*The Annals of Statistics*
2014, Vol. 42, No. 3, 819–849
DOI: 10.1214/13-AOS1198


# STRONG ORACLE OPTIMALITY OF FOLDED CONCAVE PENALIZED ESTIMATION

By Jianqing Fan[1], Lingzhou Xue[2] and Hui Zou[3]

*Princeton University, Pennsylvania State University and University of Minnesota*

Folded concave penalization methods have been shown to enjoy the strong oracle property for high-dimensional sparse estimation. However, a folded concave penalization problem usually has multiple local solutions and the oracle property is established only for one of the unknown local solutions. A challenging fundamental issue still remains that it is not clear whether the local optimum computed by a given optimization algorithm possesses those nice theoretical properties. To close this important theoretical gap in over a decade, we provide a unified theory to show explicitly how to obtain the oracle solution via the local linear approximation algorithm. For a folded concave penalized estimation problem, we show that as long as the problem is localizable and the oracle estimator is well behaved, we can obtain the oracle estimator by using the one-step local linear approximation. In addition, once the oracle estimator is obtained, the local linear approximation algorithm converges, namely it produces the same estimator in the next iteration. The general theory is demonstrated by using four classical sparse estimation problems, that is, sparse linear regression, sparse logistic regression, sparse precision matrix estimation and sparse quantile regression.

**1. Introduction.** Sparse estimation is at the center of the stage of high-dimensional statistical learning. The two mainstream methods are the LASSO (or $\ell_1$ penalization) and the folded concave penalization [Fan and Li (2001)] such as the SCAD and the MCP. Numerous papers have been devoted to the

Received October 2012; revised December 2013.
[1]Supported by NIH Grant R01-GM072611 and NSF Grants DMS-12-06464 and 0704337.
[2]Supported by NIH Grant R01-GM100474 as a postdoctor at Princeton University.
[3]Supported by NSF Grant DMS-08-46068 and a grant from Office of Naval Research.
*AMS 2000 subject classifications.* Primary 62J07.
*Key words and phrases.* Folded concave penalty, local linear approximation, nonconvex optimization, oracle estimator, sparse estimation, strong oracle property.

numerical and theoretical study of both methods. A strong irrepresentable condition is necessary for the LASSO to be selection consistent [Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006)]. The folded concave penalization, unlike the LASSO, does not require the irrepresentable condition to achieve the variable selection consistency and can correct the intrinsic estimation bias of the LASSO [Fan and Li (2001), Fan and Peng (2004), Zhang (2010a), Fan and Lv (2011)]. The LASSO owns its popularity largely to its computational properties. For certain learning problems, such as the LASSO penalized least squares, the solution paths are piecewise linear which allows one to employ a LARS-type algorithm to compute the entire solution path efficiently [Efron et al. (2004)]. For a more general class of $\ell_1$ penalization problems, the coordinate descent algorithm has been shown to be very useful and efficient [Friedman, Hastie and Tibshirani (2008, 2010)].

The computation for folded concave penalized methods is much more involved, because the resulting optimization problem is usually nonconvex and has multiple local minimizers. Several algorithms have been developed for computing the folded concave penalized estimators. Fan and Li (2001) worked out the local quadratic approximation (LQA) algorithm as a unified method for computing the folded concave penalized maximum likelihood. Zou and Li (2008) proposed the local linear approximation (LLA) algorithm which turns a concave penalized problem into a series of reweighed $\ell_1$ penalization problems. Both LQA and LLA are related to the MM principle [Hunter and Lange (2004), Hunter and Li (2005)]. Recently, coordinate descent was applied to solve the folded concave penalized least squares [Mazumder, Friedman and Hastie (2011), Fan and Lv (2011)]. Zhang (2010a) devised a PLUS algorithm for solving the penalized least squares using the MCP and proved the oracle property. Zhang (2010b, 2013) analyzed the capped-$\ell_1$ penalty for solving the penalized least squares and proved the oracle property as well. With these advances in computing algorithms, one can now at least efficiently compute a local solution of the folded concave penalized problem. It has been shown repeatedly that the folded concave penalty performs better than the LASSO in various high-dimensional sparse estimation problems. Examples include sparse linear regression [Fan and Li (2001), Zhang (2010a)], sparse generalized linear model [Fan and Lv (2011)], sparse Cox's proportional hazards model [Bradic, Fan and Jiang (2011)], sparse precision matrix estimation [Lam and Fan (2009)], sparse Ising model [Xue, Zou and Cai (2012)], and sparse quantile regression [Wang, Wu and Li (2012), Fan, Fan and Barut (2014)], among others.

Before declaring that the folded concave penalty is superior to the LASSO, we need to resolve a missing puzzle in the picture. Theoretical properties of the folded concave penalization are established for a theoretic local solution. However, we have to employ one of these local minimization algorithms to find such a local optimal solution. It remains to prove that the computed

local solution is the desired theoretic local solution to make the theory fully relevant. Many have tried to address this issue [Zhang (2010a), Fan and Lv (2011), Zhang and Zhang (2012)]. The basic idea there is to find conditions under which the folded concave penalization actually has a unique sparse local minimizer, and hence eliminate the problem of multiple local solutions. Although this line of thoughts is natural and logically intuitive, the imposed conditions for the unique sparse local minimizer are too strong to be realistic.

In this paper, we offer a different approach to directly deal with the multiple local solutions issue. We outline a general procedure based on the LLA algorithm for solving a specific local solution of the folded concave penalization, and then derive a lower bound on the probability that this specific local solution exactly equals the oracle estimator. This probability lower bound equals $1-\delta_0-\delta_1-\delta_2$, where $\delta_0$ corresponds to the exception probability of the localizability of the underlying model, $\delta_1$ and $\delta_2$ represent the exception probability of the regularity of the oracle estimator and they have nothing to do with any actual estimation method. Explicit expressions of $\delta_0$, $\delta_1$ and $\delta_2$ are given in Section 2. Under weak regularity conditions, $\delta_1$ and $\delta_2$ are very small. Thus, if $\delta_0$ goes to zero then the computed solution is the oracle estimator with an overwhelming probability. On the other hand, if $\delta_0$ cannot go to zero, then it means that the model is extremely difficult to estimate no matter how clever an estimator is. Thus, our theory suggests a "bet-on-folded-concave-penalization" principle: as long as there is a reasonable initial estimator, our procedure can deliver an optimal estimator using the folded concave penalization via the one-step LLA implementation. Once the oracle estimator is obtained, the LLA algorithm converges in the next iteration, namely, the oracle estimator is a fixed point. Furthermore, we use four concrete examples to show that exception probabilities $\delta_0$, $\delta_1$ and $\delta_2$ go to zero at a fast rate under the ultra-high-dimensional setting where $\log(p)=O(n^\eta)$ for $\eta\in(0,1)$.

Throughout this paper, the following notation is used. For $\mathbf{U}=(u_{ij})_{k\times l}$, let $\|\mathbf{U}\|_{\min}=\min_{(i,j)}|u_{ij}|$ be its minimum absolute value, and let $\lambda_{\min}(\mathbf{U})$ and $\lambda_{\max}(\mathbf{U})$ be its smallest and largest eigenvalues. We introduce several matrix norms: the $\ell_1$ norm $\|\mathbf{U}\|_{\ell_1}=\max_j\sum_i|u_{ij}|$, the $\ell_2$ norm $\|\mathbf{U}\|_{\ell_2}=\lambda_{\max}^{1/2}(\mathbf{U}'\mathbf{U})$, the $\ell_\infty$ norm $\|\mathbf{U}\|_{\ell_\infty}=\max_i\sum_j|u_{ij}|$, the entrywise $\ell_1$ norm $\|\mathbf{U}\|_1=\sum_{(i,j)}|u_{ij}|$ and the entrywise $\ell_\infty$ norm $\|\mathbf{U}\|_{\max}=\max_{(i,j)}|u_{ij}|$.

**2. Main results.** We begin with an abstract presentation of the sparse estimation problem. Consider estimating a model based on $n$ *i.i.d.* observations. The target of estimation is "parameter" $\boldsymbol{\beta}^\star=(\beta_1^\star,\ldots,\beta_p^\star)'$, that is, the model is parameterized by $\boldsymbol{\beta}^\star$. The dimension $p$ is larger than the sample size $n$. In some problems, the target $\boldsymbol{\beta}^\star$ can be a matrix (e.g., an inverse covariance matrix). In such cases, it is understood that $(\beta_1^\star,\ldots,\beta_p^\star)'$ is the

vectorization of the matrix $\boldsymbol{\beta}^\star$. Denote the support set as $\mathcal{A} = \{j : \beta_j^\star \neq 0\}$ and its cardinality is $s = |\mathcal{A}|$. The sparsity assumption means that $s \ll p$.

Suppose that our estimation scheme is to get a local minimizer of the following folded concave penalized estimation problem:

$$\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + P_\lambda(|\boldsymbol{\beta}|) \tag{1}$$

with $\ell_n(\boldsymbol{\beta})$ is a convex loss and $P_\lambda(|\boldsymbol{\beta}|) = \sum_j P_\lambda(|\beta_j|)$ is a folded concave penalty. In our general theory, $\ell_n(\boldsymbol{\beta})$ in (1) does not need to be differentiable. The above formulation is a bit abstract but covers many important statistical models. For example, $\ell_n(\boldsymbol{\beta})$ can be the squared error loss in penalized least squares, the check loss in penalized quantile regression and the negative log-quasi-likelihood function in penalized maximum quasi-likelihood.

An oracle knows the true support set, and defines the oracle estimator as

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}) = \arg \min_{\boldsymbol{\beta} : \boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}} \ell_n(\boldsymbol{\beta}). \tag{2}$$

We assume that (2) is regular such that the oracle solution is unique, namely,

$$\nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = 0 \qquad \forall j \in \mathcal{A}, \tag{3}$$

where $\nabla_j$ denotes the subgradient with respect to the $j$th element of $\boldsymbol{\beta}$. If the convex loss is differentiable, the subgradient is the usual gradient. The oracle estimator is not a feasible estimator but it can be used as a theoretic benchmark for other estimators to compare with. An estimator is said to have the oracle property if it has the same asymptotic distribution as the oracle estimator [Fan and Li (2001), Fan and Peng (2004)]. Moreover, an estimator is said to have the strong oracle property if the estimator equals the oracle estimator with overwhelming probability [Fan and Lv (2011)].

Throughout this paper, we also assume that the penalty $P_\lambda(|t|)$ is a general folded concave penalty function defined on $t \in (-\infty, \infty)$ satisfying:

(i) $P_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$ with $P_\lambda(0) = 0$;
(ii) $P_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $P'_\lambda(0) := P'_\lambda(0+) \geq a_1\lambda$;
(iii) $P'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;
(iv) $P'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with the pre-specified constant $a > a_2$.

Where $a_1$ and $a_2$ are two fixed positive constants. The above definition follows and extends previous works on the SCAD and the MCP [Fan and Li (2001), Zhang (2010a), Fan and Lv (2011)]. The derivative of the SCAD penalty is

$$P'_\lambda(t) = \lambda I_{\{t \leq \lambda\}} + \frac{(a\lambda - t)_+}{a - 1} I_{\{t > \lambda\}} \qquad \text{for some } a > 2$$

and the derivative of the MCP is $P'_\lambda(t) = (\lambda - \frac{t}{a})_+$, for some $a > 1$. It is easy to see that $a_1 = a_2 = 1$ for the SCAD, and $a_1 = 1 - a^{-1}$, $a_2 = 1$ for the MCP.

The hard-thresholding penalty $P_\lambda(t) = \lambda^2 - (t-\lambda)^2 I_{\{t<\lambda\}}$ [Antoniadis and Fan (2001)] is another special case of the general folded concave penalty with $a = a_1 = 1$, $a_2 = \frac{1}{2}$.

Numerical results in the literature show that the folded concave penalty outperforms the $\ell_1$ penalty in terms of estimation accuracy and selection consistency. To provide understanding of their differences, it is important to show that the obtained solution of the folded concave penalization has better theoretical properties than the $\ell_1$-penalization. The technical difficulty here is to show that the computed local solution is the local solution with proven properties. Zhang (2010a) and Fan and Lv (2011) proved the restricted global optimality that the oracle estimator is the unique global solution in the subspace $\mathbb{S}_s$, which is the union of all $s$-dimensional coordinate subspaces in $\mathbb{R}^p$. Under strong conditions, Zhang and Zhang (2012) proved that the global solution leads to desirable recovery performance and corresponds to the unique sparse local solution, and hence any algorithm finding a sparse local solution will find the desired global solution. The fundamental problem with these arguments is that in reality it is very rare that the concave regularization actually has a unique sparse local solution, which in turn implies that these strong conditions are too stringent to hold in practice. Evidence is given in the simulation studies in Section 4 where we show that the concave penalization has multiple sparse local solutions.

We argue that, although the estimator is defined via the folded concave penalization, we only care about properties of the computed estimator. It is perfectly fine that the computed local solution is not the global minimizer, as long as it has the desired properties. In this paper, we directly analyze a specific estimator by the local linear approximation (LLA) algorithm [Zou and Li (2008)]. The LLA algorithm takes advantage of the special folded concave structure and utilizes the majorization–minimization (MM) principle to turn a concave regularization problem into a sequence of weighted $\ell_1$ penalized problems. Within each LLA iteration, the local linear approximation is the best convex majorization of the concave penalty function [see Theorem 2 of Zou and Li (2008)]. Moreover, the MM principle has provided theoretical guarantee to the convergence of the LLA algorithm to a stationary point of the folded concave penalization. By analyzing a logarithmic number of the LLA iterations, Zhang (2010b) and Huang and Zhang (2012) proved that the LLA algorithm helps the folded concave penalization reduce the estimation error of the LASSO in sparse linear and generalized linear regression. In contrast, thresholding Lasso will not have the oracle property if the irresponsentable condition does not hold: once some important variables are missed in the Lasso fit, they can not be rescued by thresholding. This is a very different operation from the LLA algorithm.

Here, we summarize the details of the LLA algorithm as in Algorithm 1.

**Algorithm 1** The local linear approximation (LLA) algorithm

1. Initialize $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{\text{initial}}$ and compute the adaptive weight
$$\hat{\mathbf{w}}^{(0)} = (\hat{w}_1^{(0)}, \ldots, \hat{w}_p^{(0)})' = (P'_\lambda(|\hat{\beta}_1^{(0)}|), \ldots, P'_\lambda(|\hat{\beta}_p^{(0)}|))'.$$
2. For $m = 1, 2, \ldots,$ repeat the LLA iteration till convergence
   (2.a) Obtain $\hat{\boldsymbol{\beta}}^{(m)}$ by solving the following optimization problem
$$\hat{\boldsymbol{\beta}}^{(m)} = \min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + \sum_j \hat{w}_j^{(m-1)} \cdot |\beta_j|,$$
   (2.b) Update the adaptive weight vector $\hat{\mathbf{w}}^{(m)}$ with $\hat{w}_j^{(m)} = P'_\lambda(|\hat{\beta}_j^{(m)}|)$.

