

## CALIBRATING NONCONVEX PENALIZED REGRESSION IN ULTRA-HIGH DIMENSION

BY LAN WANG<sup>1</sup>, YONGDAI KIM<sup>2</sup> AND RUNZE LI<sup>3</sup>

*University of Minnesota, Seoul National University and  
 Pennsylvania State University*

We investigate high-dimensional nonconvex penalized regression, where the number of covariates may grow at an exponential rate. Although recent asymptotic theory established that there exists a local minimum possessing the oracle property under general conditions, it is still largely an open problem how to identify the oracle estimator among potentially multiple local minima. There are two main obstacles: (1) due to the presence of multiple minima, the solution path is nonunique and is not guaranteed to contain the oracle estimator; (2) even if a solution path is known to contain the oracle estimator, the optimal tuning parameter depends on many unknown factors and is hard to estimate. To address these two challenging issues, we first prove that an easy-to-calculate calibrated CCCP algorithm produces a consistent solution path which contains the oracle estimator with probability approaching one. Furthermore, we propose a high-dimensional BIC criterion and show that it can be applied to the solution path to select the optimal tuning parameter which asymptotically identifies the oracle estimator. The theory for a general class of nonconvex penalties in the ultra-high dimensional setup is established when the random errors follow the sub-Gaussian distribution. Monte Carlo studies confirm that the calibrated CCCP algorithm combined with the proposed high-dimensional BIC has desirable performance in identifying the underlying sparsity pattern for high-dimensional data analysis.

**1. Introduction.** High-dimensional data, where the number of covariates  $p$  greatly exceeds the sample size  $n$ , arise frequently in modern applications in biology, chemometrics, economics, neuroscience and other scientific fields. To facilitate the analysis, it is often useful and reasonable to assume that only a small number of covariates are relevant for modeling the response variable. Under this sparsity assumption, a widely used approach for analyzing high-dimensional data

---

Received September 2012; revised June 2013.

<sup>1</sup>Supported in part by NSF Grant DMS-13-08960.

<sup>2</sup>Supported in part by National Research Foundation of Korea Grant number 20100012671 funded by the Korea government.

<sup>3</sup>Supported in part by National Natural Science Foundation of China, 11028103 and NIH Grants P50 DA10075, R21 DA024260, R01 CA168676 and R01 MH096711.

*MSC2010 subject classifications.* Primary 62J05; secondary 62J07.

*Key words and phrases.* High-dimensional regression, LASSO, MCP, SCAD, variable selection, penalized least squares.

is regularized or penalized regression. This approach estimates the unknown regression coefficients by solving the following penalized regression problem

$$(1.1) \quad \min_{\boldsymbol{\beta} \in \mathcal{R}^p} \left\{ (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\},$$

where  $\mathbf{y}$  is the vector of responses,  $\mathbf{X}$  is an  $n \times p$  matrix of covariates,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is the vector of unknown regression coefficients,  $\|\cdot\|$  denotes the  $L_2$  norm (Euclidean norm), and  $p_\lambda(\cdot)$  is a penalty function which depends on a tuning parameter  $\lambda > 0$ . Many commonly used variable selection procedures in the literature can be cast into the above framework, including the best subset selection,  $L_1$  penalized regression or Lasso [Tibshirani (1996)], Bridge regression [Frank and Friedman (1993)], SCAD [Fan and Li (2001)], MCP [Zhang (2010a)], among others.

The Lasso penalized regression is computationally attractive and enjoys great performance in prediction. However, it is known that Lasso requires rather stringent conditions on the design matrix to be variable selection consistent [Zou (2006), Zhao and Yu (2006)]. Focusing on identifying the unknown sparsity pattern, nonconvex penalized high-dimensional regression has recently received considerable attention. Fan and Li (2001) first systematically studied nonconvex penalized likelihood for fixed finite dimension  $p$ . In particular, they recommended the SCAD penalty which enjoys the oracle property for variable selection. That is, it can estimate the zero coefficients as exact zero with probability approaching one, and estimate the nonzero coefficients as efficiently as if the true sparsity pattern is known in advance. Fan and Peng (2004) extended these results by allowing  $p$  to grow with  $n$  at the rate  $p = o(n^{1/5})$  or  $p = o(n^{1/3})$ . For high dimensional nonconvex penalized regression with  $p \gg n$ , Kim, Choi and Oh (2008) proved that the oracle estimator itself is a local minimum of SCAD penalized least squares regression under very relaxed conditions; Zhang (2010a) proposed a minimax concave penalty (MCP) and devised a novel PLUS algorithm which when used together can achieve the oracle property under certain regularity conditions. Important insight has also been gained through the recent work on theoretical analysis of the global solution [Kim and Kwon (2012), Zhang and Zhang (2012)]. However, direct computation of the global solution to the nonconvex penalized regression is infeasible in high dimensional setting.

For practical data analysis, it is critical to find an easy-to-implement procedure which can find a local solution with satisfactory theoretical property even when the number of covariates greatly exceeds the sample size. Two challenging issues remain unsolved. One is the problem of multiple local minima; the other is the problem of optimal tuning parameter selection.

