonempty and compact.*

Proof: Since $[0,1]\subseteq{\rm int}({\rm dom}\phi)$, from [28, Theorem 23.4] $\partial\phi(t)=[\phi'_-(t),\phi'_+(t)]$ is nonempty and bounded for each $t\in[0,1]$. We first argue that $(\partial\phi)^{-1}(\frac{1}{1-t_\phi^*})\cap[t_\phi^*,1)\ne\emptyset$. Assume that there exists $\overline{t}\in(t_\phi^*,1)$ such that $\phi'_-(\overline{t})<\phi'_-(1)$ (if not, we will have $\partial\phi(t)=\{\phi'(t)\}=\{\phi'_-(1)\}$ for all $t\in(t_\phi^*,1)$, and hence there exists $\xi\in(t_\phi^*,1)$ such that $\phi'(\xi)=\frac{\phi(1)}{1-t_\phi^*}$, which implies the desired statement). Together with the convexity of ϕ and [28, Theorem 24.1], we have $\phi'_-(t)\le\phi'_-(\overline{t})$ for all $t\in[t_\phi^*,\overline{t}]$. By [28, Corollary 24.2.1],

$$\begin{aligned}\phi(1)=\phi(1)-\phi(t_\phi^*)&=\int_{t_\phi^*}^{1}\phi'_-(t)dt=\int_{t_\phi^*}^{\overline{t}}\phi'_-(t)dt+\int_{\overline{t}}^{1}\phi'_-(t)dt\\&<\phi'_-(1)(\overline{t}-t_\phi^*)+\int_{\overline{t}}^{1}\phi'_-(t)dt\le\phi'_-(1)(1-t_\phi^*).\end{aligned}$$

In addition, by the convexity of ϕ, $\phi(1) \geq \phi(t^*_\phi)+\phi'_+(t^*_\phi)(1-t^*_\phi) = \phi'_+(t^*_\phi)(1-t^*_\phi)$. Thus, $a := \frac{\phi(1)}{1-t^*_\phi} = \frac{1}{1-t^*_\phi} \in [\phi'_+(t^*_\phi), \phi'_-(1))$. If $a = \phi'_+(t^*_\phi)$, clearly, $t^*_\phi \in (\partial\phi)^{-1}(\frac{1}{1-t^*_\phi}) \cap [t^*_\phi, 1)$. So, it suffices to consider the case $a \in (\phi'_+(t^*_\phi), \phi'_-(1))$. Now $(\partial\phi)^{-1}(a) \cap [0, 1) \neq \emptyset$ (if not, $a \in \partial\phi(t') = [\phi'_-(t'), \phi'_+(t')]$ for $t' \geq 1$ or $t' < t^*_\phi$, which contradicts $a \in (\phi'_+(t^*_\phi), \phi'_-(1))$).

Next we show that $(\partial\phi)^{-1}(\frac{1}{1-t^*_\phi}) \cap [t^*_\phi, 1)$ is compact. Fix an arbitrary $b \in (\partial\phi)^{-1}(\frac{1}{1-t^*_\phi})$. Since $(\partial\phi)^{-1}(\frac{1}{1-t^*_\phi})$ is compact, we only need to argue that $b < 1$. This clearly holds by noting that $a = \frac{1}{1-t^*_\phi} \in \partial\phi(b) = [\phi'_-(b), \phi'_+(b)]$ and $a \in [\phi'_+(t^*_\phi), \phi'_-(1))$. □

Lemma 5 *Let $\phi \in \Phi$. For any given $\omega \geq 0$, define $\upsilon^* := \min\limits_{t\in[0,1]} \{\phi(t) + \omega(1-t)\}$. Then,*

$$\begin{cases} \upsilon^* = 1 & \text{if } \omega \in (\phi'_-(1), +\infty); \\ \upsilon^* \geq \frac{\omega(1-\bar{t}_\phi)}{\phi'_-(1)(1-t^*_\phi)} & \text{if } \omega \in \big[\frac{1}{1-t^*_\phi}, \phi'_-(1)\big]; \\ \upsilon^* \geq \omega(1-\bar{t}_\phi) & \text{if } \omega \in \big[0, \frac{1}{1-t^*_\phi}\big). \end{cases}$$

Proof: When $\omega > \phi'_-(1)$, clearly, $\upsilon^* = \phi(1)$ since $\phi(t) + \omega(1-t)$ is nonincreasing in $[0, 1]$. When $\omega \in \big[0, \frac{1}{1-t^*_\phi}\big)$, since $\phi'_-(t) \geq \phi'_+(\bar{t}_\phi) > \omega$ for any $t > \bar{t}_\phi$ by Lemma 4, the optimal solution $\widehat{t}$ of $\min_{t\in[0,1]}\{\phi(t) + \omega(1-t)\}$ satisfies $\widehat{t} \leq \bar{t}_\phi$. By the convexity of ϕ,

$$\phi(t) + \omega(1-t) \geq \phi(\widehat{t}) + \omega(1-\widehat{t}) \geq \omega(1-\bar{t}_\phi) \quad \forall t \in [0, 1].$$

This shows that $\upsilon^* \geq \omega(1-\bar{t}_\phi)$ for this case. When $\omega \in \big[\frac{1}{1-t^*_\phi}, \phi'_-(1)\big]$, by Lemma 4

$$\upsilon^* \geq \min_{t\in[0,1]} \Big\{\phi(t) + \frac{1}{1-t^*_\phi}(1-t)\Big\} = \phi(\bar{t}_\phi) + \frac{1}{1-t^*_\phi}(1-\bar{t}_\phi) \geq \frac{1-\bar{t}_\phi}{1-t^*_\phi} \geq \frac{\omega(1-\bar{t}_\phi)}{\phi'_-(1)(1-t^*_\phi)},$$

where the last inequality is due to $\omega \leq \phi'_-(1)$. The proof is completed. □

For every $x \in \Omega$, with a parameter $\rho > 0$ we define a truncated vector $x^\rho \in \Omega$ by

$$(x^\rho)_{\mathcal{J}_i} = \begin{cases} x_{\mathcal{J}_i} & \text{if } \|x_{\mathcal{J}_i}\| > \frac{\phi'_-(1)}{\rho}, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the following result holds for the objective value of (2) at x^ρ and that of (6) at x.