In the following theorems, we provide the nonasymptotic analysis of the LLA algorithm for obtaining the oracle estimator in the folded concave penalized problem if it is initialized by some appropriate initial estimator. In particular, the convex loss $\ell_n(\boldsymbol{\beta})$ is not required to be differentiable. To simplify notation, define $\nabla \ell_n(\boldsymbol{\beta}) = (\nabla_1 \ell_n(\boldsymbol{\beta}), \ldots, \nabla_p \ell_n(\boldsymbol{\beta}))$ as the subgradient vector of $\ell_n(\boldsymbol{\beta})$. Denote by $\mathcal{A}^c = \{j : \beta_j^\star = 0\}$ the complement of the true support set $\mathcal{A}$, and set $\nabla_{\mathcal{A}^c} \ell_n(\boldsymbol{\beta}) = (\nabla_j \ell_n(\boldsymbol{\beta}) : j \in \mathcal{A}^c)$ with respect to $\mathcal{A}^c$.

THEOREM 1. *Suppose the minimal signal strength of $\boldsymbol{\beta}^\star$ satisfies that*

(A0) $\|\boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\min} > (a+1)\lambda$.

*Consider the folded concave penalized problem with $P_\lambda(\cdot)$ satisfying* (i)–(iv). *Let $a_0 = \min\{1, a_2\}$. Under the event*
$$\mathcal{E}_1 = \{\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\|_{\max} \le a_0\lambda\} \cap \{\|\nabla_{\mathcal{A}^c} \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} < a_1\lambda\},$$
*the LLA algorithm initialized by $\hat{\boldsymbol{\beta}}^{\text{initial}}$ finds $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ after one iteration.*

Applying the union bound to $\mathcal{E}_1$, we easily get the following corollary.

COROLLARY 1. *With probability at least $1 - \delta_0 - \delta_1$, the LLA algorithm initialized by $\hat{\boldsymbol{\beta}}^{\text{initial}}$ finds $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ after one iteration, where*
$$\delta_0 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\|_{\max} > a_0\lambda)$$
*and*
$$\delta_1 = \Pr(\|\nabla_{\mathcal{A}^c} \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} \ge a_1\lambda).$$

REMARK 1. By its definition, $\delta_0$ represents the localizability of the underlying model. To apply Theorem 1, we need to have an appropriate initial

estimator to make $\delta_0$ go to zero as $n$ and $p$ diverge to infinity, namely the underlying problem is localizable. In Section 3, we will show by concrete examples on how to find a good initial estimator to make the problem localizable. $\delta_1$ represents the regularity behavior of the oracle estimator, that is, its closeness to the true "parameter" measured by the score function. Note that $\nabla_{\mathcal{A}^c}\ell_n(\boldsymbol{\beta}^\star)$ is concentrated around zero. Thus, $\delta_1$ is usually small. In summary, Theorem 1 and its corollary state that as long as the problem is localizable and regular, we can find an oracle estimator by using the one-step local linear approximation, which is a generalization of the one-step estimation idea [Zou and Li (2008)] to the high-dimensional setting.

THEOREM 2. *Consider the folded concave penalized problem (1) with* $P_\lambda(\cdot)$ *satisfying* (i)–(iv). *Under the event*

$$\mathcal{E}_2 = \{\|\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} < a_1\lambda\} \cap \{\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} > a\lambda\},$$

*if* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *is obtained, the LLA algorithm will find* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *again in the next iteration, that is, it converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *in the next iteration and is a fixed point.*

Now we combine Theorems 1 and 2 to derive the nonasymptotic probability bound for the LLA algorithm to exactly converge to the oracle estimator.

COROLLARY 2. *Consider the folded concave penalized problem (1) with* $P_\lambda(\cdot)$ *satisfying* (i)–(iv). *Under assumption* (A0), *the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{initial}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \delta_0 - \delta_1 - \delta_2$, *where*

$$\delta_2 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \le a\lambda).$$

REMARK 2. The localizable probability $1-\delta_0$ and regularity probability $1-\delta_1$ have been defined before. $\delta_2$ is a probability on the magnitude of the oracle estimator. Both $\delta_1$ and $\delta_2$ are related to the regularity behavior of the oracle estimator and will be referred to the oracle regularity condition. Under assumption (A0), it requires only the uniform convergence of $\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}$. Namely,

$$\delta_2 \le \Pr(\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\max} > \lambda).$$

Thus, we can regard $\delta_2$ as a direct measurement of the closeness of the oracle estimator to the true "parameter" and is usually small because of the small intrinsic dimensionality $s$. This will indeed be shown in Section 3.

REMARK 3. The philosophy of our work follows the well-known one-step estimation argument [Bickel (1975)] in the maximum likelihood estimation. In some likelihood models, the log-likelihood function is not concave. One of the local maximizers of the log-likelihood is shown to be asymptotic efficient, but how to compute that estimator is very challenging. Bickel (1975) overcame this difficulty by focusing on a specially designed one-step Newton–Raphson estimator initialized by a root-$n$ estimator. This one-step estimator is asymptotically efficient, just like the theoretical MLE. Note that Bickel's theory did not try to get the global maximizer nor the theoretical local maximizer of the log-likelihood, although the log-likelihood was used to construct the explicit estimator. Our general theory follows this line of thinking. Theorems 1–2 show how to construct the explicit estimator that possesses the desired strong oracle property. This is all we need to close the theoretical gap. Following Bickel (1975), we can just use the two-step LLA solution and do not need to care about whether the LLA algorithm converges. Of course, Theorem 2 does offer a statistical convergence proof of the LLA algorithm, which differs from its numeric convergence argument. Moreover, Theorem 2 also proves the statistical equivalence between the two-step and fully converged LLA solutions. Thus, we recommend using the two-step LLA solution as the folded concave penalized estimator in applications.

**3. Theoretical examples.** This section outlines four classical examples to demonstrate interesting and powerful applications of Theorems 1 and 2. We consider the linear regression, logistic regression, precision matrix estimation and quantile regression. We basically need to check the localizable condition and the regularity condition for these problems.

3.1. *Sparse linear regression.* The first example is the canonical problem of the folded concave penalized least square estimation, that is,

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \sum_j P_\lambda(|\beta_j|), \tag{4}$$

where $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)' \in \mathbb{R}^{n\times p}$. Let $\boldsymbol{\beta}^\star$ be the true parameter vector in the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \varepsilon$, and the true support set of $\boldsymbol{\beta}^\star = (\beta_j^\star)_{1\le j\le p}$ is $\mathcal{A} = \{j : \beta_j^\star \neq 0\}$. For the folded concave penalized least square problem, the oracle estimator has an explicit form of

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}) \qquad \text{with } \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}} = (\mathbf{X}_{\mathcal{A}}'\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}'\mathbf{y}$$

and the Hessian matrix of $\ell_n(\boldsymbol{\beta})$ is $n^{-1}\mathbf{X}'\mathbf{X}$ regardless of $\boldsymbol{\beta}$. Applying Theorems 1–2, we can derive the following theorem with explicit upper bounds for $\delta_1$ and $\delta_2$, which depends only on the behavior of the oracle estimator.

THEOREM 3. *Recall that* $\delta_0 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\|_{\max} > a_0\lambda)$. *Suppose*

(A1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \varepsilon$ *with* $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ *being i.i.d. sub-Gaussian* $(\sigma)$ *for some fixed constant* $\sigma > 0$, *that is,* $E[\exp(t\varepsilon_i^2)] \le \exp(\sigma^2 t^2/2)$.

*The LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{initial}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \delta_0 - \delta_1^{\text{linear}} - \delta_2^{\text{linear}}$, *where*

$$\delta_1^{\text{linear}} = 2(p-s) \cdot \exp\left( -\frac{a_1^2 n \lambda^2}{2M\sigma^2} \right)$$

*and*

$$\delta_2^{\text{linear}} = 2s \cdot \exp\left( -\frac{n\lambda_{\min}}{2\sigma^2} \cdot (\|\boldsymbol{\beta}_{\mathcal{A}}^\star\|_{\min} - a\lambda)^2 \right),$$

*where* $\lambda_{\min} = \lambda_{\min}(\frac{1}{n}\mathbf{X}_{\mathcal{A}}'\mathbf{X}_{\mathcal{A}})$ *and* $M = \max_j \frac{1}{n}\|\mathbf{x}_{(j)}\|_{\ell_2}^2$, *which is usually 1 due to normalization, with* $\mathbf{x}_{(j)} = (x_{1j}, \ldots, x_{nj})'$.

By Theorem 3, $\delta_1^{\text{linear}}$ and $\delta_2^{\text{linear}}$ go to zero very quickly. Then it remains to bound $\delta_0$. To analyze $\delta_0$, we should decide the initial estimator. Here, we use the LASSO [Tibshirani (1996)] to initialize the LLA algorithm, which is

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda_{\text{lasso}}\|\boldsymbol{\beta}\|_{\ell_1}. \tag{5}$$

Note that $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ is the one-step LLA solution initialized by zero. In order to bound $\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star$, we invoke the following restricted eigenvalue condition:

(C1) $\kappa_{\text{linear}} = \min_{\mathbf{u}\neq 0:\, \|\mathbf{u}_{\mathcal{A}^c}\|_{\ell_1} \le 3\|\mathbf{u}_{\mathcal{A}}\|_{\ell_1}} \frac{\|\mathbf{X}\mathbf{u}\|_{\ell_2}^2}{n\|\mathbf{u}\|_{\ell_2}^2} \in (0, \infty)$.

This condition was studied in Bickel, Ritov and Tsybakov (2009), van de Geer and Bühlmann (2009) and Negahban et al. (2012). Under the assumptions of (A1) and (C1), the LASSO yields the unique optimum $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ satisfying

$$\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\ell_2} \le \frac{4s^{1/2}\lambda_{\text{lasso}}}{\kappa_{\text{linear}}}$$

with probability at least $1 - 2p\exp(-\frac{n\lambda_{\text{lasso}}^2}{2M\sigma^2})$. Thus, using this as an upper bound for $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\max}$, it is easy for us to obtain the following corollary.

COROLLARY 3. *Under assumptions* (A0), (A1) *and* (C1), *if we pick* $\lambda \ge \frac{4s^{1/2}\lambda_{\text{lasso}}}{a_0\kappa_{\text{linear}}}$, *the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - 2p\exp(-\frac{n\lambda_{\text{lasso}}^2}{2M\sigma^2}) - \delta_1^{\text{linear}} - \delta_2^{\text{linear}}$.

REMARK 4. Corollary 3 also suggests that sometimes it is good to use zero to initialize the LLA algorithm. If $\hat{\boldsymbol{\beta}}^{\text{initial}} = \mathbf{0}$, the first LLA iteration

gives a LASSO estimator with $\lambda_{\text{lasso}} = P'_\lambda(0)$. For both SCAD and MCP, $P'_\lambda(0) = \lambda$. If $\lambda_{\text{lasso}} = \lambda$ satisfies requirements in Corollary 3, then after two more LLA iterations, the LLA algorithm converges to the oracle estimator with high probability. To be more specific, we have the following corollary.

COROLLARY 4. *Consider the SCAD or MCP penalized linear regression. Under assumptions* (A0), (A1) *and* (C1), *if* $a_0\kappa_{\text{linear}} \geq 4s^{1/2}$, *the LLA algorithm initialized by zero converges to the oracle estimator after three iterations with probability at least* $1 - 2p \cdot \exp(-\frac{n\lambda^2}{2M\sigma^2}) - \delta_1^{\text{linear}} - \delta_2^{\text{linear}}$.

REMARK 5. The $s^{1/2}$ factor appears in Corollaries 3–4 because $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\ell_2}$ is used to bound $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\max}$. It is possible to get rid of the $s^{1/2}$ factor by using the $\ell_\infty$ loss of the LASSO in Zhang (2009) and Ye and Zhang (2010). Ye and Zhang (2010) introduced the cone invertability factor condition, that is,

(C1′) $\zeta_{\text{linear}} = \min_{\mathbf{u}\neq 0:\, \|\mathbf{u}_{\mathcal{A}^c}\|_{\ell_1} \leq 3\|\mathbf{u}_{\mathcal{A}}\|_{\ell_1}} \frac{\|\mathbf{X}'\mathbf{X}\mathbf{u}\|_{\max}}{n\|\mathbf{u}\|_{\max}} \in (0, \infty)$.

Under assumptions (A1) and (C1′), the LASSO yields $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ satisfying $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\max} \leq \frac{3\lambda_{\text{lasso}}}{2\zeta_{\text{linear}}}$ with probability at least $1 - 2p\exp(-\frac{n\lambda_{\text{lasso}}^2}{8M\sigma^2})$.

REMARK 6. Although we have considered using the LASSO as the initial estimator, we can also use the Dantzig selector [Candes and Tao (2007)] as the initial estimator, and the same analysis can still go through under a similar restricted eigenvalue condition as in Bickel, Ritov and Tsybakov (2009) or a similar cone invertability factor condition as in Ye and Zhang (2010).

3.2. *Sparse logistic regression.* The second example is the folded concave penalized logistic regression. Assume that

(A2) the conditional distribution of $y_i$ given $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ is a Bernoulli distribution with $\Pr(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}^\star) = \exp(\mathbf{x}_i'\boldsymbol{\beta}^\star)/(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta}^\star))$.

Then the penalized logistic regression is given by

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_i \{-y_i\mathbf{x}_i'\boldsymbol{\beta} + \psi(\mathbf{x}_i'\boldsymbol{\beta})\} + \sum_j P_\lambda(|\beta_j|) \tag{6}$$

with the canonical link $\psi(t) = \log(1 + \exp(t))$. This is a canonical model for high-dimensional classification problems, and it is a classical example of generalized linear models. The oracle estimator is given by

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}) = \arg\min_{\boldsymbol{\beta}:\, \boldsymbol{\beta}_{\mathcal{A}^c}=\mathbf{0}} \frac{1}{n}\sum_i \{-y_i\mathbf{x}_i'\boldsymbol{\beta} + \psi(\mathbf{x}_i'\boldsymbol{\beta})\}.$$

For ease of presentation, we define $\boldsymbol{\mu}(\boldsymbol{\beta}) = (\psi'(\mathbf{x}_1'\boldsymbol{\beta}), \ldots, \psi'(\mathbf{x}_n'\boldsymbol{\beta}))'$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \operatorname{diag}\{\psi''(\mathbf{x}_1'\boldsymbol{\beta}), \ldots, \psi''(\mathbf{x}_n'\boldsymbol{\beta})\}$. We also define three useful quantities: $Q_1 = \max_j \lambda_{\max}(\frac{1}{n}\mathbf{X}_{\mathcal{A}}' \operatorname{diag}\{|\mathbf{x}_{(j)}|\}\mathbf{X}_{\mathcal{A}})$, $Q_2 = \|(\frac{1}{n}\mathbf{X}_{\mathcal{A}}'\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}})^{-1}\|_{\ell_\infty}$, and $Q_3 = \|\mathbf{X}_{\mathcal{A}^c}'\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}'\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}})^{-1}\|_{\ell_\infty}$, where $\operatorname{diag}\{|\mathbf{x}_{(j)}|\}$ is a diagonal matrix with elements $\{|x_{ij}|\}_{i=1}^n$.