A direct consequence of the multiple local minima problem is that the solution path is not unique and is not guaranteed to contain the oracle estimator. This problem is due to the nature of the nonconvexity of the penalty. To understand it,

1 we note that the penalized objective function in (1.1) is nonconvex in  $\beta$  whenever 1  
2 the convexity of the least squares loss function does not dominate the concavity 2  
3 of the penalty part. In general, the occurrence of multiple minima is unavoidable 3  
4 unless strong assumptions are imposed on both the design matrix and the penalty 4  
5 function. The recent theory for SCAD penalized linear regression [Kim, Choi and 5  
6 Oh (2008)] and for general nonconcave penalized generalized linear models [Fan 6  
7 and Lv (2011)] indicates that one of the local minima enjoys the oracle property 7  
8 but it is still an unsolved problem how to identify the oracle estimator among mul- 8  
9 tiple minima when  $p \gg n$ . Popularly used algorithms generally only ensure the 9  
10 convergence to a local minimum, which is not necessarily the oracle estimator. 10  
11 Numerical evidence in Section 4 suggests that the local minima identified by some 11  
12 of the popular algorithms have a relatively low probability to recover the unknown 12  
13 sparsity pattern although it may have small estimation error. 13

14 Even if a solution path is known to contain the oracle estimator, identifying 14  
15 such a desirable estimator from the path is itself a challenging problem in ultra- 15  
16 high dimension. The main issue is to find the optimal tuning parameter which 16  
17 yields the oracle estimator. The theoretically optimal tuning parameter does not 17  
18 have an explicit representation and depends on unknown factors such as the vari- 18  
19 ance of the unobserved random noise. Cross-validation is commonly adopted in 19  
20 practice to select the tuning parameter but is observed to often result in overfitting. 20  
21 In the case of fixed  $p$ , Wang, Li and Tsai (2007) rigorously proved that gener- 21  
22 alized cross-validation leads to an overfitted model with a positive probability for 22  
23 SCAD-penalized regression. Effective BIC-type criterion for nonconvex penalized 23  
24 regression has been investigated in Wang, Li and Tsai (2007) and Zhang, Li and 24  
25 Tsai (2010) for fixed  $p$ ; and in Wang, Li and Leng (2009) for diverging  $p$  (but 25  
26  $p < n$ ). However, to the best of our knowledge, there is still no satisfactory tuning 26  
27 parameter selection procedure for nonconvex penalized regression in ultra-high 27  
28 dimension. 28

29 The above two main concerns motivate us to consider calibrating nonconvex 29  
30 penalized regression in ultra-high dimension with the goal to identify the oracle 30  
31 estimator with high probability. To achieve this, we first prove that a calibration 31  
32 of the CCCP algorithm [Kim, Choi and Oh (2008)] for nonconvex penalized re- 32  
33 gression produces a consistent solution path with probability approaching one in 33  
34 merely two steps under conditions much more relaxed than what would be required 34  
35 for the Lasso estimator to be model selection consistent. Furthermore, extending 35  
36 the recent work of Chen and Chen (2008) and Kim, Kwon and Choi (2012) for 36  
37 Bayesian information criterion (BIC) on high dimensional least squares regression, 37  
38 we propose a high-dimensional BIC for a nonconvex penalized solution path and 38  
39 prove its validity under more general conditions when  $p$  grows at an exponential 39  
40 rate. The recent independent work of Zhang (2010a, 2013) devised a multi-stage 40  
41 convex relaxation scheme and proved that for the capped  $L_1$  penalty the algorithm 41  
42 can find a consistent solution path with probability approaching one under cer- 42  
43 tain conditions. Despite the similar flavor shared with the algorithm proposed in 43

1 this paper, his algorithm takes multiple steps (which can be very large in practice 1  
 2 depending on the design condition) and the paper has not studied the problem of 2  
 3 tuning parameter selection. 3

4 To deepen our understanding of the nonconvex penalized regression, we also de- 4  
 5 rive an interesting auxiliary theoretical result of an upper bound on the  $L_2$  distance 5  
 6 between a sparse local solution of nonconvex penalized regression and the oracle 6  
 7 estimator. This result is new and insightful. It suggests that under general regular- 7  
 8 ity conditions a sparse local minimum can often have small estimation error even 8  
 9 though it may not be the oracle estimator. Overall, the theoretical results in this 9  
 10 paper fill in important gaps in the literature, thus substantially enlarge the scope of 10  
 11 applications of nonconvex penalized regression in ultra-high dimension. In Monte 11  
 12 Carlo studies, we demonstrate that the calibrated CCCP algorithm combined with 12  
 13 the proposed high-dimensional BIC is effective in identifying the underlying spar- 13  
 14 sity pattern. 14