Lemma 6 *Let $\phi \in \Phi$. Then, for any $x \in \Omega$ and $w \in [0, e]$, when $\rho > \overline{\rho} = \nu L_f \frac{(1-t^*_\phi)\phi'_-(1)}{1-\bar{t}_\phi}$,*

$$\nu f(x^\rho) + \|\mathcal{G}(x^\rho)\|_0 \leq \nu f(x) + \textstyle\sum_{i=1}^m \big[\phi(w_i) + \rho(1-w_i)\|x_{\mathcal{J}_i}\|\big], \tag{29}$$

and moreover, $x^\rho = x$ and $\|\mathcal{G}(x)\|_1 - \langle w, \mathcal{G}(x)\rangle = 0$ provided that (29) becomes an equality.

Proof: Fix arbitrary $x \in \Omega$, $w \in [0, e]$ and $\rho > \overline{\rho}$. Applying Lemma 5 with $\omega = \rho\|x_{\mathcal{J}_i}\|$ for every $i \in \{1, \ldots, m\}$ delivers

$$\begin{cases} \phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\| \geq 1 & \text{if } i \in I_1; \\ \phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\| \geq \frac{(1-\overline{t}_\phi)\rho\|x_{\mathcal{J}_i}\|}{\phi'_-(1)(1-t^*_\phi)} & \text{if } i \in I_2; \\ \phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\| \geq \rho\|x_{\mathcal{J}_i}\|(1-\overline{t}_\phi) & \text{if } i \in I_3. \end{cases} \tag{30}$$

where $I_1 := \big\{i : \rho\|x_{\mathcal{J}_i}\| > \phi'_-(1)\big\}$, $I_2 := \big\{i : \rho\|x_{\mathcal{J}_i}\| \in [\frac{1}{1-t^*_\phi}, \phi'_-(1)]\big\}$ and $I_3 := (I_1 \cup I_2)^c$. From the expression of x^ρ and $\rho > \nu L_f \frac{(1-t^*_\phi)\phi'_-(1)}{1-\overline{t}_\phi} \geq \frac{\nu L_f}{1-\overline{t}_\phi}$, it immediately follows that

$$\sum_{i\in I_1} \big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big] = \|\mathcal{G}(x^\rho)\|_0 \;\text{ and }\; \sum_{i\in I_2\cup I_3} \big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big] \geq \nu L_f\|x_{\mathcal{J}_i}\|.$$

Together with $|f(x)-f(x^\rho)| \leq L_f\|x-x^\rho\|$ by the Lipschitz continuity of f, we have

$$\begin{aligned} &\sum_{i=1}^{m} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] - \|\mathcal{G}(x^\rho)\|_0 \\ &= \sum_{i\in I_1\cup I_2\cup I_3} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] - \|\mathcal{G}(x^\rho)\|_0 \\ &\geq \sum_{i\in I_2\cup I_3} \nu L_f\|x_{\mathcal{J}_i}\| = \nu L_f\|\mathcal{G}(x) - \mathcal{G}(x^\rho)\|_1 \\ &\geq \nu L_f\|x - x^\rho\| \geq \nu|f(x) - f(x^\rho)|. \end{aligned} \tag{31}$$

By the arbitrariness of $x \in \Omega, w \in [0, e]$ and $\rho > \overline{\rho}$, the first part of conclusions follows.

Next we prove the second part. Since inequality (29) becomes an equality, i.e.,

$$\nu|f(x^\rho) - f(x)| = \textstyle\sum_{i=1}^{m} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] - \|\mathcal{G}(x^\rho)\|_0, \tag{32}$$

together with inequality (31) it immediately follows that

$$\sum_{i\in I_2\cup I_3} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] = \sum_{i\in I_2\cup I_3} \nu L_f\|x_{\mathcal{J}_i}\|. \tag{33}$$

Suppose on the contradiction that $x \neq x^\rho$. Then there exists an index $k \in I_2 \cup I_3$ such that $\|x_{\mathcal{J}_k}\| \neq 0$. By (30) and $\rho > \nu L_f \frac{(1-t^*_\phi)\phi'_-(1)}{1-\overline{t}_\phi}$, $\phi(w_k) + \rho(1 - w_k)\|x_{\mathcal{J}_k}\| > \nu L_f\|x_{\mathcal{J}_k}\|$. Together with $\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\| \geq \nu L_f\|x_{\mathcal{J}_i}\|$ for all $i \in I_2 \cup I_3$, we obtain

$$\sum_{i\in I_2\cup I_3} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] > \sum_{i\in I_2\cup I_3} \nu L_f\|x_{\mathcal{J}_i}\|,$$

which contradicts (33). Substituting $x = x^\rho$ into (32) and using the definition of x^ρ yields

$$\sum_{i\in I_1\cup I_3} \big[\phi(w_i) + \rho(1 - w_i)\|x_{\mathcal{J}_i}\|\big] = \|\mathcal{G}(x^\rho)\|_0.$$

Notice that $\phi(w_i)\geq\phi(1)+\phi'_{-}(1)(w_i-1)\geq\phi(1)-\rho\|x_{\mathcal{J}_i}\|(1-w_i)$ for every $i\in I_1$, and hence $\sum_{i\in I_1}\big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big]\geq|I_1|=\|\mathcal{G}(x)\|_0$. Together with $\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\geq 0$ for $i\in I_3$, the last equality implies that $\phi(w_i)=1$ for $i\in I_1$ and $\sum_{i\in I_3}\phi(w_i)=0$. Clearly, the latter is equivalent to saying that $w_i=t_\phi^*$ for $i\in I_3$. Now from (31) we get

$$\textstyle\sum_{i=1}^m\big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big]=|I_1|+\sum_{i=1}^m\rho(1-w_i)\|x_{\mathcal{J}_i}\|=\|\mathcal{G}(x^\rho)\|_0.$$

This means that $\sum_{i=1}^m\rho(1-w_i)\|x_{\mathcal{J}_i}\|=0$. Thus, we complete the proof. □