THEOREM 4. *Recall that* $\delta_0 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\|_{\max} > a_0\lambda)$. *Under assumption* (A2), *the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{initial}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \delta_0 - \delta_1^{\text{logit}} - \delta_2^{\text{logit}}$, *where*

$$\delta_1^{\text{logit}} = 2s \cdot \exp\left(-\frac{n}{M}\min\left\{\frac{2}{Q_1^2 Q_2^4 s^2}, \frac{a_1^2\lambda^2}{2(1+2Q_3)^2}\right\}\right) + 2(p-s)\cdot\exp\left(-\frac{a_1^2 n\lambda^2}{2M}\right)$$

*with* $M = \max_j n^{-1}\|\mathbf{x}_{(j)}\|_{\ell_2}^2$ *and*

$$\delta_2^{\text{logit}} = 2s \cdot \exp\left(-\frac{n}{MQ_2^2}\min\left\{\frac{2}{Q_1^2 Q_2^2 s^2}, \frac{1}{2}(\|\boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\min} - a\lambda)^2\right\}\right).$$

Under fairly weak assumptions, $\delta_1^{\text{logit}}$ and $\delta_2^{\text{logit}}$ go to zero very quickly. The remaining challenge is to bound $\delta_0$. We consider using the $\ell_1$-penalized maximum likelihood estimator as the initial estimator, that is,

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_i \{-y_i\mathbf{x}_i'\boldsymbol{\beta} + \psi(\mathbf{x}_i'\boldsymbol{\beta})\} + \lambda_{\text{lasso}}\|\boldsymbol{\beta}\|_{\ell_1}.$$

THEOREM 5. *Let* $m = \max_{(i,j)}|x_{ij}|$. *Under assumption* (A2) *and*

(C2) $\kappa_{\text{logit}} = \min_{\mathbf{u}\neq\mathbf{0}:\,\|\mathbf{u}_{\mathcal{A}^c}\|_{\ell_1}\le 3\|\mathbf{u}_{\mathcal{A}}\|_{\ell_1}} \frac{\mathbf{u}'\nabla^2\ell_n^{\text{logit}}(\boldsymbol{\beta}^\star)\mathbf{u}}{\mathbf{u}'\mathbf{u}} \in (0,\infty)$,

*if* $\lambda_{\text{lasso}} \le \frac{\kappa_{\text{logit}}}{20ms}$, *with probability at least* $1 - 2p\cdot\exp(-\frac{n}{2M}\lambda_{\text{lasso}}^2)$, *we have*

$$\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\ell_2} \le 5\kappa_{\text{logit}}^{-1} s^{1/2}\lambda_{\text{lasso}}.$$

In light of Theorem 5, we can obtain the following corollary.

COROLLARY 5. *Under assumptions* (A0), (A2) *and* (C2), *if we pick* $\lambda \ge \frac{5s^{1/2}\lambda_{\text{lasso}}}{a_0\kappa_{\text{logit}}}$, *the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - 2p\exp(-\frac{n}{2M}\lambda_{\text{lasso}}^2) - \delta_1^{\text{logit}} - \delta_2^{\text{logit}}$.

Again we can use zero to initialize the LLA algorithm and do three LLA iterations, because the first LLA iteration gives a $\ell_1$ penalized logistic regression with $\lambda_{\text{lasso}} = P_\lambda'(0)$ which equals $\lambda$ for both SCAD and MCP.

COROLLARY 6. *Consider the SCAD/MCP penalized logistic regression. Under assumptions* (A0), (A2) *and* (C2), *if* $a_0\kappa_{\text{logit}} \geq 5s^{1/2}$ *holds, the LLA algorithm initialized by zero converges to the oracle estimator after three iterations with probability at least* $1 - 2p\exp(-\frac{n}{2M}\lambda^2) - \delta_1^{\text{logit}} - \delta_2^{\text{logit}}$.

REMARK 7. The $s^{1/2}$ factor appears in Corollaries 5–6 because $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\ell_2}$ is used to bound $\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\max}$. To remove the $s^{1/2}$ factor, we can use the general invertability factor condition [Huang and Zhang (2012)] to obtain the $\ell_\infty$ loss of the LASSO. For space consideration, details are omitted.

3.3. *Sparse precision matrix estimation.* The third example is the folded concave penalized Gaussian quasi-likelihood estimator for the sparse precision matrix estimation problem, that is,

$$\min_{\boldsymbol{\Theta}\succ 0} -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_n\rangle + \sum_{(j,k)\,:\,j\neq k} P_\lambda(|\theta_{jk}|) \tag{7}$$

with the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_n = (\hat{\sigma}_{ij}^n)_{q\times q}$. Under the Gaussian assumption, the sparse precision matrix is translated into a sparse Gaussian graphical model. In this example, the target "parameter" is the true precision matrix $\boldsymbol{\Theta}^\star = (\theta_{jk}^\star)_{q\times q}$ with the support set $\mathcal{A} = \{(j,k) : \theta_{jk}^\star \neq 0\}$. Due to the symmetric structure of $\boldsymbol{\Theta}$, the dimension of the target "parameter" is $p = q(q+1)/2$, and the cardinality of $\mathcal{A}$ is $s = \#\{(j,k) : j \leq k, \theta_{jk}^\star \neq 0\}$. Moreover, we denote the maximal degree of $\boldsymbol{\Theta}^\star$ as $d = \max_j \#\{k : \theta_{jk}^\star \neq 0\}$.

In the sparse precision matrix estimation, the oracle estimator is given by

$$\widehat{\boldsymbol{\Theta}}^{\text{oracle}} = \arg\min_{\boldsymbol{\Theta}\succ\mathbf{0}\,:\,\boldsymbol{\Theta}_{\mathcal{A}^c}=\mathbf{0}} -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_n\rangle.$$

The Hessian matrix of $\ell_n(\boldsymbol{\Theta})$ is $\mathbf{H}^\star = \boldsymbol{\Sigma}^\star \otimes \boldsymbol{\Sigma}^\star$. To simplify notation, we let

$$K_1 = \|\boldsymbol{\Sigma}^\star\|_{\ell_\infty}, \qquad K_2 = \|(\mathbf{H}_{\mathcal{A}\mathcal{A}}^\star)^{-1}\|_{\ell_\infty} \quad \text{and} \quad K_3 = \|\mathbf{H}_{\mathcal{A}^c\mathcal{A}}^\star(\mathbf{H}_{\mathcal{A}\mathcal{A}}^\star)^{-1}\|_{\ell_\infty}.$$

In the next theorem, we derive the explicit bounds for $\delta_1$ and $\delta_2$ under the Gaussian assumption. Similar results hold under the exponential/polynomial tail condition in Cai, Liu and Luo (2011). Under the normality, we cite a large deviation result [Saulis and Statulevičius (1991), Bickel and Levina (2008)]:

$$\Pr(|\hat{\sigma}_{ij}^n - \sigma_{ij}^\star| \geq \nu) \leq C_0\exp(-c_0 n\nu^2) \tag{8}$$

for any $\nu$ such that $|\nu| \leq \nu_0$, where $\nu_0$, $c_0$ and $C_0$ depend on $\max_i \sigma_{ii}^\star$ only.

THEOREM 6. *Let* $\delta_0^G = \Pr(\|\widehat{\boldsymbol{\Theta}}^{\text{initial}} - \boldsymbol{\Theta}^\star\|_{\max} > a_0\lambda)$. *Assume that*

(A0′) $\|\boldsymbol{\Theta}_{\mathcal{A}}^\star\|_{\min} > (a+1)\lambda$, *and*

(A3) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are i.i.d. Gaussian samples with the true covariance* $\mathbf{\Sigma}^\star$.

*The LLA algorithm initialized by* $\widehat{\mathbf{\Theta}}^{\text{initial}}$ *converges to* $\widehat{\mathbf{\Theta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \delta_0^G - \delta_1^G - \delta_2^G$, *where*

$$\delta_1^G = C_0 s \cdot \exp\left(-\frac{c_0}{4} n \cdot \min\left\{\frac{a_1^2\lambda^2}{(2K_3+1)^2}, \frac{1}{9K_1^2K_2^2d^2}, \frac{1}{9K_1^6K_2^4d^2}\right\}\right) + C_0(p-s)\cdot\exp\left(-\frac{c_0 a_1^2}{4} n\lambda^2\right)$$

*and*

$$\delta_2^G = C_0 s \cdot \exp\left(-\frac{c_0 n}{4K_2^2} \cdot \min\left\{\frac{1}{9K_1^2d^2}, \frac{1}{9K_1^6K_2^2d^2}, (\|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - a\lambda)^2\right\}\right).$$

Theorem 6 shows that both $\delta_1^G$ and $\delta_2^G$ go to zero very quickly. Now we only need to deal with $\delta_0^G$. To initialize the LLA algorithm, we consider using the constrained $\ell_1$ minimization estimator (CLIME) by Cai, Liu and Luo (2011), that is,

$$\widehat{\mathbf{\Theta}}^{\text{clime}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1 \qquad \text{subject to } \|\widehat{\mathbf{\Sigma}}_n\mathbf{\Theta} - \mathbf{I}\|_{\max} \le \lambda_{\text{clime}}.$$

Define $L = \|\mathbf{\Theta}^\star\|_{\ell_1}$. As discussed in Cai, Liu and Luo (2011), it is reasonable to assume that $L$ is upper bounded by a constant or $L$ is some slowly diverging quantity, because $\mathbf{\Theta}^\star$ has a few nonzero entries in each row. We combine the concentration bound (8) and the same line of proof as in Cai, Liu and Luo (2011) to show that with probability at least $1 - C_0 p \cdot \exp(-\frac{c_0 n}{L^2}\lambda_{\text{clime}}^2)$, we have

$$\|\widehat{\mathbf{\Theta}}^{\text{clime}} - \mathbf{\Theta}^\star\|_{\max} \le 4L\lambda_{\text{clime}}.$$

Thus we have the following corollary.

COROLLARY 7. *Under assumptions* (A0′) *and* (A3), *if* $\lambda \ge \frac{4L}{a_0}\lambda_{\text{clime}}$, *the LLA algorithm initialized by* $\widehat{\mathbf{\Theta}}^{\text{clime}}$ *converges to* $\widehat{\mathbf{\Theta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - C_0 p \exp(-\frac{c_0 n}{L^2}\lambda_{\text{clime}}^2) - \delta_1^G - \delta_2^G$.

In the current literature, the $\ell_1$ penalized likelihood estimator GLASSO [Yuan and Lin (2007)] is perhaps the most popular estimator for sparse precision matrix estimation. However, it requires a strong irrepresentable condition [Ravikumar et al. (2011)] stating that there exists a fixed constant $\gamma_G \in (0,1)$ such that $\|\mathbf{H}^\star_{\mathcal{A}^c\mathcal{A}}(\mathbf{H}^\star_{\mathcal{A}\mathcal{A}})^{-1}\|_{\ell_\infty} \le \gamma_G$. This condition is very restrictive. If we replace the $\ell_1$ penalty with a folded concave penalty, it is interesting to see that we can obtain the oracle precision matrix estimator by using CLIME as the initial estimator in the LLA algorithm without requiring any strong structure assumption such as the irrepresentable condition.

3.4. *Sparse quantile regression.* The fourth example is the folded concave penalized quantile regression. Quantile regression [Koenker (2005)] has wide applications in statistics and econometrics. Recently, the spare quantile regression has received much attention [Zou and Yuan (2008), Li and Zhu (2008), Wu and Liu (2009), Belloni and Chernozhukov (2011), Wang, Wu and Li (2012), Fan, Fan and Barut (2014)]. We consider estimating the conditional $\tau$ quantile under

(A4) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \varepsilon$ with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ being the independent errors satisfying $\Pr(\varepsilon_i \le 0) = \tau$ for some fixed constant $\tau \in (0,1)$. Let $f_i(\cdot)$ be the density function of $\varepsilon_i$, and define $F_i(\cdot)$ as its distribution function.

Denote by $\rho_\tau(u) = u \cdot (\tau - I_{\{u \le 0\}})$ the check loss function [Koenker and Bassett (1978)]. The folded concave penalized quantile regression is given by

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_i \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \sum_j P_\lambda(|\beta_j|). \tag{9}$$

The oracle estimator of the sparse quantile regression is given by

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} = (\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}) = \arg \min_{\boldsymbol{\beta}:\, \boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}} \frac{1}{n} \sum_i \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}).$$

Note that the check loss $\rho_\tau(\cdot)$ is convex but nondifferentiable. Thus, we need to handle the subgradient $\nabla \ell_n(\boldsymbol{\beta}) = (\nabla_1 \ell_n(\boldsymbol{\beta}), \ldots, \nabla_p \ell_n(\boldsymbol{\beta}))$, where

$$\nabla_j \ell_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_i x_{ij} \cdot ((1-\tau) I_{\{y_i - \mathbf{x}_i'\boldsymbol{\beta} \le 0\}} - z_j I_{\{y_i - \mathbf{x}_i'\boldsymbol{\beta} = 0\}} - \tau I_{\{y_i - \mathbf{x}_i'\boldsymbol{\beta} > 0\}})$$

with $z_j \in [\tau - 1, \tau]$ is the subgradient of $\rho_\tau(u)$ when $u = 0$. To simplify notation, we let $M_{\mathcal{A}} = \max_i \frac{1}{s} \|\mathbf{x}_{i\mathcal{A}}\|_{\ell_2}^2$, and $m_{\mathcal{A}^c} = \max_{(i,j):\, j \in \mathcal{A}^c} |x_{ij}|$.

THEOREM 7. *Recall that* $\delta_0 = \Pr(\|\hat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\|_{\max} > a_0\lambda)$. *Suppose*

(C3) *there exist constants* $u_0 > 0$ *and* $0 < f_{\min} \le f_{\max} < \infty$ *such that for any* $u$ *satisfying* $|u| \le u_0$, $f_{\min} \le \min_i f_i(u) \le \max_i f_i(u) \le f_{\max}$.