15 The rest of the paper is organized as follows. In Section 2, we define the nota- 15  
 16 tion, review the CCCP algorithm and introduce the new methodology. In Section 3, 16  
 17 we establish that the proposed calibrated CCCP solution path contains the oracle 17  
 18 estimator with probability approaching one under general conditions, and that the 18  
 19 proposed high-dimensional BIC is able to select the optimal tuning parameter with 19  
 20 probability tending to one. In Section 4, we report numerical results from Monte 20  
 21 Carlo simulations and a real data example. In Section 5, we present an auxiliary 21  
 22 theoretical result which sheds light on the estimation accuracy of a local minimum 22  
 23 of nonconvex penalized regression if it is not the oracle estimator. The proofs are 23  
 24 given in Section 6. 24

## 25 2. Calibrated nonconvex penalized least squares method. 25

26 2.1. *Notation and setup.* Suppose that  $\{(Y_i, \mathbf{x}_i)\}_{i=1}^n$  is a random sample from 26  
 27 the linear regression model: 27

$$28 \quad (2.1) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \quad 28$$

29 where  $\mathbf{y} = (Y_1, \dots, Y_n)^T$ ,  $\mathbf{X}$  is the  $n \times p$  nonstochastic design matrix with the 29  
 30  $i$ th row  $\mathbf{x}_i^T$ ,  $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^T$  is the vector of unknown true parameters, and 30  
 31  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$  is a vector of independent and identically distributed random 31  
 32 errors. 32

33 We are interested in the case where  $p = p_n$  greatly exceeds the sample size  $n$ . 33  
 34 The vector of the true parameters  $\boldsymbol{\beta}^*$  is assumed to be sparse in the sense that the 34  
 35 majority of its components are exactly zero. Let  $A_0 = \{j : \beta_j^* \neq 0\}$  be the index 35  
 36 set of covariates with nonzero coefficients and let  $|A_0| = q$  denote the cardinal- 36  
 37 ity of  $A_0$ . We use  $d_* = \min\{|\beta_j^*| : \beta_j^* \neq 0\}$  to denote the minimal absolute value 37  
 38 of the nonzero coefficients. Without loss of generality, we may assume that the 38  
 39 first  $q$  components of  $\boldsymbol{\beta}^*$  are nonzero, thus we can write  $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \mathbf{0}^T)^T$ , where 39  
 40 41 42 43

$\mathbf{0}$  represents a zero vector of length  $p - q$ . The oracle estimator is defined as  $\widehat{\boldsymbol{\beta}}^{(o)} = (\widehat{\boldsymbol{\beta}}_1^{(o)T}, \mathbf{0}^T)^T$ , where  $\widehat{\boldsymbol{\beta}}_1^{(o)}$  is the least squares estimator fitted using only the covariates whose indices are in  $A_0$ .

To handle the high-dimensional covariates, we consider the penalized regression in (1.1). The penalty function  $p_\lambda(t)$  is assumed to be increasing and concave for  $t \in [0, +\infty)$  with a continuous derivative  $\dot{p}_\lambda(t)$  on  $(0, +\infty)$ . To induce sparsity of the penalized estimator, it is generally necessary for the penalty function to have a singularity at the origin, that is,  $\dot{p}_\lambda(0+) > 0$ . Without loss of generality, the penalty function can be standardized such that  $\dot{p}_\lambda(0+) = \lambda$ . Furthermore, it is required that

$$(2.2) \quad \dot{p}_\lambda(t) \leq \lambda \quad \forall 0 < t < a_0\lambda,$$

$$(2.3) \quad \dot{p}_\lambda(t) = 0 \quad \forall t > a_0\lambda$$

for some positive constant  $a_0$ . Condition (2.3) plays the key role of not over-penalizing large coefficients, thus alleviating the bias problem associated with Lasso.

The above class of penalty functions include the popularly used SCAD penalty and MCP. The SCAD penalty is defined by

$$(2.4) \quad \dot{p}_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a - 1)\lambda} I(t > \lambda) \right\}$$

for some  $a > 2$ , where the notation  $b_+$  stands for the positive part of  $b$ , that is,  $b_+ = bI(b > 0)$ . Fan and Li (2001) recommended to use  $a = 3.7$  from a Bayesian perspective. On the other hand, the MCP is defined by  $\dot{p}_\lambda(t) = a^{-1}(a\lambda - t)_+$  for some  $a > 0$  (as  $a \downarrow 1$ , it amounts to hard-thresholding, thus in the following we assume  $a > 1$ ).