The proof of Theorem 2.1: Fix an arbitrary $\rho>\overline{\rho}$. Let $\mathcal{S}$ be the feasible set of (4), and let $\mathcal{S}_\rho$ be that of (6) associated to ρ. We first prove that $\mathcal{S}^*\subseteq\mathcal{S}_\rho^*$. Fix an arbitrary $(\overline{x},\overline{w})\in\mathcal{S}^*$. Then, $\overline{x}$ is globally optimal to (2) and $\|\mathcal{G}(\overline{x})\|_0=\sum_{i=1}^m\phi(\overline{w}_i)$. Let (x,w) be an arbitrary point from $\mathcal{S}_\rho$. Assume that x^ρ be defined as in Lemma 6. Then

$$\begin{aligned}\nu f(x)+\textstyle\sum_{i=1}^m\big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big]&\geq\nu f(x^\rho)+\|\mathcal{G}(x^\rho)\|_0\geq\nu f(\overline{x})+\|\mathcal{G}(\overline{x})\|_0\\&=\nu f(\overline{x})+\sum_{i=1}^m\big[\phi(\overline{w}_i)+\rho(1-\overline{w}_i)\|x_{\mathcal{J}_i}\|\big],\end{aligned}$$

where the second inequality is due to $x^\rho\in\Omega$. Notice that $(\overline{x},\overline{w})\in\mathcal{S}_\rho$ and (x,w) is an arbitrary point from $\mathcal{S}_\rho$. The last inequality shows that $(\overline{x},\overline{w})\in\mathcal{S}_\rho^*$, and then $\mathcal{S}^*\subseteq\mathcal{S}_\rho^*$.

We next prove $\mathcal{S}_\rho^*\subseteq\mathcal{S}^*$. Fix an arbitrary $(\overline{x},\overline{w})\in\mathcal{S}_\rho^*$. Define $\overline{w}^\rho\in\mathbb{R}^m$ by

$$\overline{w}_i^\rho:=\begin{cases}1 & \text{if } \rho\|\overline{x}_{\mathcal{J}_i}\|>\phi'_{-}(1);\\ t_\phi^* & \text{if } \rho\|\overline{x}_{\mathcal{J}_i}\|\leq\phi'_{-}(1),\end{cases}\quad\text{for}\ \ i=1,2,\ldots,m.$$

Then, we have $\sum_{i=1}^m\phi(\overline{w}_i^\rho)=\|\mathcal{G}(\overline{x})\|_0$ and $\|\mathcal{G}(\overline{x}^\rho)\|_1=\sum_{i=1}^m\overline{w}_i^\rho\|\overline{x}_{\mathcal{J}_i}^\rho\|$, where $\overline{x}^\rho$ is defined as in Lemma 6 with $x=\overline{x}$. From the results of Lemma 6, it follows that

$$\begin{aligned}&\nu f(\overline{x})+\textstyle\sum_{i=1}^m\big[\phi(\overline{w}_i)+\rho(1-\overline{w}_i)\|\overline{x}_{\mathcal{J}_i}\|\big]\geq\nu f(\overline{x}^\rho)+\|\mathcal{G}(\overline{x}^\rho)\|_0\\&=\nu f(\overline{x}^\rho)+\textstyle\sum_{i=1}^m\big[\phi(\overline{w}_i^\rho)+\rho(1-\overline{w}_i^\rho)\|\overline{x}_{\mathcal{J}_i}^\rho\|\big]\\&\geq\nu f(\overline{x})+\textstyle\sum_{i=1}^m\big[\phi(\overline{w}_i)+\rho(1-\overline{w}_i)\|\overline{x}_{\mathcal{J}_i}\|\big],\end{aligned}$$

where the last inequality is due to $(\overline{x}^\rho,\overline{w}_i^\rho)\in\mathcal{S}_\rho$. The last inequality implies that

$$\nu f(\overline{x})+\textstyle\sum_{i=1}^m\big[\phi(\overline{w}_i)+\rho(1-\overline{w}_i)\|\overline{x}_{\mathcal{J}_i}\|\big]=\nu f(\overline{x}^\rho)+\|\mathcal{G}(\overline{x}^\rho)\|_0.$$

Using Lemma 6 again, we have $\overline{x}=\overline{x}^\rho$ and $\|\mathcal{G}(\overline{x})\|_1-\sum_{i=1}^m\overline{w}_i\|\overline{x}_{\mathcal{J}_i}\|=0$, which implies $(\overline{x},\overline{w})\in\mathcal{S}$. Now let (x,w) be an arbitrary point from $\mathcal{S}$. Then $(x,w)\in\mathcal{S}_\rho$, and we have

$$\begin{aligned}\nu f(x)+\textstyle\sum_{i=1}^m\phi(w_i)&=\nu f(x)+\textstyle\sum_{i=1}^m\big[\phi(w_i)+\rho(1-w_i)\|x_{\mathcal{J}_i}\|\big]\\&\geq\nu f(\overline{x})+\textstyle\sum_{i=1}^m\big[\phi(\overline{w}_i)+\rho(1-\overline{w}_i)\|\overline{x}_{\mathcal{J}_i}\|\big]\\&=\nu f(\overline{x})+\textstyle\sum_{i=1}^m\phi(\overline{w}_i).\end{aligned}$$

Notice that $(\overline{x},\overline{w})\in\mathcal{S}$. From the last inequality and the arbitrariness of (x,w) in $\mathcal{S}$, it follows that $(\overline{x},\overline{w})\in\mathcal{S}^*$. Thus, by the arbitrariness of $(\overline{x},\overline{w})$ in $\mathcal{S}_\rho^*$, we obtain that $\mathcal{S}_\rho^*\subseteq\mathcal{S}^*$. Together with $\mathcal{S}^*\subseteq\mathcal{S}_\rho^*$, we complete the proof of theorem. □

Appendix C.

To achieve the results of Theorem 4.1 and Theorem 4.2, we need to establish the following two lemmas where $\delta^k := x^k-\overline{x}$ and $v^k = e - w^k$ for $k\ge 1$. The first one states a relation between $\sum_{i\in(S^{k-1})^c}\|\delta^k_{\mathcal{J}_i}\|$ and $\sum_{i\in S^{k-1}}\|\delta^k_{\mathcal{J}_i}\|$ where $S^{k-1}\supset\overline{S}$ is an index set.