*If* $\lambda \gg 1/n$ *such that* $\log p = o(n\lambda^2)$, $(M_{\mathcal{A}} s)^{1/2}(\|\boldsymbol{\beta}_{\mathcal{A}}^\star\|_{\min} - a\lambda) \le u_0$, *and* $m_{\mathcal{A}^c} M_{\mathcal{A}} s = o(\frac{n^{1/2}\lambda}{\log^{1/2} n})$, *the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{initial}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \delta_0 - \delta_1^Q - \delta_2^Q$, *where* $\delta_1^Q = 4n^{-1/2} + C_1(p-s) \cdot \exp(-\frac{a_1 n\lambda}{104 m_{\mathcal{A}^c}}) + 2(p-s) \cdot \exp(-\frac{a_1^2 n\lambda^2}{32 m_{\mathcal{A}^c}^2})$, *and* $\delta_2^Q = 4\exp(-\frac{\lambda_{\min}^2 f_{\min}^2}{72 M_{\mathcal{A}}} \cdot \frac{n}{s}(\|\boldsymbol{\beta}_{\mathcal{A}}^\star\|_{\min} - a\lambda)^2)$ *with* $\lambda_{\min} = \lambda_{\min}(\frac{1}{n}\mathbf{X}_{\mathcal{A}}'\mathbf{X}_{\mathcal{A}})$ *and* $C_1 > 0$ *that does not depend on* $n$, $p$ *or* $s$.

Under fairly weak assumptions, both $\delta_1^Q$ and $\delta_2^Q$ go to zero very quickly. Next, we only need to bound $\delta_0$. We consider using the $\ell_1$-penalized quantile regression as the initial estimator in the LLA algorithm, that is,

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_i \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \lambda_{\text{lasso}} \|\boldsymbol{\beta}\|_{\ell_1}.$$

To bound $\delta_0$, we use the estimation bound for $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ by Belloni and Chernozhukov (2011) and summarize the main result in the following lemma.

LEMMA 1. *Under assumption* (A4) *and the assumption of Theorem 2 in Belloni and Chernozhukov (2011), which implies* $\gamma \to 0$, *for any* $A > 1$ *and* $p^{-1} \le \alpha \to 0$, *if* $\lambda_{\text{lasso}}$ *satisfies (3.8) of Belloni and Chernozhukov (2011), with probability at least* $1 - \alpha - 4\gamma - 3p^{-A^2}$, $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *satisfies*

$$\|\hat{\boldsymbol{\beta}}^{\text{lasso}} - \boldsymbol{\beta}^\star\|_{\ell_2} \le C_2 \sqrt{\frac{s \log p}{n}},$$

*where* $C_2 > 0$ *is a fixed constant that does not depend on* $s$, $p$ *and* $n$.

Thus we have the following corollary.

COROLLARY 8. *Under the assumptions of Lemma 1, for any* $A > 1$ *and* $p^{-1} \le \alpha \to 0$, *if* $\lambda$ *such that* $\lambda \ge \frac{C_2}{a_0}\sqrt{s \log p/n}$ *and* $\lambda$ *also satisfies the conditions in Theorem 7, the LLA algorithm initialized by* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ *converges to* $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ *after two iterations with probability at least* $1 - \alpha - 4\gamma - 3p^{-A^2} - \delta_1^Q - \delta_2^Q$.

**4. Simulation studies.** In this section, we use simulation to examine the finite sample properties of the folded concave penalization for solving four classical problems. We fixed $a = 3.7$ in the SCAD and $a = 2$ in the MCP as suggested in Fan and Li (2001) and Zhang (2010a), respectively.

4.1. *Sparse regression models.* We first considered sparse linear, logistic and quantile regression models. In all examples, we simulated $n$ training data and $n$ validation data and generated $\mathbf{x} \sim N_p(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{p\times p}$.

MODEL 1 (Sparse linear regression). Set $n = 100$ and $p = 1000$. The response $y = \mathbf{x}'\boldsymbol{\beta}^\star + \varepsilon$, where $\boldsymbol{\beta}^\star = (3, 1.5, 0, 0, 2, 0_{p-5})$ and $\varepsilon \sim N(0, 1)$.

The validation error of a generic estimator $\hat{\boldsymbol{\beta}}$ for Model 1 is defined as

$$\sum_{i \in \text{validation}} (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2.$$

MODEL 2 (Sparse logistic regression). Set $n = 200$ and $p = 1000$. The response $y$ follows a Bernoulli distribution with success probability as $\frac{\exp(\mathbf{x}'\boldsymbol{\beta}^\star)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}^\star)}$, where $\boldsymbol{\beta}^\star$ is constructed by randomly choosing 10 elements in $\boldsymbol{\beta}^\star$ as $t_1 s_1, \ldots, t_{10} s_{10}$ and setting the other $p - 10$ elements as zero, where $t_j$'s are independently drawn from $\mathrm{Unif}(1, 2)$, and $s_j$'s are independent Bernoulli samples with $\Pr(s_j = 1) = \Pr(s_j = -1) = 0.5$.

The validation error of a generic estimator $\hat{\boldsymbol{\beta}}$ for Model 2 is defined as

$$\sum_{i \in \text{validation}} (-y_i \mathbf{x}_i' \hat{\boldsymbol{\beta}} + \log(1 + \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}}))).$$

MODEL 3 (Sparse quantile regression). Set $n = 100$ and $p = 400$. The response $y = \mathbf{x}'\boldsymbol{\beta}^\star + \varepsilon$ where $\boldsymbol{\beta}^\star$ is constructed in the same way as in Model 2, and $\varepsilon$ follows the standard Cauchy distribution.

We considered $\tau = 0.3, 0.5$ in the simulation. The validation error of a generic estimator $\hat{\boldsymbol{\beta}}$ for Model 3 is defined as

$$\sum_{i \in \text{validation}} \rho_\tau(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}).$$

We include $\ell_1$ penalization in the simulation study, and computed the $\ell_1$ penalized linear/logistic regression and quantile regression by R packages *glmnet* and *quantreg*. We implemented three local solutions of SCAD/MCP. The first one, denoted by SCAD-cd/MCP-cd, was the fully convergent solutions computed by coordinate descent. The second one, denoted by SCAD-lla0/MCP-lla0, was computed by the LLA algorithm initialized by zero. The third one, denoted by SCAD-lla$^\star$/MCP-lla$^\star$, was computed by the LLA algorithm initialized by the tuned LASSO estimator. SCAD-lla$^\star$/MCP-lla$^\star$ is the fully iterative LLA solution designed according to the theoretical analysis in Sections 3.1, 3.2 and 3.4. We also computed the three-step LLA solution of SCAD-lla0/MCP-lla0, denoted by SCAD-3slla0/MCP-3slla0, and the two-step LLA solution of SCAD-lla$^\star$/MCP-lla$^\star$, denoted by SCAD-2slla$^\star$/MCP-2slla$^\star$. Especially, SCAD-2slla$^\star$/MCP-2slla$^\star$ is the recommended two-step LLA solution. For each competitor, its penalization parameter was chosen by minimizing the validation error.

We conducted 100 independent runs for each model. Estimation accuracy is measured by the average $\ell_1$ loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_{\ell_1}$ and $\ell_2$ loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_{\ell_2}$, and selection accuracy is evaluated by the average counts of false positive and false negative. The simulation results are summarized in Table 1.

We can draw the following main conclusions:

(1) The local solutions solved by coordinate descent and LLA are different. LLA using different initial values are technically different algorithms, and the corresponding fully converged solutions are different, also. This mes-

TABLE 1

*Numerical comparison of LASSO, SCAD & MCP in Models 1–3. Estimation accuracy is measured by the $\ell_1$ loss and the $\ell_2$ loss, and selection accuracy is measured by counts of false negative (#FN) or false positive (#FP). Each metric is averaged over* 100 *replications with its standard error shown in the parenthesis*

| **Method** | $\ell_1$ **loss** | $\ell_2$ **loss** | **#FP** | **#FN** | $\ell_1$ **loss** | $\ell_2$ **loss** | **#FP** | **#FN** |
|---|---|---|---|---|---|---|---|---|
| | Model 1 (linear regression) | | | | Model 2 (logistic regression) | | | |
| LASSO | 1.20 | 0.46 | 14.68 | 0 | 15.08 | 3.62 | 55.92 | 0.59 |
| | (0.05) | (0.01) | (0.74) | (0) | (0.06) | (0.02) | (0.93) | (0.04) |
| SCAD-cd | 0.34 | 0.20 | 2.22 | 0 | 9.12 | 2.40 | 27.72 | 0.58 |
| | (0.02) | (0.01) | (0.40) | (0) | (0.15) | (0.04) | (0.46) | (0.04) |
| SCAD-3slla0 | 0.29 | 0.20 | 0 | 0 | 6.79 | 2.44 | 0.90 | 2.72 |
| | (0.01) | (0.01) | (0) | (0) | (0.13) | (0.03) | (0.06) | (0.08) |
| SCAD-lla0 | 0.29 | 0.20 | 0 | 0 | 6.42 | 2.34 | 0.79 | 2.73 |
| | (0.01) | (0.01) | (0) | (0) | (0.13) | (0.03) | (0.06) | (0.08) |
| SCAD-2slla$^\star$ | 0.30 | 0.20 | 0 | 0 | 6.65 | 2.41 | 0.76 | 2.58 |
| | (0.01) | (0.01) | (0) | (0) | (0.15) | (0.04) | (0.08) | (0.08) |
| SCAD-lla$^\star$ | 0.29 | 0.19 | 0 | 0 | 6.41 | 2.33 | 0.74 | 2.74 |
| | (0.01) | (0.01) | (0) | (0) | (0.14) | (0.04) | (0.06) | (0.09) |
| MCP-cd | 0.31 | 0.20 | 0.75 | 0 | 6.97 | 2.30 | 3.62 | 1.46 |
| | (0.02) | (0.01) | (0.14) | (0) | (0.16) | (0.05) | (0.14) | (0.08) |
| MPC-3slla0 | 0.30 | 0.20 | 0 | 0 | 7.10 | 2.52 | 0.94 | 2.86 |
| | (0.02) | (0.01) | (0) | (0) | (0.14) | (0.04) | (0.09) | (0.09) |
| MPC-lla0 | 0.29 | 0.20 | 0 | 0 | 6.88 | 2.45 | 1.11 | 2.81 |
| | (0.02) | (0.01) | (0) | (0) | (0.14) | (0.04) | (0.09) | (0.09) |
| MCP-2slla$^\star$ | 0.29 | 0.19 | 0 | 0 | 6.79 | 2.43 | 0.96 | 2.49 |
| | (0.02) | (0.01) | (0) | (0) | (0.14) | (0.04) | (0.08) | (0.09) |
| MCP-lla$^\star$ | 0.28 | 0.19 | 0 | 0 | 6.30 | 2.30 | 0.78 | 2.64 |
| | (0.02) | (0.01) | (0) | (0) | (0.14) | (0.04) | (0.07) | (0.08) |
| | Model 3 (quantile regression) | | | | | | | |
| | $\tau = 0.3$ | | | | $\tau = 0.5$ | | | |
| LASSO | 14.33 | 2.92 | 39.31 | 1.03 | 13.09 | 2.62 | 41.42 | 0.61 |
| | (0.35) | (0.07) | (1.29) | (0.14) | (0.33) | (0.06) | (1.18) | (0.09) |
| SCAD-3slla0 | 9.08 | 2.31 | 22.79 | 1.27 | 6.58 | 1.65 | 22.43 | 0.62 |
| | (0.43) | (0.10) | (1.02) | (0.15) | (0.38) | (0.09) | (1.03) | (0.11) |
| SCAD-lla0 | 7.70 | 2.20 | 16.08 | 1.66 | 4.47 | 1.37 | 13.48 | 0.68 |
| | (0.46) | (0.12) | (0.94) | (0.20) | (0.31) | (0.09) | (0.72) | (0.12) |
| SCAD-2slla$^\star$ | 7.43 | 2.13 | 13.26 | 1.43 | 4.80 | 1.50 | 11.43 | 0.74 |
| | (0.45) | (0.10) | (1.02) | (0.15) | (0.29) | (0.08) | (0.75) | (0.11) |
| SCAD-lla$^\star$ | 5.92 | 1.93 | 8.89 | 1.63 | 3.96 | 1.27 | 10.18 | 0.69 |
| | (0.37) | (0.11) | (0.68) | (0.19) | (0.39) | (0.08) | (0.74) | (0.11) |
| MCP-3slla0 | 10.09 | 2.71 | 19.34 | 1.58 | 7.44 | 1.93 | 17.54 | 0.88 |
| | (0.44) | (0.09) | (1.10) | (0.17) | (0.44) | (0.10) | (0.77) | (0.14) |
| MCP-lla0 | 9.86 | 2.69 | 11.63 | 2.18 | 5.79 | 1.70 | 9.45 | 1.04 |
| | (0.53) | (0.12) | (0.95) | (0.21) | (0.44) | (0.10) | (0.77) | (0.14) |
| MCP-2slla$^\star$ | 6.13 | 2.15 | 5.10 | 1.75 | 4.48 | 1.54 | 3.53 | 1.03 |
| | (0.44) | (0.10) | (0.54) | (0.16) | (0.29) | (0.09) | (0.47) | (0.13) |
| MCP-lla$^\star$ | 5.95 | 2.05 | 2.92 | 1.91 | 3.88 | 1.39 | 2.04 | 1.00 |
| | (0.42) | (0.11) | (0.50) | (0.18) | (0.29) | (0.09) | (0.29) | (0.13) |

sage clearly shows that the concave regularization problem does have multiple local minimizers and multiple sparse local minimizers.

(2) SCAD-2slla$^\star$/MCP-2slla$^\star$ are recommended based on our philosophy and theory (see Remark 3). SCAD-2slla$^\star$/MCP-2slla$^\star$ is asymptotically equivalent to SCAD-lla$^\star$/MCP-lla$^\star$, but two-step solutions are cheaper to compute than fully converged ones. We do expect to see they differ with finite sample size, but the difference is ignorable as in Table 1.

(3) We included SCAD-3slla0/MCP-3slla0 because of Corollaries 4 and 6 that justify the use of zero as a good initial value in the LLA algorithm under some extra conditions. The simulation results show that zero can be a good initial value but it is not the best choice one would try.

4.2. *Sparse Gaussian graphical model.* We drew $n = 100$ training data and $n$ validation data from $N_{q=100}(\mathbf{0}, \mathbf{\Sigma}^\star)$ with a sparse precision matrix $\mathbf{\Theta}^\star$.

MODEL 4. $\mathbf{\Theta}^\star$ is a tridiagonal matrix by setting $\mathbf{\Sigma}^\star = (\sigma_{ij}^\star)_{q\times q}$ as an AR(1) covariance matrix with $\sigma_{ij}^\star = \exp(-|s_i - s_j|)$ for $s_1 < \cdots < s_q$, which draws $s_q - s_{q-1}, s_{q-1} - s_{q-2}, \ldots, s_2 - s_1$ independently from Unif$(0.5, 1)$.