Let  $\mathbf{x}_{(j)}$  be the  $j$ th column vector of  $\mathbf{X}$ . Without loss of generality, we assume that  $\mathbf{x}_{(j)}^T \mathbf{x}_{(j)} / n = 1$  for all  $j$ . Throughout this paper, the following notation is used. For an arbitrary index set  $A \subseteq \{1, 2, \dots, p\}$ ,  $\mathbf{X}_A$  denotes the  $n \times |A|$  submatrix of  $\mathbf{X}$  formed by those columns of  $\mathbf{X}$  whose indices are in  $A$ . For a vector  $\mathbf{v} = (v_1, \dots, v_p)'$ , we use  $\|\mathbf{v}\|$  to denote its  $L_2$  norm; on the other hand  $\|\mathbf{v}\|_0 = \#\{j : v_j \neq 0\}$  denotes the  $L_0$  norm,  $\|\mathbf{v}\|_1 = \sum_j |v_j|$  denotes the  $L_1$  norm and  $\|\mathbf{v}\|_\infty = \max_j |v_j|$  denotes the  $L_\infty$  norm. We use  $\mathbf{v}_A$  to represent the size- $|A|$  subvector of  $\mathbf{v}$  formed by the entries  $v_j$  with indices in  $A$ . For a symmetric matrix  $\mathbf{B}$ ,  $\lambda_{\min}(\mathbf{B})$  and  $\lambda_{\max}(\mathbf{B})$  stand for the smallest and largest eigenvalues of  $\mathbf{B}$ , respectively. Furthermore, we let

$$(2.5) \quad \xi_{\min}(m) = \min_{|B| \leq m, A_0 \subseteq B} \lambda_{\min}(n^{-1} \mathbf{X}_B^T \mathbf{X}_B).$$

Finally,  $p$ ,  $q$ ,  $\lambda$  and other related quantities are all allowed to depend on  $n$ , but we suppress such dependence for notational simplicity.

1     2.2. *The CCCP algorithm.* It is challenging to solve the penalized regression 1  
2 problem in (1.1) when the penalty function is nonconvex. Kim, Choi and Oh (2008) 2  
3 proposed a fast optimization algorithm called the SCAD–CCCP (CCCP stands for 3  
4 ConCave Convex procedure) algorithm for solving the SCAD-penalized regres- 4  
5 sion. The key idea is to update the solution with the minimizer of the tight con- 5  
6 vex upper bound of the objective function obtained at the current solution. What 6  
7 makes a fast algorithm practical relies on the possibility of decomposing the non- 7  
8 convex penalized least squares objective function as the sum of a convex func- 8  
9 tion and a concave function. To be specific, suppose we want to minimize an ob- 9  
10 jective function  $C(\boldsymbol{\beta})$  which has the representation  $C(\boldsymbol{\beta}) = C_{\text{vex}}(\boldsymbol{\beta}) + C_{\text{cav}}(\boldsymbol{\beta})$  10  
11 for a convex function  $C_{\text{vex}}(\boldsymbol{\beta})$  and a concave function  $C_{\text{cav}}(\boldsymbol{\beta})$ . Given a cur- 11  
12 rent solution  $\boldsymbol{\beta}^{(k)}$ , the tight convex upper bound of  $C(\boldsymbol{\beta})$  is given by  $Q(\boldsymbol{\beta}) =$  12  
13  $C_{\text{vex}}(\boldsymbol{\beta}) + \nabla C_{\text{cav}}(\boldsymbol{\beta}^{(k)})' \boldsymbol{\beta}$  where  $\nabla C_{\text{cav}}(\boldsymbol{\beta}) = \partial C_{\text{cav}}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ . We then update the 13  
14 solution by minimizing  $Q(\boldsymbol{\beta})$ . Since  $Q(\boldsymbol{\beta})$  is a convex function, it can be easily 14  
15 minimized. 15

16 For the penalized regression in (1.1), we consider a penalty function  $p_\lambda(|\beta_j|)$  16  
17 which has the decomposition 17

$$18 \quad (2.6) \quad p_\lambda(|\beta_j|) = J_\lambda(|\beta_j|) + \lambda|\beta_j|, \quad 18$$

19 where  $J_\lambda(|\beta_j|)$  is a differentiable concave function. For example, for the SCAD 19  
20 penalty, 20

$$21 \quad J_\lambda(|\beta_j|) = -\frac{\beta_j^2 - 2\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda \leq |\beta_j| \leq a\lambda) \quad 22$$

$$23 \quad + \left[ \frac{(a+1)\lambda^2}{2} - \lambda|\beta_j| \right] I(|\beta_j| > a\lambda), \quad 23$$

$$24 \quad 24$$

$$25 \quad 25$$

$$26 \quad 26$$

27 while for the MCP penalty, 27

$$28 \quad J_\lambda(|\beta_j|) = \frac{\beta_j^2}{2a} I(0 \leq |\beta_j| < a\lambda) + \left[ \frac{a\lambda^2}{2} - \lambda|\beta_j| \right] I(|\beta_j| \geq a\lambda). \quad 28$$

29 Hence, using the decomposition in (2.6), the penalized objective function in (1.1) 29  
30 can be rewritten as 30

$$31 \quad \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p J_\lambda(|\beta_j|) + \lambda \sum_{j=1}^p |\beta_j|, \quad 31$$

32 which is the sum of convex and concave functions. The CCCP algorithm is applied 32  
33 as follows. Given a current solution  $\boldsymbol{\beta}^{(k)}$ , the tight convex upper bound is 33  
34 34

$$35 \quad (2.7) \quad Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}, \lambda) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \nabla J_\lambda(|\beta_j^{(k)}|) \beta_j + \lambda \sum_{j=1}^p |\beta_j|. \quad 35$$