Lemma 7 *For $k\ge 1$, if there is an index set $S^{k-1}\supseteq\overline{S}$ such that $\min_{i\in(S^{k-1})^c} w_i^{k-1}\le\overline{t}_\phi$, then with $\lambda^{k-1}\ge\frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$ it holds that $\sum_{i\in(S^{k-1})^c}\|\delta^k_{\mathcal{J}_i}\|\le\frac{2}{1-\overline{t}_\phi}\sum_{i\in S^{k-1}}\|\delta^k_{\mathcal{J}_i}\|$.*

Proof: By the optimality of x^k and the feasibility of $\overline{x}$ to the subproblem (12), we have

$$\frac{1}{2n}\big\|Ax^k-b\big\|^2+\lambda^{k-1}\sum_{i=1}^m v_i^{k-1}\|x^k_{\mathcal{J}_i}\|\le\frac{1}{2n}\big\|A\overline{x}-b\big\|^2+\lambda^{k-1}\sum_{i=1}^m v_i^{k-1}\|\overline{x}_{\mathcal{J}_i}\|$$

which, by using $\delta^k=x^k-\overline{x}$, $\varepsilon=b-A\overline{x}$ and $\widehat{\varepsilon}=\frac{1}{n}A^{\mathbb{T}}\varepsilon$, can be rearranged as follows:

$$\frac{1}{2n}\big\|A\delta^k\big\|^2\le\langle\widehat{\varepsilon},\delta^k\rangle+\lambda^{k-1}\sum_{i=1}^m v_i^{k-1}\Big(\|\overline{x}_{\mathcal{J}_i}\|-\|x^k_{\mathcal{J}_i}\|\Big).$$

Together with $\overline{x}_{\mathcal{J}_i}=0$ for all $i\in\overline{S}^c$ and the definition of $\mathcal{G}(\cdot)$, we obtain that

$$\begin{aligned}\frac{1}{2n}\big\|A\delta^k\big\|^2&\le\sum_{i=1}^m\langle\widehat{\varepsilon}_{\mathcal{J}_i},\delta^k_{\mathcal{J}_i}\rangle+\lambda^{k-1}\sum_{i\in\overline{S}}v_i^{k-1}\Big(\|\overline{x}_{\mathcal{J}_i}\|-\|x^k_{\mathcal{J}_i}\|\Big)-\lambda^{k-1}\sum_{i\in\overline{S}^c}v_i^{k-1}\|x^k_{\mathcal{J}_i}\|\\&\le\langle\mathcal{G}(\widehat{\varepsilon}),\mathcal{G}(\delta^k)\rangle+\lambda^{k-1}\sum_{i\in\overline{S}}v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\|-\lambda^{k-1}\sum_{i\in(S^{k-1})^c}v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\|\\&\le\sum_{i\in S^{k-1}\backslash\overline{S}}\|\widehat{\varepsilon}_{\mathcal{J}_i}\|\|\delta^k_{\mathcal{J}_i}\|+\big(\lambda^{k-1}+\|\mathcal{G}(\widehat{\varepsilon})\|_\infty\big)\sum_{i\in\overline{S}}\|\delta^k_{\mathcal{J}_i}\|\\&\quad+\big[\|\mathcal{G}(\widehat{\varepsilon})\|_\infty-\lambda^{k-1}(1-\overline{t}_\phi)\big]\textstyle\sum_{i\in(S^{k-1})^c}\|\delta^k_{\mathcal{J}_i}\|\end{aligned}\tag{34}$$

where the last inequality are due to $\min_{i\in(S^{k-1})^c}v_i^{k-1}\ge 1-\overline{t}_\phi$. Then, it holds that

$$\begin{aligned}&\frac{1}{2n}\|A\delta^k\|^2+\Big[\lambda^{k-1}(1-\overline{t}_\phi)-\|\mathcal{G}(\widehat{\varepsilon})\|_\infty\Big]\sum_{i\in(S^{k-1})^c}\|\delta^k_{\mathcal{J}_i}\|\\&\le\sum_{i\in S^{k-1}\backslash\overline{S}}\|\widehat{\varepsilon}_{\mathcal{J}_i}\|\|\delta^k_{\mathcal{J}_i}\|+(\lambda^{k-1}+\|\mathcal{G}(\widehat{\varepsilon})\|_\infty)\sum_{i\in\overline{S}}\|\delta^k_{\mathcal{J}_i}\|\le(\lambda^{k-1}+\|\mathcal{G}(\widehat{\varepsilon})\|_\infty)\sum_{i\in S^{k-1}}\|\delta^k_{\mathcal{J}_i}\|\end{aligned}$$

Together with $\frac{1}{2n}\|A\delta^k\|^2\ge 0$ and $\lambda^{k-1}\ge\frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$, we get the desired inequality. □

When S^{k-1} in Lemma 7 is also such that the matrix A satisfies the RSC in $\mathcal{C}(\overline{S},|S^{k-1}|)$ with constant $\gamma_k>0$, the result of Lemma 7 can be strengthened as follows.

Lemma 8 *For $k \ge 1$, if there is an index set $S^{k-1} \supseteq \overline{S}$ such that $\min_{i\in(S^{k-1})^c} w_i^{k-1} \le \overline{t}_\phi$ and A has the RSC over $\mathcal{C}(\overline{S}, |S^{k-1}|)$ with constant $\gamma_k > 0$, then with $\lambda^{k-1} \ge \frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$*

$$\|\delta^k\| \le \frac{1}{\gamma_k}\Big(\big\|\,[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\,\big\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\Big).$$

Proof: Using inequality (34) and noting that $\delta^k \in \mathcal{C}(\overline{S}, |S^{k-1}|)$ by Lemma 7, we have

$$\begin{aligned}
\gamma_k\|\delta^k\|^2 \le \frac{1}{2n}\|A\delta^k\|^2 &\le \sum_{i=1}^m \|\widehat{\varepsilon}_{\mathcal{J}_i}\|\|\delta^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\| - \lambda^{k-1}\sum_{i\notin S^{k-1}} v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\| \\
&\le \sum_{i=1}^m \|\widehat{\varepsilon}_{\mathcal{J}_i}\|\|\delta^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\| - \lambda^{k-1}(1-\overline{t}_\phi)\sum_{i\notin S^{k-1}} \|\delta^k_{\mathcal{J}_i}\| \\
&\le \textstyle\sum_{i\in S^{k-1}} \|\widehat{\varepsilon}_{\mathcal{J}_i}\|\|\delta^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\delta^k_{\mathcal{J}_i}\| \\
&\le \sqrt{\textstyle\sum_{i\in S^{k-1}}\|\widehat{\varepsilon}_{\mathcal{J}_i}\|^2}\,\|\delta^k\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\,\|\delta^k\|
\end{aligned}$$

where the third inequality is by $\lambda^{k-1} \ge \frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$. This implies the desired result. □