MODEL 5. $\mathbf{\Theta}^\star = \mathbf{U}'_{q\times q}\mathbf{U}_{q\times q} + \mathbf{I}_{q\times q}$ where $\mathbf{U} = (u_{ij})_{q\times q}$ has zero diagonals and 100 nonzero off-diagonal entries. The nonzero entries are generated by $u_{ij} = t_{ij}s_{ij}$ where $t_{ij}$'s are independently drawn from Unif$(1, 2)$, and $s_{ij}$'s are independent Bernoulli variables with $\Pr(s_{ij} = \pm 1) = 0.5$.

The validation error of a generic estimator $\widehat{\mathbf{\Theta}}$ for Models 4–5 is defined as

$$-\log\det(\widehat{\mathbf{\Theta}}) + \langle \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Sigma}}_n^{\text{validation}} \rangle.$$

We computed the GLASSO and the CLIME by the R packages *glasso* and *clime*. We computed two local solutions of the SCAD/MCP penalized estimator denoted by GSCAD/GMCP. The first one, denoted by GSCAD-lla0/GMCP-lla0, used $\text{diag}(\widehat{\Sigma}_{jj}^{-1})$ as the initial solution in the LLA algorithm. The second one, denoted by GSCAD-lla$^\star$/GMCP-lla$^\star$, used the tuned CLIME to initialize the LLA algorithm. GSCAD-lla$^\star$/GMCP-lla$^\star$ was designed according to the theoretical analysis in Section 3.3. We computed the three-step LLA solution of GSCAD-lla0/GMCP-lla0, denoted by GSCAD-3slla0/GMCP-3slla0, and the two-step LLA solution of GSCAD-lla$^\star$/GMCP-lla$^\star$, denoted by GSCAD-2slla$^\star$/GMCP-2slla$^\star$. For each competitor, its penalization parameter was chosen by minimizing the validation error.

For each model, we conducted 100 independent runs. Estimation accuracy is measured by the average Operator norm loss $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\ell_2}$ and Frobenius

TABLE 2

*Numerical comparison of GLASSO, CLIME, GSCAD and GMCP in Model 5. Estimation accuracy is measured by the Operator norm ($\|\cdot\|_{\ell_2}$) and the Frobenius norm ($\|\cdot\|_F$), and selection accuracy is measured by counts of false negative (#FN) or false positive (#FP). Each metric is averaged over 100 replications with its standard error in the parenthesis*

| **Method** | $\|\cdot\|_{\ell_2}$ | $\|\cdot\|_F$ | **#FP** | **#FN** | $\|\cdot\|_{\ell_2}$ | $\|\cdot\|_F$ | **#FP** | **#FN** |
|---|---|---|---|---|---|---|---|---|
| | Model 4 | | | | Model 5 | | | |
| GLASSO | 1.45 | 6.12 | 743.56 | 1.34 | 11.63 | 25.45 | 236.76 | 56.16 |
| | (0.01) | (0.02) | (10.75) | (0.17) | (0.02) | (0.03) | (5.19) | (0.52) |
| CLIME | 1.40 | 5.89 | 741.16 | 2.42 | 8.56 | 18.40 | 323.04 | 12.26 |
| | (0.01) | (0.03) | (12.80) | (0.24) | (0.05) | (0.08) | (7.22) | (0.38) |
| GSCAD-3slla0 | 1.20 | 4.59 | 659.04 | 1.78 | 10.84 | 22.05 | 225.36 | 54.98 |
| | (0.02) | (0.03) | (9.41) | (0.20) | (0.05) | (0.12) | (4.92) | (0.58) |
| GSCAD-lla0 | 1.16 | 4.42 | 641.82 | 1.96 | 10.73 | 20.68 | 228.70 | 54.54 |
| | (0.02) | (0.03) | (9.41) | (0.20) | (0.05) | (0.12) | (4.92) | (0.58) |
| GSCAD-2slla$^\star$ | 1.20 | 4.60 | 660.84 | 1.74 | 6.49 | 13.69 | 203.52 | 28.78 |
| | (0.02) | (0.03) | (9.39) | (0.19) | (0.13) | (0.15) | (5.27) | (0.57) |
| GSCAD-lla$^\star$ | 1.16 | 4.42 | 635.49 | 1.94 | 6.42 | 13.36 | 196.60 | 30.02 |
| | (0.02) | (0.03) | (9.39) | (0.19) | (0.13) | (0.15) | (5.27) | (0.57) |
| GMCP-3slla0 | 1.57 | 4.62 | 349.36 | 3.56 | 10.41 | 20.40 | 201.06 | 54.12 |
| | (0.04) | (0.04) | (7.03) | (0.21) | (0.07) | (0.17) | (4.26) | (0.68) |
| GMCP-lla0 | 1.53 | 4.56 | 291.04 | 6.45 | 10.34 | 19.20 | 200.37 | 52.24 |
| | (0.04) | (0.04) | (5.12) | (0.32) | (0.07) | (0.12) | (4.26) | (0.60) |
| GMCP-2slla$^\star$ | 1.40 | 4.37 | 290.10 | 3.96 | 5.98 | 12.74 | 68.14 | 23.84 |
| | (0.04) | (0.04) | (4.71) | (0.20) | (0.17) | (0.17) | (3.42) | (0.61) |
| GMCP-lla$^\star$ | 1.39 | 4.31 | 229.87 | 6.29 | 5.80 | 12.72 | 44.79 | 25.18 |
| | (0.03) | (0.04) | (4.56) | (0.33) | (0.15) | (0.17) | (3.12) | (0.53) |

norm loss $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F$, and selection accuracy is evaluated by the average counts of false positive and false negative. The simulation results are summarized in Table 2. The main conclusions are the same as those in spare regression models. First, the fully converged LLA solutions are different with different initial values, so it is impractical to try to prove that this problem has a unique minimizer. Second, with the CLIME as the initial value, the two-step LLA solutions perform as well as the fully converged LLA solutions.

## 5. Technical proofs.

5.1. *Proof of Theorem 1.* Define $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{\text{initial}}$. Under the event $\{\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^\star\|_{\max} \le a_0\lambda\}$, due to assumption (A0), we have $|\hat{\beta}_j^{(0)}| \le \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^\star\|_{\max} \le a_0\lambda \le a_2\lambda$ for $j \in \mathcal{A}^c$, and $|\hat{\beta}_j^{(0)}| \ge \|\boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\min} - \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^\star\|_{\max} > a\lambda$ for $j \in \mathcal{A}$.

By property (iv), $P'_\lambda(|\hat{\beta}_j^{(0)}|) = 0$ for all $j \in \mathcal{A}$. Thus $\hat{\boldsymbol{\beta}}^{(1)}$ is the solution to the problem

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + \sum_{j \in \mathcal{A}^c} P'_\lambda(|\hat{\beta}_j^{(0)}|) \cdot |\beta_j|. \tag{10}$$

By properties (ii) and (iii), $P'_\lambda(|\hat{\beta}_j^{(0)}|) \geq a_1\lambda$ holds for $j \in \mathcal{A}^c$. We now show that $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ is the unique global solution to (10) under the additional condition $\{\|\nabla_{\mathcal{A}^c} \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} < a_1\lambda\}$. To see this, note that by convexity, we have

$$\begin{aligned} \ell_n(\boldsymbol{\beta}) &\geq \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) + \sum_j \nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})(\beta_j - \hat{\beta}_j^{\text{oracle}}) \\ &= \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) + \sum_{j \in \mathcal{A}^c} \nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})(\beta_j - \hat{\beta}_j^{\text{oracle}}), \end{aligned} \tag{11}$$

where the last equality used (3). By (11) and $\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}^c} = \mathbf{0}$, for any $\boldsymbol{\beta}$ we have

$$\begin{aligned} &\left\{ \ell_n(\boldsymbol{\beta}) + \sum_{j \in \mathcal{A}^c} P'_\lambda(|\hat{\beta}_j^{(0)}|)|\beta_j| \right\} - \left\{ \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) + \sum_{j \in \mathcal{A}^c} P'_\lambda(|\hat{\beta}_j^{(0)}|)|\hat{\beta}_j^{\text{oracle}}| \right\} \\ &\quad \geq \sum_{j \in \mathcal{A}^c} \{P'_\lambda(|\hat{\beta}_j^{(0)}|) - \nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) \cdot \operatorname{sign}(\beta_j)\} \cdot |\beta_j| \\ &\quad \geq \sum_{j \in \mathcal{A}^c} \{a_1\lambda - \nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) \cdot \operatorname{sign}(\beta_j)\} \cdot |\beta_j| \\ &\quad \geq 0. \end{aligned}$$

The strict inequality holds unless $\beta_j = 0$, $\forall j \in \mathcal{A}^c$. This together with the uniqueness of the solution to (2) concludes that $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ is the unique solution to (10). Hence, $\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{\text{oracle}}$, which completes the proof of Theorem 1.

5.2. *Proof of Theorem 2.* Given that the LLA algorithm finds $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ at the current iteration, we denote $\hat{\boldsymbol{\beta}}$ as the solution to the convex optimization problem in the next iteration of the LLA algorithm. Using $\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}^c} = \mathbf{0}$ and $P'_\lambda(|\hat{\beta}_j^{\text{oracle}}|) = 0$ for $j \in \mathcal{A}$ under the event $\{\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} > a\lambda\}$, we have

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + \sum_{j \in \mathcal{A}^c} \gamma \cdot |\beta_j|, \tag{12}$$

where $\gamma = P'_\lambda(0) \geq a_1\lambda$. This problem is very similar to (10). We can follow the proof of Theorem 1 to show that under the additional condition $\{\|\nabla_{\mathcal{A}^c} \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} < a_1\lambda\}$, $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ is the unique solution to (12). Hence, the LLA algorithm converges, which completes the proof of Theorem 2.

5.3. *Proof of Theorem 3.* Let $\mathbf{H}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}$. Since $\mathbf{y} = \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}^{\star}_{\mathcal{A}} + \varepsilon$, we have $\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \frac{1}{n}\mathbf{X}'_{\mathcal{A}^c}(\mathbf{I}_{n\times n} - \mathbf{H}_{\mathcal{A}})\varepsilon$. By the Chernoff bound, we have

$$\begin{aligned}\delta_1 &\le \sum_{j\in\mathcal{A}^c} \Pr(|\mathbf{x}'_{(j)}(\mathbf{I}_{n\times n} - \mathbf{H}_{\mathcal{A}})\varepsilon| > a_1 n\lambda)\\ &\le 2\sum_{j\in\mathcal{A}^c} \exp\left(-\frac{a_1^2 n^2\lambda^2}{2\sigma^2\cdot\|\mathbf{x}'_{(j)}(\mathbf{I}_{n\times n} - \mathbf{H}_{\mathcal{A}})\|_{\ell_2}^2}\right).\end{aligned}$$

Since $\|\mathbf{x}'_{(j)}(\mathbf{I}_{n\times n} - \mathbf{H}_{\mathcal{A}})\|_{\ell_2}^2 = \mathbf{x}'_{(j)}(\mathbf{I}_{n\times n} - \mathbf{H}_{\mathcal{A}})\mathbf{x}_{(j)} \le nM$, we conclude that

$$\delta_1 \le 2(p-s)\exp\left(-\frac{a_1^2 n\lambda^2}{2M\sigma^2}\right).$$

Now we bound $\delta_2$. Note that $\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}} = \boldsymbol{\beta}^{\star}_{\mathcal{A}} + (\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\varepsilon$, and then $\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \ge \|\boldsymbol{\beta}^{\star}_{\mathcal{A}}\|_{\min} - \|(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\varepsilon\|_{\max}$. Thus, we have

$$\delta_2 \le \Pr(\|(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\varepsilon\|_{\max} \ge \|\boldsymbol{\beta}^{\star}_{\mathcal{A}}\|_{\min} - a\lambda). \tag{13}$$

It remains to derive an explicit bound for (13). To simplify notation, we let $(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_s)'$, with $\mathbf{u}_j = \mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{e}_j$, where $\mathbf{e}_j$ is the unit vector with $j$th element 1. It is obvious to see that $\|\mathbf{u}_j\|_{\ell_2}^2 = \mathbf{e}'_j(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{e}'_j \le (n\lambda_{\min})^{-1}$. By the Chernoff bound, we have

$$\begin{aligned}\delta_2 &\le \Pr(\|(\mathbf{X}'_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}'_{\mathcal{A}}\varepsilon\|_{\max} \ge \|\boldsymbol{\beta}^{\star}_{\mathcal{A}}\|_{\min} - a\lambda)\\ &\le 2\sum_{j=1}^{s}\exp\left(-\frac{(\|\boldsymbol{\beta}^{\star}_{\mathcal{A}}\|_{\min} - a\lambda)^2}{2\sigma^2\|\mathbf{u}_j\|_{\ell_2}^2}\right)\\ &\le 2s\exp\left(-\frac{n\lambda_{\min}}{2\sigma^2}(\|\boldsymbol{\beta}^{\star}_{\mathcal{A}}\|_{\min} - a\lambda)^2\right).\end{aligned}$$

Thus, we complete the proof of Theorem 3.

5.4. *Proof of Theorem 4.* A translation of (3) into our setting becomes

$$\mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \mathbf{X}'_{\mathcal{A}}\mathbf{y}. \tag{14}$$

We now use this to bound $\delta_2$.