36 We then update the current solution by  $\boldsymbol{\beta}^{(k+1)} = \arg \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(k)}, \lambda)$ . 36  
37 37  
38 38  
39 39  
40 40  
41 41  
42 42  
43 43

1 An important property of the CCCP algorithm is that the objective function 1  
 2 always decreases after each iteration [Yuille and Rangarajan (2003), and Tao and 2  
 3 An (1997)], from which it can be deduced that the solution converges to a local 3  
 4 minimum. See, for example, Corollary 3.2 of Hunter and Li (2005). However, 4  
 5 there is no guarantee that the local minimum found is the oracle estimator itself 5  
 6 because there are multiple local minima and the solution of the CCCP algorithm 6  
 7 depends on the choice of the initial solution. 7

8  
 9 *2.3. Calibrated nonconvex penalized regression.* In this paper, we propose and 9  
 10 study a calibrated CCCP estimator. More specifically, we start with the initial value 10  
 11  $\boldsymbol{\beta}^{(0)} = \mathbf{0}$  and a tuning parameter  $\lambda > 0$  and let  $Q$  be the tight convex upper bound 11  
 12 defined in (2.7). The calibrated algorithm consists of the following two steps. 12

- 13 1. Let  $\widehat{\boldsymbol{\beta}}^{(1)}(\lambda) = \arg \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(0)}, \tau\lambda)$ , where the choice of  $\tau > 0$  will be dis- 13  
 14 cussed later. 14
- 15 2. Let  $\widehat{\boldsymbol{\beta}}(\lambda) = \arg \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}^{(1)}(\lambda), \lambda)$ . 15  
 16

17 When we consider a sequence of tuning parameter values, we obtain a solu- 17  
 18 tion path  $\{\widehat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$ . The calculation of the path is fast even for very high- 18  
 19 dimensional  $p$  as for each of the two steps a convex minimization problem is 19  
 20 solved. In step 1, a smaller tuning parameter  $\tau\lambda$  is adopted to increase the esti- 20  
 21 mation accuracy, see Section 3.1 for discussions on the practical choice of  $\tau$ . We 21  
 22 call a solution path “*path consistent*” if it contains the oracle estimator. In Sec- 22  
 23 tion 3.1, we will prove that the calibrated CCCP algorithm produces a consistent 23  
 24 solution path under rather weak conditions. 24

25 Given such a solution path, a critical question is how to tune the regularization 25  
 26 parameter  $\lambda$  in order to identify the oracle estimator. The performance of a penal- 26  
 27 ized regression estimator is known to heavily depend on the choice of the tuning 27  
 28 parameter. To further calibrate nonconvex penalized regression, we consider the 28  
 29 following high-dimensional BIC criterion (HBIC) to compare the estimators from 29  
 30 the above solution path: 30

$$(2.8) \quad \text{HBIC}(\lambda) = \log(\widehat{\sigma}_{\lambda}^2) + |M_{\lambda}| \frac{C_n \log(p)}{n},$$

31 where  $M_{\lambda} = \{j : \widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0\}$  is the model identified by  $\widehat{\boldsymbol{\beta}}(\lambda)$ ,  $|M_{\lambda}|$  denotes the 31  
 32 cardinality of  $M_{\lambda}$ , and  $\widehat{\sigma}_{\lambda}^2 = n^{-1} \text{SSE}_{\lambda}$  with  $\text{SSE}_{\lambda} = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\|^2$ . As we are 32  
 33 interested in the case where  $p$  greatly exceeds  $n$ , the penalty term also depends 33  
 34 on  $p$ ; and  $C_n$  is a sequence of numbers that diverges to  $\infty$ , which will be discussed 34  
 35 later. 35  
 36  
 37  
 38

39 We compare the value of the above HBIC criterion for  $\lambda \in \Lambda_n = \{\lambda : |M_{\lambda}| \leq 39  
 40 K_n\}$ , where  $K_n > q$  represents a rough estimate of an upper bound of the sparsity 40  
 41 of the model and is allowed to diverge to  $\infty$ . We select the tuning parameter 41

$$\widehat{\lambda} = \arg \min_{\lambda \in \Lambda_n} \text{HBIC}(\lambda).$$

The above criterion extends the recent works of [Chen and Chen \(2008\)](#) and [Kim, Kwon and Choi \(2012\)](#) on the high-dimensional BIC for the least squares regression to tuning parameter selection for nonconvex penalized regression. In Sections 3.1–3.3, we study asymptotic properties under conditions such as sub-Gaussian random errors, dimension of the covariates growing at the exponential rate and diverging  $K_n$ .

**3. Theoretical properties.** The main theory comprises two parts. We first show that under some general regularity conditions the calibrated CCCP algorithm yields a solution path with the “*path consistency*” property. We next verify that when the proposed high-dimensional BIC is applied to this solution path to choose the tuning parameter  $\lambda$ , with probability tending to one the resulted estimator is the oracle estimator itself.