The proof of Theorem 4.1: For each $k \in \mathbb{N}$, define $S^{k-1} := \overline{S} \cup \{i \notin \overline{S} \colon w_i^{k-1} > \overline{t}_\phi\}$. Notice that $\nu \le \frac{1-\overline{t}_\phi}{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}$ and $\frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi} \le \rho\nu^{-1}$. We have $\lambda^{k-1} \ge \frac{(3-\overline{t}_\phi)\|\mathcal{G}(\widehat{\varepsilon})\|_\infty}{1-\overline{t}_\phi}$ for all $k \in \mathbb{N}$. If $|S^{k-1}| \le 1.5\overline{r}$ for some $k \in \mathbb{N}$, from Lemma 8 it follows that

$$\begin{aligned}
\|x^k - \overline{x}\| &\le \frac{1}{\kappa}\Big(\big\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\big\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\Big) \\
&\le \frac{1}{\kappa}\Big(\big\|\mathcal{G}(\widehat{\varepsilon})\big\|_\infty\sqrt{1.5\overline{r}} + \lambda^{k-1}\sqrt{\overline{r}}\Big) \le \frac{\lambda^{k-1}(4-2\overline{t}_\phi)}{\kappa(3-\overline{t}_\phi)}\sqrt{1.5\overline{r}},
\end{aligned} \tag{35}$$

where the last inequality is due to $\|\mathcal{G}(\widehat{\varepsilon})\|_\infty \le \frac{1-\overline{t}_\phi}{3-\overline{t}_\phi}\lambda^{k-1}$ for all $k \in \mathbb{N}$. So, it suffices to argue that $|S^{k-1}| \le 1.5\overline{r}$ for all $k \in \mathbb{N}$. When $k = 1$, it automatically holds since $S^0 = \overline{S}$ by $w^0 \le \overline{t}_\phi e$. Now assume that $|S^{k-1}| \le 1.5\overline{r}$ for all $k = l$ with $l \ge 1$. We shall prove that $|S^l| \le 1.5\overline{r}$. Using (35) with $k = l$, we have $\|x^l - \overline{x}\| \le \frac{\lambda^{l-1}(4-2\overline{t}_\phi)\sqrt{1.5\overline{r}}}{\kappa(3-\overline{t}_\phi)}$. Notice that $i \in S^l\backslash\overline{S}$ implies $i \notin \overline{S}$ and $w_i^l \in (\overline{t}_\phi, 1]$. From $w_i^l \in \partial\psi^*(\rho\|x^l_{\mathcal{J}_i}\|) = (\partial\psi)^{-1}(\rho\|x^l_{\mathcal{J}_i}\|)$,

$$\rho\|x^l_{\mathcal{J}_i}\| \ge \psi'_-(w_i^l) = \phi'_-(w_i^l) \ge \phi'_+(\overline{t}_\phi) \ge \frac{1}{1-t^*_\phi},$$

where the equality is due to $\psi'_-(t) = \phi'(t)$ for all $t \in (0,1]$. This inequality implies

$$\begin{aligned}
\sqrt{|S^l\backslash\overline{S}|} &\le \sqrt{\textstyle\sum_{i\in S^l\backslash\overline{S}}\rho^2(1-t^*_\phi)^2\|x^l_{\mathcal{J}_i}\|^2} \le \rho(1-t^*_\phi)\|x^l - \overline{x}\| \\
&\le \frac{\rho\lambda^{l-1}(1-t^*_\phi)(4-2\overline{t}_\phi)}{\kappa(3-\overline{t}_\phi)}\sqrt{1.5\overline{r}} \le \sqrt{0.5\overline{r}},
\end{aligned} \tag{36}$$

where the last inequality is due to $\rho\lambda^{l-1} \le \frac{(3-\bar{t}_\phi)\kappa}{\sqrt{3}(4-2\bar{t}_\phi)(1-t^*_\phi)}$ implied by $\lambda^{l-1} = \nu^{-1}$ for $l = 1$ and $\lambda^{l-1} = \rho\nu^{-1}$ for $l > 1$. So, $|S^l| \le 1.5\overline{r}$. Thus, $|S^{k-1}| \le 1.5\overline{r}$ holds for all $k \in \mathbb{N}$. □

In the following, we upper bound $(v_i^k)^2$ for $i \in \overline{S}$ by means of $\mathbb{I}_\Delta(i)$ and $\mathbb{I}_{F^k}(i)$.

Lemma 9 *For each $k \ge 1$, let F^k be the index set defined as in (15). Then, it holds that*

$$\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^k)^2} \le \sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)} + \sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_{F^k}(i)}.$$

Proof: Notice that $v_i^k = 1 - w_i^k \le 1$. If $i \in F^k$, clearly, $v_i^k \le \mathbb{I}_{F^k}(i)$. Otherwise, together with Remark 3.1(a) and $v_i^k = 1 - w_i^k$, it follows that

$$v_i^k \le \mathbb{I}_{\left\{i:\, \|x^k_{\mathcal{J}_i}\| \le \phi'_-(1)/\rho\right\}}(i) \le \mathbb{I}_{\left\{i:\, \|\overline{x}_{\mathcal{J}_i}\| \le \frac{1}{\rho(1-t^*_\phi)} + \frac{\phi'_-(1)}{\rho}\right\}}(i) \le \mathbb{I}_\Delta(i).$$

Hence, for each i, it holds that $0 \le v_i^k \le \mathbb{I}_\Delta(i) + \mathbb{I}_{F^k}(i)$. The desired result follows by noting that $\|a + b\| \le \|a\| + \|b\|$ for all vectors a and b. □

The proof of Theorem 4.2: For each $k \in \mathbb{N}$, define $S^{k-1} := \overline{S} \cup \{i \notin \overline{S} : w_i^{k-1} > \bar{t}_\phi\}$. Since the conclusion holds for $k = 1$, it suffices to consider $k \ge 2$. Now, from (36),

$$\begin{aligned}
\big\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\big\| &\le \big\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\big\| + \big\|\mathcal{G}(\widehat{\varepsilon})\big\|_\infty \sqrt{|S^{k-1}\backslash\overline{S}|} \le \|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\| + \frac{\lambda^{k-1}(1-\bar{t}_\phi)}{3-\bar{t}_\phi}\sqrt{|S^{k-1}\backslash\overline{S}|} \\
&\le \big\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\big\| + \rho(1-t^*_\phi)\lambda^{k-1}\frac{1-\bar{t}_\phi}{3-\bar{t}_\phi}\big\|x^{k-1} - \overline{x}\big\|. \qquad (37)
\end{aligned}$$