Define a map $F:\mathbb{B}(r)\subset\mathbb{R}^p \to \mathbb{R}^p$ satisfying $F(\boldsymbol{\Delta}) = ((F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}}))', \mathbf{0}')'$ with $F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}}) = (\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^{\star})\mathbf{X}_{\mathcal{A}})^{-1}\cdot\mathbf{X}'_{\mathcal{A}}(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{\star} + \boldsymbol{\Delta})) + \boldsymbol{\Delta}_{\mathcal{A}}$ and the convex compact set $\mathbb{B}(r) = \{\boldsymbol{\Delta}\in\mathbb{R}^p : \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\max} \le r, \boldsymbol{\Delta}_{\mathcal{A}^c} = \mathbf{0}\}$ with $r = 2Q_2\cdot\|\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^{\star}) - \mathbf{y})\|_{\max}$. Our aim is to show

$$F(\mathbb{B}(r)) \subset \mathbb{B}(r), \tag{15}$$

when

$$\left\|\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star)-\mathbf{y})\right\|_{\max} \le \frac{1}{Q_1Q_2^2 s}. \tag{16}$$

If (15) holds, by the Brouwer's fixed-point theorem, there always exists a fixed point $\widehat{\boldsymbol{\Delta}} \in \mathbb{B}(r)$ such that $F(\widehat{\boldsymbol{\Delta}}) = \widehat{\boldsymbol{\Delta}}$. It immediately follows that $\mathbf{X}'_{\mathcal{A}}\mathbf{y} = \mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star + \widehat{\boldsymbol{\Delta}})$ and $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}^c} = \mathbf{0}$, which implies $\boldsymbol{\beta}^\star + \widehat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}}^{\text{oracle}}$ by uniqueness of the solution to (14). Thus,

$$\|\hat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^\star\|_{\max} = \|\widehat{\boldsymbol{\Delta}}\|_{\max} \le r. \tag{17}$$

If further

$$\left\|\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star)-\mathbf{y})\right\|_{\max} \le \frac{1}{2Q_2}(\|\boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\min} - a\lambda),$$

we have $r \le \|\boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\min} - a\lambda$, and then $\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} \ge a\lambda$. Therefore, we have

$$\delta_2 \le \Pr\left(\left\|\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star)-\mathbf{y})\right\|_{\max} > \min\left\{\frac{1}{Q_1Q_2^2 s}, \frac{1}{2Q_2}(\|\boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\min} - a\lambda)\right\}\right).$$

By the Hoeffding's bound in Proposition 4(a) of Fan and Lv (2011), we have

$$\delta_2 \le 2s \cdot \exp\left(-\frac{n}{MQ_2^2}\cdot \min\left\{\frac{2}{Q_1^2Q_2^2 s^2}, \frac{1}{2}(\|\boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\min} - a\lambda)^2\right\}\right).$$

We now derive (15). Using its Taylor expansion around $\boldsymbol{\Delta} = \mathbf{0}$, we have

$$\mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) = \mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star) + \mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}\boldsymbol{\Delta} + \mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}}),$$

where $\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}}) = \mathbf{X}'_{\mathcal{A}}(\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}^\star))\mathbf{X}\boldsymbol{\Delta}$ with $\widetilde{\boldsymbol{\Delta}}$ on the line segment joining $\mathbf{0}$ and $\boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}_{\mathcal{A}^c} = \mathbf{0}$, we have $\mathbf{X}\boldsymbol{\Delta} = \mathbf{X}_{\mathcal{A}}\boldsymbol{\Delta}_{\mathcal{A}}$. By the mean-value theorem, we have $\|\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})\|_{\max} \le \max_j \boldsymbol{\Delta}'_{\mathcal{A}}\mathbf{X}'_{\mathcal{A}}\operatorname{diag}\{|\mathbf{x}_{(j)}| \circ |\boldsymbol{\mu}''(\bar{\boldsymbol{\beta}})|\}\mathbf{X}_{\mathcal{A}}\boldsymbol{\Delta}_{\mathcal{A}}$ for $\bar{\boldsymbol{\beta}}$ being on the line segment joining $\boldsymbol{\beta}^\star$ and $\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}$. Using the fact that $|\psi'''(t)| = \theta(t)(1-\theta(t))|2\theta(t)-1| \le \frac{1}{4}$ with $\theta(t) = (1+\exp(t))^{-1}$, we have

$$\|\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})\|_{\max} \le \frac{n}{4}Q_1 \cdot \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_2}^2 \le \frac{n}{4}Q_1 s r^2. \tag{18}$$

Notice that

$$\begin{aligned} F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}}) &= (\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}})^{-1}(\mathbf{X}'_{\mathcal{A}}\mathbf{y} - \mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star + \boldsymbol{\Delta})) + \boldsymbol{\Delta}_{\mathcal{A}} \\ &= (\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}})^{-1} \cdot (\mathbf{X}'_{\mathcal{A}}\mathbf{y} - \mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star) - \mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})), \end{aligned}$$

we then use the triangle inequality to obtain

$$\begin{aligned} \|F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}})\|_{\max} &= \|(\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}_{\mathcal{A}})^{-1} \cdot (\mathbf{X}'_{\mathcal{A}}\mathbf{y} - \mathbf{X}'_{\mathcal{A}}\boldsymbol{\mu}(\boldsymbol{\beta}^\star) - \mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}}))\|_{\max} \\ &\le Q_2 \cdot \left(\left\|\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star)-\mathbf{y})\right\|_{\max} + \frac{1}{n}\|\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})\|_{\max}\right). \end{aligned}$$

By (18) and the definition of $r$, we have $\|F_{\mathcal{A}}(\boldsymbol{\Delta}_{\mathcal{A}})\|_{\max} \le \frac{r}{2} + \frac{1}{4}Q_1Q_2sr^2 \le r$. This establishes the desired contraction (15).

Next, we prove the upper bound for $\delta_1$. Recall that $\widehat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^\star$. Recall that $\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n\{-y_i\mathbf{x}_i'\boldsymbol{\beta} + \psi(\mathbf{x}_i'\boldsymbol{\beta})\}$. By a Taylor expansion,

$$\nabla\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \nabla\ell_n(\boldsymbol{\beta}^\star) + \nabla^2\ell_n(\boldsymbol{\beta}^\star)\cdot\widehat{\boldsymbol{\Delta}} + (\nabla^2\ell_n(\tilde{\boldsymbol{\beta}}) - \nabla^2\ell_n(\boldsymbol{\beta}^\star))\cdot\widehat{\boldsymbol{\Delta}}, \tag{19}$$

where $\tilde{\boldsymbol{\beta}}$ is on the line segment joining $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ and $\boldsymbol{\beta}^\star$. Observe that the first and second derivatives of $\ell_n(\boldsymbol{\beta})$ can be explicitly written as

$$\nabla\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}'(\boldsymbol{\mu}(\boldsymbol{\beta}) - \mathbf{y}) \quad \text{and} \quad \nabla^2\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\beta})\mathbf{X}. \tag{20}$$

We define $\mathbf{R}(\boldsymbol{\Delta}) = (\nabla^2\ell_n(\tilde{\boldsymbol{\beta}}) - \nabla^2\ell_n(\boldsymbol{\beta}^\star))\cdot\widehat{\boldsymbol{\Delta}} = \mathbf{X}'(\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - \boldsymbol{\Sigma}(\boldsymbol{\beta}^\star))\mathbf{X}\widehat{\boldsymbol{\Delta}}$. We rewrite $\mathbf{R}(\boldsymbol{\Delta})$ as $(\mathbf{R}'_{\mathcal{A}}(\boldsymbol{\Delta}), \mathbf{R}'_{\mathcal{A}^c}(\boldsymbol{\Delta}))'$. Let $\widetilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$. Then, using $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}^c} = \mathbf{0}$, we have $\mathbf{X}\widehat{\boldsymbol{\Delta}} = \mathbf{X}_{\mathcal{A}}\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}$. Substituting this into (19), we obtain

$$\nabla_{\mathcal{A}}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star) + \frac{1}{n}\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}}\widehat{\boldsymbol{\Delta}}_{\mathcal{A}} + \frac{1}{n}\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}}) \tag{21}$$

and

$$\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \nabla_{\mathcal{A}^c}\ell_n(\boldsymbol{\beta}^\star) + \frac{1}{n}\mathbf{X}'_{\mathcal{A}^c}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}}\widehat{\boldsymbol{\Delta}}_{\mathcal{A}} + \frac{1}{n}\mathbf{R}_{\mathcal{A}^c}(\widetilde{\boldsymbol{\Delta}}). \tag{22}$$

Using (20) for $\boldsymbol{\beta}^\star$ and $\nabla_{\mathcal{A}}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = \mathbf{0}$, we solve for $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}$ from (21) and substitute it into (22) to obtain

$$\begin{aligned}
&\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) \\
&\quad= \mathbf{X}'_{\mathcal{A}^c}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}}(\mathbf{X}'_{\mathcal{A}}\boldsymbol{\Sigma}(\boldsymbol{\beta}^\star)\mathbf{X}_{\mathcal{A}})^{-1}\left(-\frac{1}{n}\mathbf{X}'_{\mathcal{A}}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star) - \mathbf{y}) - \frac{1}{n}\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})\right) \\
&\qquad+ \frac{1}{n}\mathbf{X}'_{\mathcal{A}^c}(\boldsymbol{\mu}(\boldsymbol{\beta}^\star) - \mathbf{y}) + \frac{1}{n}\mathbf{R}_{\mathcal{A}^c}(\widetilde{\boldsymbol{\Delta}}).
\end{aligned}$$

Recall that we have proved that (17) holds under the condition (16). Now under the condition (16) and the additional event

$$\left\{\|\nabla_{\mathcal{A}^c}\ell_n(\boldsymbol{\beta}^\star)\|_{\max} < \frac{a_1\lambda}{2}\right\} \cap \left\{\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max} \le \frac{a_1\lambda}{4Q_3 + 2}\right\},$$

we can follow the same lines of proof as in (18) to show that

$$\|\mathbf{R}(\widetilde{\boldsymbol{\Delta}})\|_{\max} \le \frac{n}{4}Q_1\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\|_{\ell_2}^2 \le \frac{n}{4}Q_1sr^2,$$

where $r = 2Q_2\cdot\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}$. Noticing that under condition (16)

$$\frac{n}{4}Q_1sr^2 = snQ_1Q_2^2\cdot\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}^2 \le n\cdot\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max},$$

under the same event we have

$$\begin{aligned}\|\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\mathrm{oracle}})\| &\le Q_3\cdot\left(\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}+\frac{1}{n}\|\mathbf{R}_{\mathcal{A}}(\widetilde{\boldsymbol{\Delta}})\|_{\max}\right)\\ &\quad+\|\nabla_{\mathcal{A}^c}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}+\frac{1}{n}\|\mathbf{R}_{\mathcal{A}^c}(\widetilde{\boldsymbol{\Delta}})\|_{\max}\\ &\le (2Q_3+1)\cdot\|\nabla_{\mathcal{A}}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}+\|\nabla_{\mathcal{A}^c}\ell_n(\boldsymbol{\beta}^\star)\|_{\max}\\ &< a_1\lambda.\end{aligned}$$

The desired probability bound can be obtained by using Proposition 4(a) of Fan and Lv (2011). This completes the proof of Theorem 4.

5.5. *Proof of Theorem 5.* The proof is relegated to a supplementary file [Fan, Xue and Zou (2014)] for the sake of space constraint.

5.6. *Proof of Theorem 6.* We first derive bound $\delta_2=\Pr(\|\widehat{\boldsymbol{\Theta}}^{\mathrm{oracle}}_{\mathcal{A}}\|_{\min}\le a\lambda)$. A translation of (3) into the precision matrix estimation setting becomes $\widehat{\boldsymbol{\Sigma}}^{\mathrm{oracle}}_{\mathcal{A}}=\widehat{\boldsymbol{\Sigma}}^{n}_{\mathcal{A}}$. Let $\boldsymbol{\Sigma}^{\boldsymbol{\Delta}}=(\boldsymbol{\Theta}^\star+\boldsymbol{\Delta})^{-1}$ and $r=2K_2\|\widehat{\boldsymbol{\Sigma}}^{n}_{\mathcal{A}}-\boldsymbol{\Sigma}^\star_{\mathcal{A}}\|_{\max}$. We define a map $F:\mathbb{B}(r)\subset\mathbb{R}^{p^2}\to\mathbb{R}^{p^2}$ satisfying $F(\mathrm{vec}(\boldsymbol{\Delta}))=((F_{\mathcal{A}}(\mathrm{vec}(\boldsymbol{\Delta}_{\mathcal{A}})))',\mathbf{0}')'$ with

$$F_{\mathcal{A}}(\mathrm{vec}(\boldsymbol{\Delta}_{\mathcal{A}}))=(\mathbf{H}^\star_{\mathcal{A}\mathcal{A}})^{-1}\cdot(\mathrm{vec}(\boldsymbol{\Sigma}^{\boldsymbol{\Delta}}_{\mathcal{A}})-\mathrm{vec}(\widehat{\boldsymbol{\Sigma}}^{n}_{\mathcal{A}}))+\mathrm{vec}(\boldsymbol{\Delta}_{\mathcal{A}}) \tag{23}$$

and $\mathbb{B}(r)=\{\boldsymbol{\Delta}:\|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\max}\le r,\boldsymbol{\Delta}_{\mathcal{A}^c}=\mathbf{0}\}$. We will show that

$$F(\mathbb{B}(r))\subset\mathbb{B}(r) \tag{24}$$

under the condition

$$\|\widehat{\boldsymbol{\Sigma}}^{n}_{\mathcal{A}}-\boldsymbol{\Sigma}^\star_{\mathcal{A}}\|_{\max}<\min\left\{\frac{1}{6K_1K_2d},\frac{1}{6K_1^3K_2^2d}\right\}. \tag{25}$$

If (24) holds, an application of the Brouwer's fixed-point theorem yields a fixed-point $\widehat{\boldsymbol{\Delta}}$ in the convex compact set $\mathbb{B}(r)$ satisfying $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}^c}=\mathbf{0}$ and $F_{\mathcal{A}}(\mathrm{vec}(\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}))=\mathrm{vec}(\widehat{\boldsymbol{\Delta}}_{\mathcal{A}})$. Thus, $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}=\widehat{\boldsymbol{\Theta}}^{\mathrm{oracle}}_{\mathcal{A}}-\boldsymbol{\Theta}^\star_{\mathcal{A}}$ by the uniqueness and

$$\|\widehat{\boldsymbol{\Theta}}^{\mathrm{oracle}}-\boldsymbol{\Theta}^\star\|_{\max}=\|\widehat{\boldsymbol{\Delta}}\|_{\max}\le r. \tag{26}$$

We now establish (24). For any $\boldsymbol{\Delta}\in\mathbb{B}(r)$, by using (25) we have

$$\|\boldsymbol{\Sigma}^\star\boldsymbol{\Delta}\|_{\ell_\infty}\le K_1\cdot\|\boldsymbol{\Delta}\|_{\ell_1}\le K_1\cdot dr=2K_1K_2d\cdot\|\widehat{\boldsymbol{\Sigma}}^{n}_{\mathcal{A}}-\boldsymbol{\Sigma}^\star_{\mathcal{A}}\|_{\max}<\tfrac{1}{3}.$$

Thus, $\mathbf{J}=\sum_{j=0}^{\infty}(-1)^j(\boldsymbol{\Sigma}^\star\boldsymbol{\Delta})^j$ is a convergent matrix series of $\boldsymbol{\Delta}$. Hence,

$$\boldsymbol{\Sigma}^{\boldsymbol{\Delta}}=(\mathbf{I}+\boldsymbol{\Sigma}^\star\boldsymbol{\Delta})^{-1}\cdot\boldsymbol{\Sigma}^\star=\boldsymbol{\Sigma}^\star-\boldsymbol{\Sigma}^\star\boldsymbol{\Delta}\boldsymbol{\Sigma}^\star+\mathbf{R}^{\boldsymbol{\Delta}}, \tag{27}$$

where $\mathbf{R}^{\mathbf{\Delta}} = (\mathbf{\Sigma}^\star\mathbf{\Delta})^2 \cdot \mathbf{J}\mathbf{\Sigma}^\star$. Then it immediately yields that

$$\begin{aligned}(28)\qquad &\operatorname{vec}(\mathbf{\Sigma}^{\mathbf{\Delta}}_{\mathcal{A}}) - \operatorname{vec}(\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}}) \\ &\quad = (\operatorname{vec}(\mathbf{\Sigma}^\star_{\mathcal{A}}) - \operatorname{vec}(\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}})) - \operatorname{vec}(\mathbf{\Sigma}^\star\mathbf{\Delta}\mathbf{\Sigma}^\star) + \operatorname{vec}(\mathbf{R}^{\mathbf{\Delta}}_{\mathcal{A}}).\end{aligned}$$