To facilitate the presentation, we specify a set of regularity conditions.

(A1) There exists a positive constant  $C_1$  such that  $\lambda_{\min}(n^{-1}\mathbf{X}_{A_0}^T\mathbf{X}_{A_0}) \geq C_1$ .

(A2) The random errors  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d. mean zero sub-Gaussian random variables with a scale factor  $0 < \sigma < \infty$ , that is,  $E[\exp(t\varepsilon_i)] \leq e^{\sigma^2 t^2/2}, \forall t$ .

(A3) The penalty function  $p_\lambda(t)$  is assumed to be increasing and concave for  $t \in [0, +\infty)$  with a continuous derivative  $\dot{p}_\lambda(t)$  on  $(0, +\infty)$ . It admits a convex-concave decomposition as in (2.6) with  $J_\lambda(\cdot)$  satisfies:  $\nabla J_\lambda(|t|) = -\lambda \text{sign}(t)$  for  $|t| > a\lambda$ , where  $a > 1$  is a constant; and  $|\nabla J_\lambda(|t|)| \leq |t|$  for  $|t| \leq b\lambda$ , where  $b \leq a$  is a positive constant.

(A4) The design matrix  $\mathbf{X}$  satisfies:  $\gamma = \min_{\delta \neq 0, \|\delta_{A_0^c}\|_1 \leq 3\|\delta_{A_0}\|_1} \frac{\|\mathbf{X}\delta\|}{\sqrt{n}\|\delta_{A_0}\|} > 0$ .

(A5) Assume that  $\lambda = o(d_*)$  and  $\tau = o(1)$ , where  $d_*$  is defined on page 4,  $\lambda$  and  $\tau$  are the two parameters in the modified CCCP algorithm given in the first paragraph of Section 2.3.

REMARK 1. Condition (A1) concerns the true model and is a common assumption in the literature on high-dimensional regression. Condition (A2) implies that for a vector  $\mathbf{a} = (a_1, \dots, a_n)^T$ ,

$$(3.1) \quad P(|\mathbf{a}^T \boldsymbol{\varepsilon}| > t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2 \|\mathbf{a}\|^2}\right), \quad t \geq 0.$$

Condition (A3) is satisfied by popular nonconvex penalty functions such as SCAD and MCP. Note that the condition  $\nabla J_\lambda(|t|) = -\lambda \text{sign}(t)$  for  $|t| > a\lambda$  is equivalent to assuming that  $\dot{p}_\lambda(|t|) = 0, \forall |t| > a\lambda$ , that is, large coefficients are not penalized, which is exactly the motivation for nonconvex penalties. Condition (A4), which is given in [Bickel, Ritov and Tsybakov \(2009\)](#), ensures a desirable bound on the  $L_1$  estimation loss of the Lasso estimator. Note that the CCCP algorithm yields the Lasso estimator after the first iteration, so the asymptotic properties of the CCCP estimator is related to that of the Lasso estimator. Condition (A4) holds under the restricted eigenvalue condition which is known to be a relatively mild condition



1 on the design matrix for high-dimensional estimation. In particular, it is known to 1  
 2 hold in some examples where the covariates are highly dependent, and is much 2  
 3 weaker than the irrepresentable condition [Zhao and Yu (2006)] which is almost 3  
 4 necessary for Lasso to be model selection consistent. 4

5  
 6 3.1. *Property of the solution path.* We first state a useful lemma that charac- 6  
 7 terizes a nonasymptotic property of the oracle estimator in high dimension. The 7  
 8 result is an extension of that in Kim, Choi and Oh (2008) under the more general 8  
 9 sub-Gaussian random error condition. 9

10  
 11 LEMMA 3.1. For any given  $0 < b_1 < 1$  and  $0 < b_2 < 1$ , consider the events 11

$$12 \quad F_{n1} = \left\{ \max_{j \in A_0} |\widehat{\beta}_j^{(o)} - \beta_j^*| \leq b_1 \lambda \right\} \quad \text{and} \quad F_{n2} = \left\{ \max_{j \in A_0^c} |S_j(\widehat{\beta}^{(o)})| \leq b_2 \lambda \right\},$$

13  
 14 where  $S_j(\beta) = -n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\beta)$ . Then under conditions (A1) and (A2), 14

$$15 \quad P(F_{n1} \cap F_{n2}) \geq 1 - 2q \exp[-C_1 b_1^2 n \lambda^2 / (2\sigma^2)] - 2(p - q) \exp[-nb_2^2 \lambda^2 / (2\sigma^2)].$$

16  
 17 The proof of Lemma 3.1 is given in the online supplementary material [Wang, 18  
 19 Kim and Li (2013)]. 19

20 Theorem 3.2 below provides a nonasymptotic bound of the probability the solu- 20  
 21 tion path contains the oracle estimator. Under general conditions, this probability 21  
 22 tends to one. 22