In addition, from Lemma 8 and Lemma 9, it follows that

$$\begin{aligned}
&\|x^k - \overline{x}\| \le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\Big) \\
&\le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)} + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_{F^k}(i)}\Big) \\
&\le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)} + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\big[(1-t^*_\phi)^2\|\|x^{k-1}_{\mathcal{J}_i}\| - \|\overline{x}_{\mathcal{J}_i}\|\|^2\rho^2\big]}\Big) \\
&\le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{S^{k-1}}\| + \rho(1-t^*_\phi)\lambda^{k-1}\|x^{k-1} - \overline{x}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)}\Big).
\end{aligned}$$

where the third inequality is by the definition of F^k. Together with (37), we obtain

$$\begin{aligned}
\|x^k - \overline{x}\| &\le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\| + \frac{\rho\lambda^{k-1}(1-t^*_\phi)(4-2\bar{t}_\phi)\|x^{k-1} - \overline{x}\|}{3-\bar{t}_\phi} + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)}\Big) \\
&\le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)}\Big) + \frac{1}{\sqrt{3}}\|\delta^{k-1}\| \\
&= \frac{1}{\kappa}\Big(\|[\mathcal{G}(\widehat{\varepsilon})]_{\overline{S}}\| + \rho\nu^{-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_\Delta(i)}\Big) + \frac{1}{\sqrt{3}}\|x^{k-1} - \overline{x}\|
\end{aligned}$$

where the second inequality is using $\rho\lambda^{k-1} \le \frac{(3-\bar{t}_\phi)\kappa}{\sqrt{3}(4-2\bar{t}_\phi)(1-t_\phi^*)}$, and the last one is using $\lambda^{k-1} = \rho\nu^{-1}$ for $k \ge 2$. The desired result follows by this recursion inequality. □

In order to achieve the result of Theorem 4.3, we also need the following two technical lemmas where $\widehat{\delta}^k := x^k - x^{\rm LS}$ for $k = 1, 2, \ldots$.

Lemma 10 *For $k \ge 1$, if there is an index set $S^{k-1} \supseteq \overline{S}$ such that $\min_{i\in(S^{k-1})^c} w_i^{k-1} \le \bar{t}_\phi$, then with $\lambda^{k-1} \ge \frac{2\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty}{1-\bar{t}_\phi}$ it holds that $\sum_{i\in(S^{k-1})^c} \|\widehat{\delta}^k_{\mathcal{J}_i}\| \le \frac{2}{1-\bar{t}_\phi}\sum_{i\in S^{k-1}} \|\widehat{\delta}^k_{\mathcal{J}_i}\|$.*

Proof: By the optimality of x^k and the feasibility of $x^{\rm LS}$ to the subproblem (12),

$$\frac{1}{2n}\|Ax^k - b\|^2 + \lambda^{k-1}\sum_{i=1}^m v_i^{k-1}\|x^k_{\mathcal{J}_i}\| \le \frac{1}{2n}\|Ax^{\rm LS} - b\|^2 + \lambda^{k-1}\sum_{i=1}^m v_i^{k-1}\|x^{\rm LS}_{\mathcal{J}_i}\|,$$

which, by the definition of $\varepsilon^{\rm LS}$, can be rearranged as follows:

$$\frac{1}{2n}\|A\widehat{\delta}^k\|^2 \le -\langle \varepsilon^{\rm LS}, \widehat{\delta}^k\rangle + \lambda^{k-1}\sum_{i=1}^m v_i^{k-1}(\|x^{\rm LS}_{\mathcal{J}_i}\| - \|x^k_{\mathcal{J}_i}\|).$$

Together with $x^{\rm LS}_{\mathcal{J}_i} = 0$ for all $i \notin \overline{S}$, it immediately follows that

$$\begin{aligned}
\frac{1}{2n}\|A\widehat{\delta}^k\|^2 &\le -\sum_{i=1}^m \langle \varepsilon^{\rm LS}_{\mathcal{J}_i}, \widehat{\delta}^k_{\mathcal{J}_i}\rangle + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}(\|x^{\rm LS}_{\mathcal{J}_i}\| - \|x^k_{\mathcal{J}_i}\|) - \lambda^{k-1}\sum_{i\notin\overline{S}} v_i^{k-1}\|x^k_{\mathcal{J}_i}\| \\
&\le \sum_{i\notin\overline{S}} \|\varepsilon^{\rm LS}_{\mathcal{J}_i}\|\|\widehat{\delta}^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\widehat{\delta}^k_{\mathcal{J}_i}\| - \lambda^{k-1}\sum_{i\notin S^{k-1}} v_i^{k-1}\|\widehat{\delta}^k_{\mathcal{J}_i}\| \\
&\le \sum_{i\notin\overline{S}} \|\varepsilon^{\rm LS}_{\mathcal{J}_i}\|\|\widehat{\delta}^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\widehat{\delta}^k_{\mathcal{J}_i}\| - \lambda^{k-1}(1-\bar{t}_\phi)\sum_{i\notin S^{k-1}} \|\widehat{\delta}^k_{\mathcal{J}_i}\|,
\end{aligned}$$

where the equality is due to $\varepsilon^{\rm LS}_{\mathcal{J}_{\overline{S}}} = 0$ implied by the optimality of $x^{\rm LS}$ to (17). Thus,

$$\begin{aligned}
&\frac{1}{2n}\|A\widehat{\delta}^k\|^2 + \big[\lambda^{k-1}(1-\bar{t}_\phi) - \|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty\big]\sum_{i\notin S^{k-1}}\|\widehat{\delta}^k_{\mathcal{J}_i}\| \\
&\le \sum_{i\in S^{k-1}\setminus\overline{S}} \|\varepsilon^{\rm LS}_{\mathcal{J}_i}\|\|\widehat{\delta}^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\widehat{\delta}^k_{\mathcal{J}_i}\| \le \lambda^{k-1}\sum_{i\in S^{k-1}}\|\widehat{\delta}^k_{\mathcal{J}_i}\|. \qquad (38)
\end{aligned}$$

Notice that $\lambda^{k-1} \ge \frac{2\|\mathcal{G}(\varepsilon^{LS})\|_\infty}{1-\bar{t}_\phi}$. The desired result directly follows from (38). □

If in addition the index set S^{k-1} in Lemma 10 is such that A satisfies the RSC over $\mathcal{C}(\overline{S}, |S^{k-1}|)$, then the conclusion of Lemma 10 can be strengthened as follows.