Note that $\mathbf{\Sigma}^\star\mathbf{\Delta}\mathbf{\Sigma}^\star = (\mathbf{\Sigma}^\star \otimes \mathbf{\Sigma}^\star) \cdot \operatorname{vec}(\mathbf{\Delta}) = \mathbf{H}^\star \cdot \operatorname{vec}(\mathbf{\Delta})$, and hence

$$\operatorname{vec}(\mathbf{\Sigma}^\star\mathbf{\Delta}\mathbf{\Sigma}^\star_{\mathcal{A}}) = \mathbf{H}^\star_{\mathcal{A}\mathcal{A}} \cdot \operatorname{vec}(\mathbf{\Delta}_{\mathcal{A}}).$$

Now we follow the proof of Lemma 5 in Ravikumar et al. (2011) to obtain

$$(29)\qquad \|\mathbf{R}^{\mathbf{\Delta}}\|_{\max} = \max_{(i,j)} |\mathbf{e}_i'((\mathbf{\Sigma}^\star\mathbf{\Delta})^2 \cdot \mathbf{J}\mathbf{\Sigma}^\star)\mathbf{e}_j| \le \frac{3}{2}K_1^3 \cdot d\|\Delta\|^2_{\max}.$$

Hence, a combination of (23), (28) and (29) yields the contraction (24), that is,

$$\begin{aligned}&\|F_{\mathcal{A}}(\operatorname{vec}(\mathbf{\Delta}_{\mathcal{A}}))\|_{\max} \\ &\quad = \|(\mathbf{H}^\star_{\mathcal{A}\mathcal{A}})^{-1} \cdot ((\operatorname{vec}(\mathbf{\Sigma}^\star_{\mathcal{A}}) - \operatorname{vec}(\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}})) + \operatorname{vec}(\mathbf{R}^{\mathbf{\Delta}}_{\mathcal{A}}))\|_{\max} \\ &\quad \le K_2 \cdot (\|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}} - \mathbf{\Sigma}^\star_{\mathcal{A}}\|_{\max} + \|\mathbf{R}^{\mathbf{\Delta}}\|_{\max}) \\ &\quad \le r.\end{aligned}$$

Under the additional condition,

$$\|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}} - \mathbf{\Sigma}^\star_{\mathcal{A}}\|_{\max} < \frac{1}{2K_2}(\|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - a\lambda)$$

by (26) and the definition or $r$, we have that

$$\begin{aligned}\|\widehat{\mathbf{\Theta}}^{\text{oracle}}_{\mathcal{A}}\|_{\min} &\ge \|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - \|\widehat{\mathbf{\Theta}}^{\text{oracle}} - \mathbf{\Theta}^\star\|_{\max} \\ &= \|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - 2K_2 \cdot \|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}} - \mathbf{\Sigma}^\star_{\mathcal{A}}\|_{\max} \\ &> a\lambda.\end{aligned}$$

Thus,

$$\delta_2 \le \Pr\left(\|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}} - \mathbf{\Sigma}^\star_{\mathcal{A}}\|_{\max} > \frac{1}{2K_2}\min\left\{\frac{1}{3K_1 d}, \frac{1}{3K_1^3 K_2 d}, \|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - a\lambda\right\}\right).$$

An application of (8) yields the bound on $\delta_2$.

We now bound $\delta_1$. Note that $\nabla_{\mathcal{A}^c}\ell_n(\widehat{\mathbf{\Theta}}^{\text{oracle}}) = \widehat{\mathbf{\Sigma}}^n_{\mathcal{A}^c} - \widehat{\mathbf{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c}$, and hence

$$(30)\qquad \|\nabla_{\mathcal{A}^c}\ell_n(\widehat{\mathbf{\Theta}}^{\text{oracle}})\|_{\max} \le \|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}^c} - \mathbf{\Sigma}^\star_{\mathcal{A}^c}\|_{\max} + \|\widehat{\mathbf{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c} - \mathbf{\Sigma}^\star_{\mathcal{A}^c}\|_{\max}.$$

Note $\|\widehat{\mathbf{\Sigma}}^n_{\mathcal{A}^c} - \mathbf{\Sigma}^\star_{\mathcal{A}^c}\|_{\max}$ is bounded by using (8). Then we only need to bound $\|\widehat{\mathbf{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c} - \mathbf{\Sigma}^\star_{\mathcal{A}^c}\|_{\max}$. Recall $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}}^{\text{oracle}} - \mathbf{\Theta}^\star$. By (27), we have

$$(31)\qquad \widehat{\mathbf{\Sigma}}^{\text{oracle}} = (\mathbf{\Theta}^\star + \widehat{\mathbf{\Delta}})^{-1} = (\mathbf{I} + \mathbf{\Sigma}^\star\mathbf{\Delta})^{-1} \cdot \mathbf{\Sigma}^\star = \mathbf{\Sigma}^\star - \mathbf{\Sigma}^\star\widehat{\mathbf{\Delta}}\mathbf{\Sigma}^\star + \widehat{\mathbf{R}},$$

where $\widehat{\mathbf{R}} = (\boldsymbol{\Sigma}^\star\widehat{\boldsymbol{\Delta}})^2 \cdot \widehat{\mathbf{J}}\boldsymbol{\Sigma}^\star$ and $\widehat{\mathbf{J}}$ is defined similar to $\mathbf{J}$ with $\boldsymbol{\Delta}$ replaced by $\widehat{\boldsymbol{\Delta}}$. Then $\widehat{\mathbf{J}}$ is a convergent matrix series under the condition (25). In terms of $\mathcal{A}$, we can equivalently write (31) as

$$\begin{aligned}\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}^{\text{oracle}}_{\mathcal{A}}) - \operatorname{vec}(\boldsymbol{\Sigma}^\star_{\mathcal{A}}) &= -\mathbf{H}^\star_{\mathcal{A}\mathcal{A}} \cdot \operatorname{vec}(\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}) + \operatorname{vec}(\widehat{\mathbf{R}}_{\mathcal{A}}),\\ \operatorname{vec}(\widehat{\boldsymbol{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c}) - \operatorname{vec}(\boldsymbol{\Sigma}^\star_{\mathcal{A}^c}) &= -\mathbf{H}^\star_{\mathcal{A}^c\mathcal{A}} \cdot \operatorname{vec}(\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}) + \operatorname{vec}(\widehat{\mathbf{R}}_{\mathcal{A}^c}),\end{aligned}$$

where we use the fact that $\widehat{\boldsymbol{\Delta}}_{\mathcal{A}^c} = \mathbf{0}$. Solving $\operatorname{vec}(\widehat{\boldsymbol{\Delta}}_{\mathcal{A}})$ from the first equation and substituting it into the second equation, we obtain

$$\begin{aligned}&\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c}) - \operatorname{vec}(\boldsymbol{\Sigma}^\star_{\mathcal{A}^c})\\ &\quad= \mathbf{H}^\star_{\mathcal{A}^c\mathcal{A}}(\mathbf{H}^\star_{\mathcal{A}\mathcal{A}})^{-1} \cdot (\operatorname{vec}(\widehat{\boldsymbol{\Sigma}}^{\text{oracle}}_{\mathcal{A}}) - \operatorname{vec}(\boldsymbol{\Sigma}^\star_{\mathcal{A}}) - \operatorname{vec}(\widehat{\mathbf{R}}_{\mathcal{A}})) + \operatorname{vec}(\widehat{\mathbf{R}}_{\mathcal{A}^c}).\end{aligned}$$

Recall (29) holds under condition (25). Thus, we have

$$\|\widehat{\mathbf{R}}\|_{\max} \le \tfrac{3}{2}K_1^3 \cdot d\|\widehat{\Delta}\|^2_{\max} = 6K_1^3K_2^2 d \cdot \|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}}{}^\star\|_{\max} \le \|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}}{}^\star\|_{\max}.$$

Under the extra event $\{\|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}^c} - \boldsymbol{\Sigma}^\star_{\mathcal{A}^c}\|_{\max} < \frac{a_1\lambda}{2}\} \cap \{\|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}} - \boldsymbol{\Sigma}^\star_{\mathcal{A}}\|_{\max} \le \frac{a_1\lambda}{4K_3+2}\}$, we derive the desired upper bound for (30) by using the triangular inequality,

$$\begin{aligned}\|\nabla_{\mathcal{A}^c}\ell_n(\widehat{\boldsymbol{\Theta}}^{\text{oracle}})\|_{\max} &\le \|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}^c} - \boldsymbol{\Sigma}^\star_{\mathcal{A}^c}\|_{\max} + \|\boldsymbol{\Sigma}^\star_{\mathcal{A}^c} - \widehat{\boldsymbol{\Sigma}}^{\text{oracle}}_{\mathcal{A}^c}\|_{\max}\\ &\le \frac{a_1\lambda}{2} + (2K_3+1) \cdot \|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}} - \boldsymbol{\Sigma}_{\mathcal{A}}{}^\star\|_{\max}\\ &< a_1\lambda.\end{aligned}$$

Therefore,

$$\begin{aligned}\delta_1 \le \Pr\left\{\|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}} - \boldsymbol{\Sigma}^\star_{\mathcal{A}}\|_{\max} \ge \min\left\{\frac{1}{6K_1K_2d}, \frac{1}{6K_1^3K_2^2d}, \frac{a_1\lambda}{4K_3+2}\right\}\right\}\\ + \Pr\left\{\|\widehat{\boldsymbol{\Sigma}}^n_{\mathcal{A}^c} - \boldsymbol{\Sigma}^\star_{\mathcal{A}^c}\|_{\max} > \frac{a_1\lambda}{2}\right\}.\end{aligned}$$

An application of (8) yields $\delta_1^G$. This completes the proof of Theorem 6.

5.7. *Proof of Theorem 7.* First, we bound $\delta_2 = \Pr(\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}}\|_{\min} \le a\lambda)$. To this end, we let

$$\mathbb{B}(r) = \{\boldsymbol{\Delta} \in \mathbb{R}^p : \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_2} \le r, \boldsymbol{\Delta}_{\mathcal{A}^c} = \mathbf{0}\}$$

with $r = \|\boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\min} - a\lambda$, and $\partial\mathbb{B}(r) = \{\boldsymbol{\Delta} \in \mathbb{R}^p : \|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_2} = r, \boldsymbol{\Delta}_{\mathcal{A}^c} = \mathbf{0}\}$. We define $F(\boldsymbol{\Delta}) = \ell_n(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - \ell_n(\boldsymbol{\beta}^\star)$ and $\widehat{\boldsymbol{\Delta}} = \arg\min_{\boldsymbol{\Delta}:\,\boldsymbol{\Delta}_{\mathcal{A}^c}=\mathbf{0}} F(\boldsymbol{\Delta})$. Then $\widehat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}}^{\text{oracle}} - \boldsymbol{\beta}^\star$. Since $F(\widehat{\boldsymbol{\Delta}}) \le F(\mathbf{0}) = 0$ holds by definition, the convexity of $F(\boldsymbol{\Delta})$ yields that $\Pr(\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\|_{\ell_2} \le r) \ge \Pr(\inf_{\boldsymbol{\Delta}\in\partial\mathbb{B}(r)} F(\boldsymbol{\Delta}) > 0)$, and thus $\delta_2 \le 1 - \Pr(\|\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}\|_{\ell_2} \le r) \le 1 - \Pr(\inf_{\boldsymbol{\Delta}\in\partial\mathbb{B}(r)} F(\boldsymbol{\Delta}) > 0)$. In what follows, it suffices to bound $\Pr(\inf_{\boldsymbol{\Delta}\in\partial\mathbb{B}(r)} F(\boldsymbol{\Delta}) > 0)$.

By the definition of $\rho_\tau(\cdot)$ and $y_i = \mathbf{x}_i'\boldsymbol{\beta}^\star + \varepsilon_i$, we can rewrite $F(\boldsymbol{\Delta})$ as

$$\begin{aligned} F(\boldsymbol{\Delta}) &= \frac{1}{n}\sum_i \{\rho_\tau(y_i - \mathbf{x}_i'(\boldsymbol{\beta}^\star + \boldsymbol{\Delta})) - \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}^\star)\} \\ &= \frac{1}{n}\sum_i \{\rho_\tau(\varepsilon_i - \mathbf{x}_i'\boldsymbol{\Delta}) - \rho_\tau(\varepsilon_i)\} \\ &= \frac{1}{n}\sum_i \{\mathbf{x}_i'\boldsymbol{\Delta}\cdot(I_{\{\varepsilon_i\le 0\}} - \tau) + (\mathbf{x}_i'\boldsymbol{\Delta} - \varepsilon_i)\cdot(I_{\{\varepsilon_i \le \mathbf{x}_i'\boldsymbol{\Delta}\}} - I_{\{\varepsilon_i\le 0\}})\}. \end{aligned}$$

Next, we bound $I_1 = \frac{1}{n}\sum_i \mathbf{x}_i'\boldsymbol{\Delta}\cdot(I_{\{\varepsilon_i\le 0\}} - \tau)$ and $I_2 = F(\boldsymbol{\Delta}) - I_1$, respectively.