23  
 24 THEOREM 3.2. (1) Assume that conditions (A1)–(A5) hold. If  $\tau \gamma^{-2} q = o(1)$ , 24  
 25 then for all  $n$  sufficiently large, 25

$$26 \quad P(\widehat{\beta}(\lambda) = \widehat{\beta}^{(o)}) \geq 1 - 8p \exp(-n\tau^2 \lambda^2 / (8\sigma^2)).$$

27  
 28 (2) Assume that conditions (A1)–(A5) hold. If  $n\tau^2 \lambda^2 \rightarrow \infty$ ,  $\log p = o(n\tau^2 \lambda^2)$  28  
 29 and  $\tau \gamma^{-2} q = o(1)$ , then 29

$$30 \quad P(\widehat{\beta}(\lambda) = \widehat{\beta}^{(o)}) \rightarrow 1$$

31  
 32 as  $n \rightarrow \infty$ . 32

33  
 34 REMARK 2. Meinshausen and Yu (2009) considered thresholding Lasso, 34  
 35 which has the oracle property under an incoherent design condition in the ultra- 35  
 36 high dimension. Zhou (2010) further proposed and investigated a multi-step 36  
 37 thresholding procedure which can accurately estimate the sparsity pattern under 37  
 38 the restricted eigenvalue condition of Bickel, Ritov and Tsybakov (2009). These 38  
 39 theoretical results are derived by assuming the initial Lasso is obtained using a 39  
 40 theoretical tuning parameter value, which depends on the unknown random noise 40  
 41 variance  $\sigma^2$ . Estimating  $\sigma^2$  is a difficult problem in high-dimensional setting, 41  
 42 particularly when the random noise is non-Gaussian. On the other hand, if the true 42  
 43 43

1 value of  $\sigma^2$  is known a priori, then it is possible to derive variable selection consistency 1  
 2 tency under somewhat more relaxed conditions on the design matrix than those in 2  
 3 the current paper. Adaptive Lasso, originally proposed by Zou (2006) for fixed di- 3  
 4 mension, was extended to high dimension by Huang, Ma and Zhang (2008) under 4  
 5 a rather strong mutual incoherence condition. Zhou, van de Geer and Bühlmann 5  
 6 (2009) derived the consistency of adaptive Lasso in high dimension under similar 6  
 7 conditions on  $X$ , but still requires complex conditions on  $s$  and  $d_*$ . Some favor- 7  
 8 able empirical performance of the multi-step thresholded Lasso versus the adaptive 8  
 9 Lasso was reported in Zhou (2010). A theoretical comparison of these two proce- 9  
 10 dures in high dimension was considered by van de Geer, Bühlmann and Zhou 10  
 11 (2011) and Chapter 7 of Bühlmann and van de Geer (2011). For both adaptive and 11  
 12 thresholded Lasso, if a covariate is deleted in the first step, it will be excluded from 12  
 13 the final selected model. Zhang (2010a) proved that selection consistency holds for 13  
 14 the MCP solution at the universal penalty level  $\sigma\sqrt{2\log p/n}$ . The LLA algorithm, 14  
 15 which Zou and Li (2008) originally proposed for fixed dimensional models, allevi- 15  
 16 ates this problem and has the potential to be extended to the ultra-high dimension 16  
 17 under conditions similar as those in this paper. Needless to say, the performances 17  
 18 of the above procedures all depend on the choice of tuning parameter. However, 18  
 19 the important issue of tuning parameter selection has not been addressed. 19  
 20

21 REMARK 3. We proved that the calibrated CCCP algorithm which involves 21  
 22 merely two iterations is guaranteed to yield a solution path that contains the or- 22  
 23 acle estimator with high probability under general conditions. To provide some 23  
 24 intuition on this theory, we first note that the first step of the algorithm yields the 24  
 25 Lasso estimator, albeit with a small penalty level  $\tau\lambda$ . If we denote the first step es- 25  
 26 timator by  $\widehat{\beta}_j^{(\text{Lasso})}(\tau\lambda)$ , then based on the optimization theory, the oracle property 26  
 27 is achieved when 27

$$\begin{aligned} & \min_{j \in A_0} |\widehat{\beta}_j^{(\text{Lasso})}(\tau\lambda)| \geq a\lambda > \lambda, \\ & \text{sign}(\widehat{\beta}_j^{(o)}) = \text{sign}(\beta_j^*), \quad j \in A_0, \\ & \max_{j \notin A_0} |\nabla J_\lambda(\widehat{\beta}_j^{(\text{Lasso})}(\tau\lambda))| + n^{-1} \|\mathbf{X}_{A_0^c}^T(\mathbf{Y} - \mathbf{X})\widehat{\beta}^{(o)}\|_\infty \leq \lambda. \end{aligned}$$

28  
 29  
 30  
 31  
 32  
 33  
 34  
 35 The proof of Theorem 3.2 relies on the following condition:

$$(3.2) \quad \|\widehat{\beta}^{(\text{Lasso})}(\tau\lambda) - \beta^*\|_\infty \leq \lambda/2, \quad \min_{\beta_j^* \neq 0} |\beta_j^*| > a\lambda + \lambda/2$$