Lemma 11 *For $k \ge 1$, if there is an index set $S^{k-1} \supseteq \overline{S}$ such that $\min_{i\in(S^{k-1})^c} w_i^{k-1} \le \bar{t}_\phi$ and A satisfies the RSC over $\mathcal{C}(\overline{S}, |S^{k-1}|)$ with constant γ_k, then with $\lambda^{k-1} \ge \frac{2\|\mathcal{G}(\varepsilon^{LS})\|_\infty}{1-\bar{t}_\phi}$*

$$\|\widehat{\delta}^k\| \le \frac{1}{\gamma_k}\Big(\|\mathcal{G}(\varepsilon^{\rm LS})_{S^{k-1}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\Big).$$

Proof: By using the first inequality in (38) and $\widehat{\delta}^k \in \mathcal{C}(|S^{k-1}|)$ by Lemma 10, we have

$$\begin{aligned}\gamma_k\|\widehat{\delta}^k\|^2 \le \frac{1}{2n}\|A\widehat{\delta}^k\|^2 &\le \sum_{i\in S^{k-1}\backslash\overline{S}} \|\varepsilon^{\mathrm{LS}}_{\mathcal{J}_i}\|\|\widehat{\delta}^k_{\mathcal{J}_i}\| + \lambda^{k-1}\sum_{i\in\overline{S}} v_i^{k-1}\|\widehat{\delta}^k_{\mathcal{J}_i}\| \\ &\le \sqrt{\textstyle\sum_{i\in S^{k-1}\backslash\overline{S}}\|\varepsilon^{\mathrm{LS}}_{\mathcal{J}_i}\|^2}\,\|\widehat{\delta}^k\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2}\,\|\widehat{\delta}^k\|, \qquad (39)\end{aligned}$$

where the last inequality is using $\|\widehat{\delta}^k\|^2 = \sum_{i=1}^m \|\widehat{\delta}^k_{\mathcal{J}_i}\|^2$. Thus, it follows that

$$\begin{aligned}\gamma_k\|\widehat{\delta}^k\| &\le \sqrt{\textstyle\sum_{i\in S^{k-1}\backslash\overline{S}}\|\varepsilon^{\mathrm{LS}}_{\mathcal{J}_i}\|^2} + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2} \\ &= \|[\mathcal{G}(\varepsilon^{\mathrm{LS}})]_{S^{k-1}}\| + \lambda^{k-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{k-1})^2},\end{aligned}$$

which implies the desired result. The proof is then completed. □

The proof of Theorem 4.3: From $\nu \le \frac{1-\overline{t}_\phi}{2\|\mathcal{G}(\varepsilon^{\mathrm{LS}})\|_\infty}$ and $\frac{\rho}{\nu} \ge \frac{2\max(\|\mathcal{G}(\varepsilon^{\mathrm{LS}})\|_\infty, 2\kappa\|\mathcal{G}(\widehat{\varepsilon}^\dagger)\|_\infty)}{1-\overline{t}_\phi}$, $\lambda^{k-1} \ge \frac{2\|\mathcal{G}(\varepsilon^{LS})\|_\infty}{1-\overline{t}_\phi}$ for all $k\in\mathbb{N}$. We prove that the inequalities in (21) hold for $k=1$. Since $w^0 \le \overline{t}_\phi e$ and $S^0 = \overline{S}$, the conditions of Lemma 11 are satisfied for $k=1$. Then,

$$\|x^1 - x^{\mathrm{LS}}\| \le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\varepsilon^{\mathrm{LS}})]_{S^0}\| + \lambda^0\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^0)^2}\Big) \le \frac{1}{\kappa}\Big(\|[\mathcal{G}(\varepsilon^{\mathrm{LS}})]_{\overline{S}}\| + \frac{\sqrt{|F^0|}}{\nu}\Big) = \frac{\sqrt{|F^0|}}{\kappa\nu}.$$

where the equality is due to $[\mathcal{G}(\varepsilon^{\mathrm{LS}})]_{\overline{S}} = 0$. Since $\|x^{\mathrm{LS}}_{\mathcal{J}_i} - \overline{x}_{\mathcal{J}_i}\| \le \|\mathcal{G}(\widehat{\varepsilon}^\dagger)\|_\infty$ for $i = 1,\ldots,m$ by (18) and $\frac{\rho}{\nu} \ge \frac{4\kappa\|\mathcal{G}(\widehat{\varepsilon}^\dagger)\|_\infty}{1-\overline{t}_\phi}$, it follows that for all $i\in F^1$,

$$\|x^{\mathrm{LS}}_{\mathcal{J}_i} - x^1_{\mathcal{J}_i}\| \ge \|\overline{x}_{\mathcal{J}_i} - x^1_{\mathcal{J}_i}\| - \|\overline{x}_{\mathcal{J}_i} - x^{\mathrm{LS}}_{\mathcal{J}_i}\| \ge \frac{1}{(1-t^*_\phi)\rho} - \frac{\rho(1-\overline{t}_\phi)}{4\nu\kappa} \ge \frac{3\sqrt{5}}{(1+3\sqrt{5})(1-t^*_\phi)\rho},$$

where the last inequality is due to $\rho \le \sqrt{\frac{4\kappa\nu}{(1-t^*_\phi)(1+3\sqrt{5})}}$. From the last two inequalities,

$$\begin{aligned}\sqrt{|F^1|} &\le \frac{(1+3\sqrt{5})(1-t^*_\phi)\rho}{3\sqrt{5}}\sqrt{\textstyle\sum_{i=1}^m\|x^{\mathrm{LS}}_{\mathcal{J}_i} - x^1_{\mathcal{J}_i}\|^2} \\ &= \frac{\rho(1-t^*_\phi)(1+3\sqrt{5})\|x^{\mathrm{LS}} - x^1\|}{3\sqrt{5}} \le \frac{\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa}\sqrt{|F^0|}.\end{aligned}$$