To bound $I_1$, we use the Cauchy–Schwarz inequality to obtain

$$|\mathbf{x}_i'\boldsymbol{\Delta}\cdot I_{\{\varepsilon_i\le 0\}}| \le |\mathbf{x}_i'\boldsymbol{\Delta}| = |\mathbf{x}_{i\mathcal{A}}\boldsymbol{\Delta}_{\mathcal{A}}| \le \|\mathbf{x}_{i\mathcal{A}}\|_{\ell_2}\cdot\|\boldsymbol{\Delta}_{\mathcal{A}}\|_{\ell_2} = M_{\mathcal{A}}^{1/2}s^{1/2}r.$$

Since $E[\mathbf{x}_i'\boldsymbol{\Delta}\cdot I_{\{\varepsilon_i\le 0\}}] = \mathbf{x}_i'\boldsymbol{\Delta}\cdot\Pr(\varepsilon_i\le 0) = \mathbf{x}_i'\boldsymbol{\Delta}\cdot\tau$, we can apply the Hoeffding's inequality to bound $I_1$ as follows:

$$\Pr\left(|I_1| > \frac{1}{6}\lambda_{\min}f_{\min}r^2\right) \le 2\exp\left(-\frac{\lambda_{\min}^2 f_{\min}^2}{72M_{\mathcal{A}}\cdot s}\cdot nr^2\right). \tag{32}$$

Now we bound $I_2$. Using Knight's identity [Knight (1998)], we write $I_2$ as

$$I_2 = \frac{1}{n}\sum_i (\mathbf{x}_i'\boldsymbol{\Delta} - \varepsilon_i)\cdot(I_{\{\varepsilon_i\le \mathbf{x}_i'\boldsymbol{\Delta}\}} - I_{\{\varepsilon_i\le 0\}}) = \frac{1}{n}\sum_i \int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} (I_{\{\varepsilon_i\le s\}} - I_{\{\varepsilon_i\le 0\}})\,ds.$$

Note that each term in the summation of $I_2$ is uniformly bounded, that is,

$$\left|\int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} (I_{\{\varepsilon_i\le s\}} - I_{\{\varepsilon_i\le 0\}})\,ds\right| \le \left|\int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} 1\,ds\right| \le |\mathbf{x}_i'\boldsymbol{\Delta}| \le M_{\mathcal{A}}{}^{1/2}s^{1/2}r.$$

Then we can use the Hoeffding's inequality to bound $I_2 - E[I_2]$ as

$$\Pr\left(|I_2 - E[I_2]| > \frac{1}{6}\lambda_{\min}f_{\min}r^2\right) \le 2\exp\left(-\frac{\lambda_{\min}^2 f_{\min}^2}{72M_{\mathcal{A}}\cdot s}\cdot nr^2\right). \tag{33}$$

Next, we apply Fubini's theorem and mean-value theorem to derive that

$$\begin{aligned} E[I_2] &= \frac{1}{n}\sum_i \int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} (F_i(s) - F_i(0))\,ds \\ &= \frac{1}{n}\sum_i \int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} f_i(\xi(s))\cdot s\,ds, \end{aligned}$$

where $\xi(s)$ is on the line segment between 0 and $s$. By the assumption of Theorem 7, $|\xi(s)| \le |\mathbf{x}_i'\boldsymbol{\Delta}| \le M_{\mathcal{A}}{}^{1/2}sr^{1/2} \le u_0$ holds, and then, by condition

(C3), we have $f_i(\xi(s)) \geq f_{\min}$ for any $i$. Using this fact, it is easy to obtain

$$E[I_2] \geq \frac{1}{n}\sum_i \int_0^{\mathbf{x}_i'\boldsymbol{\Delta}} f_{\min} \cdot s\, ds = \frac{1}{2n} f_{\min} \sum_i (\mathbf{x}_{i\mathcal{A}}\boldsymbol{\Delta}_{\mathcal{A}})^2 \geq \frac{1}{2}\lambda_{\min} f_{\min} r^2.$$

This together with (32) and (33) proves that under the event

$$\{|I_1| \leq \tfrac{1}{6}\lambda_{\min} f_{\min} r^2\} \cup \{|I_2 - E[I_2]| \leq \tfrac{1}{6}\lambda_{\min} f_{\min} r^2\},$$

we have $F(\boldsymbol{\Delta}) = I_1 + I_2 \geq -|I_1| + E[I_2] + (I_2 - E[I_2]) \geq \frac{1}{6}\lambda_{\min} f_{\min} r^2 > 0$. Therefore, an application of the union bound yields the desired bound $\delta_2^Q$.

In the sequel, we bound $\delta_1 = \Pr(\|\nabla_{\mathcal{A}^c}\ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})\|_{\max} \geq a_1\lambda)$. A translation of (3) implies the subgradient optimality condition $\nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = 0$ for $j \in \mathcal{A}$. For ease of notation, let $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{\text{oracle}}$ and $\mathcal{Z} = \{i : \hat{\varepsilon}_i = 0\}$. Now we can rewrite the subgradient $\nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})$ as

$$\begin{aligned}
\nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) &= \frac{1}{n}\sum_{i\notin\mathcal{Z}} x_{ij} \cdot (I_{\{\hat{\varepsilon}_i \leq 0\}} - \tau) - \frac{1}{n}\sum_{i\in\mathcal{Z}} x_{ij} \cdot \hat{z}_i \\
&= \frac{1}{n}\sum_{i} x_{ij} \cdot (I_{\{\hat{\varepsilon}_i \leq 0\}} - \tau) - \frac{1}{n}\sum_{i\in\mathcal{Z}} x_{ij} \cdot (\hat{z}_i + 1 - \tau) \\
&= I_{3j} + I_{4j},
\end{aligned}$$

where $\hat{z}_i \in [\tau - 1, \tau]$ $(i \in \mathcal{Z})$ satisfies the subgradient optimality condition. To bound $\nabla_j \ell_n(\hat{\boldsymbol{\beta}}^{\text{oracle}})$ for $j \in \mathcal{A}^c$, a key observation is that the quantile regression for $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ exactly interpolates $s$ observations, that is, $|\mathcal{Z}| = s$. Please see Section 2.2 of Koenker (2005) for more details. Then it is easy to derive

$$\max_{j\in\mathcal{A}^c} |I_{4j}| \leq \max_{j\in\mathcal{A}^c} \frac{1}{n}\sum_{i\in\mathcal{Z}} |x_{ij}| \cdot (\max\{1-\tau, \tau\} + 1 - \tau) \leq 2m_{\mathcal{A}^c} \cdot \frac{s}{n} \leq \frac{a_1}{4}\lambda.$$

Using this bound for $|I_{4j}|$, we can further bound $\delta_1$ as

$$\delta_1 \leq \Pr\Big(\max_{j\in\mathcal{A}^c} |I_{3j} + I_{4j}| \geq a_1\lambda\Big) \leq \Pr\Big(\max_{j\in\mathcal{A}^c} |I_{3j}| \geq \frac{3a_1}{4}\lambda\Big). \tag{34}$$

Now we only need to bound $\max_{j\in\mathcal{A}^c} |I_{3j}|$. Note that we rewrite $I_{3j}$ as

$$\begin{aligned}
I_{3j} &= \frac{1}{n}\sum_i x_{ij}((I_{\{\hat{\varepsilon}_i \leq 0\}} - I_{\{\varepsilon_i \leq 0\}}) - E[I_{\{\hat{\varepsilon}_i \leq 0\}} - I_{\{\varepsilon_i \leq 0\}}]) \\
&\quad + \frac{1}{n}\sum_i x_{ij} E[I_{\{\hat{\varepsilon}_i \leq 0\}} - I_{\{\varepsilon_i \leq 0\}}] + \frac{1}{n}\sum_i x_{ij}(I_{\{\varepsilon_i \leq 0\}} - \tau) \\
&= I_{3j.1} + I_{3j.2} + I_{3j.3}.
\end{aligned}$$

Next, we define $\varepsilon_i(\mathbf{t}) = y_i - \mathbf{x}'_{i\mathcal{A}}\mathbf{t}$. Then $\hat{\varepsilon}_i = \varepsilon_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}})$ holds by definition. We also introduce $\mathbb{B}_\star = \{\mathbf{t} \in \mathbb{R}^s : \|\mathbf{t} - \boldsymbol{\beta}_{\mathcal{A}}{}^\star\|_{\ell_2} \le r_1 = 6\sqrt{M_{\mathcal{A}} s \log n/n}\}$. As long as $\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}} \in \mathbb{B}_\star$ holds, due to the mean-value theorem, we have

$$\begin{aligned}
|I_{3j.2}| &\le \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} \left| \frac{1}{n} \sum_i x_{ij} E[I_{\{\varepsilon_i(\mathbf{t})\le 0\}} - I_{\{\varepsilon_i \le 0\}}] \right| \\
&\le \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} \frac{m_{\mathcal{A}^c}}{n} \sum_i |F_i(\mathbf{x}'_{i\mathcal{A}}(\mathbf{t} - \boldsymbol{\beta}^\star_{\mathcal{A}})) - F_i(0)| \\
&\le \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} \frac{m_{\mathcal{A}^c}}{n} \sum_i |f_i(\xi_i(\mathbf{t})) \cdot \mathbf{x}'_{i\mathcal{A}}(\mathbf{t} - \boldsymbol{\beta}^\star_{\mathcal{A}})|,
\end{aligned}$$

where $\xi_i(\mathbf{t})$ is on the line segment between 0 and $\mathbf{x}'_{i\mathcal{A}}(\mathbf{t} - \boldsymbol{\beta}^\star_{\mathcal{A}})$. Note that $\sup_{\mathbf{t}\in\mathbb{B}_\star(r)} |\mathbf{x}'_{i\mathcal{A}}(\mathbf{t} - \boldsymbol{\beta}^\star_{\mathcal{A}})| \le \|\mathbf{x}_{i\mathcal{A}}\|_{\ell_2} \cdot \|\mathbf{t} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\ell_2} \le M_{\mathcal{A}}{}^{1/2} s^{1/2} r_1 \le u_0$. Then $\sup_{\mathbf{t}\in\mathbb{B}_\star(r)} |f_i(\xi_i(\mathbf{t}))| \le f_{\max}$ holds by condition (C3). Thus, we have

$$\max_{j\in\mathcal{A}^c} |I_{3j.2}| \le \frac{m_{\mathcal{A}^c}}{n} \cdot n f_{\max} \cdot M_{\mathcal{A}}{}^{1/2} s^{1/2} r_1 \le \frac{a_1}{4}\lambda. \tag{35}$$

Let $\gamma_i(\mathbf{t}) = I_{\{\varepsilon_i(\mathbf{t})\le 0\}} - I_{\{\varepsilon_i\le 0\}} - E[I_{\{\varepsilon_i(\mathbf{t})\le 0\}} - I_{\{\varepsilon_i\le 0\}}]$ and $I_{3j.1}(\mathbf{t}) = \frac{1}{n}\sum_i x_{ij}\gamma_i(\mathbf{t})$. Again if $\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}} \in \mathbb{B}_\star$, we have $|I_{3j.1}| \le \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} |I_{3j.1}(\mathbf{t})|$. Together with this, we combine (34), (35) and the union bound to obtain

$$\begin{aligned}
\delta_1 \le{}& \Pr(\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}} \notin \mathbb{B}_\star) + \Pr\left( \max_{j\in\mathcal{A}^c} \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} |I_{3j.1}(\mathbf{t})| > \frac{a_1\lambda}{4} \right) \\
&+ \Pr\left( \max_{j\in\mathcal{A}^c} |I_{3j.3}| > \frac{a_1\lambda}{4} \right).
\end{aligned} \tag{36}$$

Note that $\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}} \in \mathbb{B}_\star$ holds under the event $\{\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\ell_2} \le r_1\}$. Then we can combine (32) and (33) to derive that

$$\Pr(\hat{\boldsymbol{\beta}}_{\mathcal{A}}{}^{\text{oracle}} \notin \mathbb{B}_\star) \le \Pr(\|\hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_{\ell_2} > r_1) = 4n^{-1/2}. \tag{37}$$

By the assumption of $\lambda$, we use Lemma A.3 of Wang, Wu and Li (2012) to obtain

$$\Pr\left( \sup_{\mathbf{t}\in\mathbb{B}_\star(r)} \left| \frac{1}{n}\sum_i x_{ij}\gamma_i(\mathbf{t}) \right| > \frac{a_1}{4}\lambda \right) \le C_1(p-s)\exp\left( -\frac{a_1 n\lambda}{104 m_{\mathcal{A}^c}} \right), \tag{38}$$

where $C_1 > 0$ is a fixed constant that does not depend on $n$, $p$, $s$, $m_{\mathcal{A}^c}$ and $M_{\mathcal{A}}$. Furthermore, we use the Hoeffding's inequality to bound $I_{3j.3}$ as

$$\Pr\left( \max_{j\in\mathcal{A}^c} |I_{3j.3}| > \frac{a_1}{4}\lambda \right) \le 2(p-s)\cdot\exp\left( -\frac{a_1^2 n\lambda^2}{32 m^2_{\mathcal{A}^c}} \right). \tag{39}$$

Therefore, we can combine (36), (37), (38) and (39) to obtain the desired probability bound $\delta_1^Q$ for $\delta_1$. This complete the proof of Theorem 7.

**Acknowledgements.** We thank the Editor, Associate Editor and referees for their helpful comments that improved an earlier version of this paper.

## SUPPLEMENTARY MATERIAL

**Supplement to "Strong oracle optimality of folded concave penalized estimation"** (DOI: 10.1214/13-AOS1198SUPP; .pdf). In this supplementary note, we give the complete proof of Theorem 5 and some comments on the simulation studies.

## REFERENCES


Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

Belloni, A. and Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120. MR3012402

Cai, T., Liu, W. and Luo, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$. *Ann. Statist.* **35** 2313–2351. MR2382644

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.* To appear.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

Fan, J., Xue, L. and Zou, H. (2014). Supplement to "Strong oracle optimality of folded concave penalized estimation." DOI:10.1214/13-AOS1198.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

Huang, J. and Zhang, C.-H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *J. Mach. Learn. Res.* **13** 1839–1864. MR2956344

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

KNIGHT, K. (1998). Limiting distributions for $L_1$ regression estimators under general conditions. *Ann. Statist.* **26** 755–770. MR1626024

KOENKER, R. (2005). *Quantile Regression*. Cambridge Univ. Press, Cambridge. MR2268657

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LI, Y. and ZHU, J. (2008). $L_1$-norm quantile regression. *J. Comput. Graph. Statist.* **17** 163–185. MR2424800

MAZUMDER, R., FRIEDMAN, J. H. and HASTIE, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. MR2894769

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

SAULIS, L. and STATULEVIČIUS, V. A. (1991). *Limit Theorems for Large Deviations*. Kluwer Academic, Dordrecht. MR1171883

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222. MR2949353

WU, Y. and LIU, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19** 801–817. MR2514189

XUE, L., ZOU, H. and CAI, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Ann. Statist.* **40** 1403–1429. MR3015030

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the $\ell_q$ loss in $\ell_r$ balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHANG, T. (2009). Some sharp performance bounds for least squares regression with $L_1$ regularization. *Ann. Statist.* **37** 2109–2144. MR2543687

ZHANG, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

ZHANG, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli* **19** 2277–2293. MR3160554

ZHANG, C.-H. and ZHANG, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.* **27** 576–593. MR3025135

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651

J. Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: jqfan@princeton.edu

L. Xue
Department of Statistics
Pennsylvania State University
University Park, Pennsylvania 16802
USA
E-mail: lzxue@psu.edu

H. Zou
School of Statistics
University of Minnesota
Minneapolis, Minnesota 55414
USA
E-mail: zouxx019@umn.edu