Consequently, the conclusion holds for $k=1$. Now assuming that the conclusion holds for $k = l-1$ with $l\ge 2$, we shall prove that the conclusion holds for $k=l$. To this end, we first argue that $|S^{l-1}| \le 1.5\overline{r}$. Indeed, if $i\in S^{l-1}\backslash\overline{S}$, we have $i\notin\overline{S}$ and $w_i^{l-1}\in(\overline{t}_\phi, 1]$. From $w_i^{l-1}\in\partial\psi^*(\rho\|x^{l-1}_{\mathcal{J}_i}\|)$, we have $\rho\|x^{l-1}_{\mathcal{J}_i}\| \ge \phi'_-(w_i^{l-1}) \ge \phi'_+(\overline{t}_\phi) \ge \frac{1}{1-t^*_\phi}$. Thus,

$$\begin{aligned}\sqrt{|S^{l-1}\backslash\overline{S}|} \le \sqrt{|F^{l-1}|} &\le \frac{\max(1,\rho)\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa}\sqrt{|F^{l-2}|} \le \cdots \\ &\le \Big(\frac{\max(1,\rho)\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa}\Big)^{l-1}\sqrt{|F^0|} \le \sqrt{(4/6)^{2l-2}|F^0|} \le \sqrt{0.5\overline{r}},\end{aligned} \qquad (40)$$

where the first inequality is due to $S^{l-1}\backslash\overline{S} \subseteq F^{l-1}$ and $l \geq 2$. This implies $|S^{l-1}| \leq 1.5\overline{r}$, and hence the conditions of Lemma 11 in Appendix C are satisfied. Consequently,

$$\begin{aligned}
\|x^l - x^{\rm LS}\| &\leq \frac{1}{\kappa}\Big(\|[\mathcal{G}(\varepsilon^{\rm LS})]_{S^{l-1}}\| + \lambda^{l-1}\sqrt{\textstyle\sum_{i\in\overline{S}}(v_i^{l-1})^2}\Big)\\
&\leq \frac{1}{\kappa}\Big(\|\mathcal{G}(\varepsilon^{\rm LS})_{S^{l-1}\backslash\overline{S}}\| + \lambda^{l-1}\sqrt{\textstyle\sum_{i\in\overline{S}}\mathbb{I}_{F^{l-1}}(i)}\Big)\\
&\leq \frac{1}{\kappa}\Big(\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty\sqrt{|S^{l-1}\backslash\overline{S}|} + \lambda^{l-1}\sqrt{|F^{l-1}\cap\overline{S}|}\Big)\\
&\leq \frac{\lambda^{l-1}}{\kappa}\Big(\frac{1}{2}\sqrt{|F^{l-1}\backslash\overline{S}|} + \sqrt{|F^{l-1}\cap\overline{S}|}\Big) \leq \frac{\rho}{\nu\kappa}\sqrt{1.25|F^{l-1}|} \leq \frac{\sqrt{5}\rho}{2\nu\kappa}\sqrt{|F^{l-1}|},
\end{aligned}$$

where the second inequality is using $\varepsilon^{\rm LS}_{\mathcal{J}_{\overline{S}}} = 0$, $\rho > \frac{2\phi'_-(1)}{\min_{i\in\overline{S}}\|\overline{x}_{\mathcal{J}_i}\|}$ and Lemma 9, the fourth one is due to $\lambda^{l-1} \geq 2\|\mathcal{G}(\varepsilon^{\rm LS})\|_\infty$, and the fifth one is since $\frac{1}{2}a + b \leq \sqrt{1.25(a^2+b^2)}$ for all $a, b \in \mathbb{R}$. In addition, by using the same argument as those for $k = 1$, for all $i \in F^l$ we have $\|x^l_{\mathcal{J}_i} - x^{\rm LS}_{\mathcal{J}_i}\| \geq \frac{3\sqrt{5}}{1+3\sqrt{5}}\frac{1}{(1-t^*_\phi)\rho}$, and hence $\sqrt{|F^l|} \leq \frac{\max(1,\rho)\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa}\sqrt{|F^{l-1}|}$. Thus, we complete the proof of the case $k = l$, and the inequalities in (21) hold.

Since $\max(1,\rho)\rho\nu^{-1}(1-t^*_\phi)(1+3\sqrt{5}) \leq 4\kappa$, we have $\frac{\max(1,\rho)\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa} \leq \frac{2}{3}$. Together with (40),

$$\sqrt{|F^{\overline{k}}|} \leq \Big(\frac{\max(1,\rho)\rho(1-t^*_\phi)(1+3\sqrt{5})}{6\nu\kappa}\Big)^{\overline{k}-1}\sqrt{|F^0|} < 1,$$

which implies that $|F^k| = 0$ when $k \geq \overline{k}$. Together with the first inequality in (21), we have $x^k = x^{\rm LS}$ when $k \geq \overline{k}$. From (18) and $\rho \leq \sqrt{\frac{4\kappa\nu}{(1-t^*_\phi)(1+3\sqrt{5})}}$, for all $i \in \overline{S}$,

$$\big|\|\overline{x}_{\mathcal{J}_i}\| - \|x^{LS}_{\mathcal{J}_i}\|\big| \leq \|\overline{x}_{\mathcal{J}_i} - x^{LS}_{\mathcal{J}_i}\| \leq \|\mathcal{G}(\widehat{\varepsilon}^\dagger)\|_\infty \leq \frac{\rho(1-\overline{t}_\phi)}{4\nu\kappa} \leq \frac{(1-\overline{t}_\phi)}{(1-t^*_\phi)(1+3\sqrt{5})\rho}.$$

This, together with $\min_{i\in\overline{S}}\|\overline{x}_{\mathcal{J}_i}\| \geq \frac{2\phi'_-(1)}{\rho} \geq \frac{2}{\rho(1-t^*_\phi)}$, implies that $\|x^{\rm LS}_{\mathcal{J}_i}\| > 0$ for all $i \in \overline{S}$. Thus, $\mathrm{supp}(\mathcal{G}(x^{\rm LS})) = \overline{S}$, and consequently $\mathrm{supp}(\mathcal{G}(x^k)) = \overline{S}$ for all $k \geq \overline{k}$. □