36  
 37  
 38 for the given  $a > 1$ . The proof proceeds by bounding the first part of (3.2) us- 38  
 39 ing a result of Bickel, Ritov and Tsybakov (2009) via  $\|\widehat{\beta}^{(\text{Lasso})}(\tau\lambda) - \beta\|_\infty \leq$  39  
 40  $\|\widehat{\beta}^{(\text{Lasso})}(\tau\lambda) - \beta\|_2$ . In Section 3.3, we considered an alternative approach using 40  
 41 the recent result of Zhang and Zhang (2012), which leads to weaker requirement 41  
 42 on the minimal signal strength under slightly stronger assumptions on the design 42  
 43 43

1 matrix. We also noted that Theorem 3.2 holds for any  $a > 1$ , although in the nu- 1  
 2 merical studies we use the familiar  $a = 3.7$ . 2

3 How fast the probability that our estimator is equal to the oracle estimator ap- 3  
 4 proaches one depends on the sparsity level, the magnitude of the smallest signal, 4  
 5 the size of the tuning parameter and the condition of the design matrix. Corol- 5  
 6 lary 3.3 below confirms that the path-consistency can hold in ultra-high dimension. 6

7  
 8 **COROLLARY 3.3.** *Assume that conditions (A1)–(A4) hold. Suppose there are 8  
 9 two positive constants  $\gamma_0$  and  $K$  such that  $\gamma \geq \gamma_0 > 0$  and  $q < K$ . If  $d_* = O(n^{-c_1})$  9  
 10 for some  $c_1 \geq 0$  and  $p = O(\exp(n^{c_2}))$  for some  $c_2 > 0$ , then 10*

$$11 \quad P(\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}) \rightarrow 1, \quad 11$$

12 provided  $\lambda = O(n^{-c_3})$  for some  $c_3 > c_1$ ,  $\tau^2 n^{1-2c_3-c_2} \rightarrow \infty$  and  $\tau = o(1)$ . 12  
 13 13

14 The above corollary indicates that if the true model is very sparse (i.e.,  $q < K$ ) 14  
 15 and the design matrix behaves well (i.e.,  $\gamma \geq \gamma_0 > 0$ ), then we can take  $\tau$  to be 15  
 16 a sequence that converges to 0 slowly, for example,  $\tau = 1/\log n$ . On the other hand, 16  
 17 if one is concerned that the true model may not be very sparse ( $q \rightarrow \infty$ ) and the 17  
 18 design matrix may not behave very well ( $\gamma \rightarrow 0$ ), then an alternative choice is to 18  
 19 take  $\tau = \lambda$  which works also quite well in practice. The following corollary estab- 19  
 20 lishes that under some general conditions, the choice of  $\tau = \lambda$  yields a consistent 20  
 21 solution path under ultra high-dimensionality. 21  
 22 22

23 **COROLLARY 3.4.** *Assume that conditions (A1)–(A4) hold. If  $q = O(n^{c_1})$  for 23  
 24 some  $c_1 \geq 0$ ,  $d_* = O(n^{-c_2})$  for some  $c_2 \geq 0$ ,  $\gamma = O(n^{-c_3})$  for some  $c_3 \geq 0$ ,  $p =$  24  
 25  $O(\exp(n^{c_4}))$  for some  $0 < c_4 < 1$ ,  $\lambda = O(n^{-c_5})$  for some  $\max(c_2, c_1 + 2c_3) <$  25  
 26  $c_5 < (1 - c_4)/4$  and  $\tau = \lambda$ , then 26*

$$27 \quad P(\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}) \rightarrow 1. \quad 27$$

28  
 29 **3.2. Property of the high-dimensional BIC.** Theorem 3.5 below establishes 29  
 30 the effectiveness of the HBIC defined in (2.8) for selecting the oracle estimator 30  
 31 along a solution path of the calibrated CCCP. 31  
 32 32

33 **THEOREM 3.5 (Property of HBIC).** *Assume that the conditions of Theo- 33  
 34 rem 3.2(2) hold, and there exists a positive constant  $\kappa$  such that 34*

$$35 \quad (3.3) \quad \lim_{n \rightarrow \infty} \min_{A \not\supseteq A_0, |A| \leq K_n} \{n^{-1} \|(\mathbf{I}_n - \mathbf{P}_A) \mathbf{X}_{A_0} \boldsymbol{\beta}_{A_0}^*\|^2\} \geq \kappa, \quad 35$$

36 where  $\mathbf{I}_n$  denotes the  $n \times n$  identity matrix and  $\mathbf{P}_A$  denotes the projection matrix 36  
 37 onto the linear space spanned by the columns of  $\mathbf{X}_A$ . If  $C_n \rightarrow \infty$ ,  $qC_n \log(p) =$  37  
 38  $o(n)$  and  $K_n^2 \log(p) \log(n) = o(n)$ , then 38  
 39 39

$$40 \quad P(M_{\widehat{\lambda}} = A_0) \rightarrow 1 \quad 40$$

41 as  $n, p \rightarrow \infty$ . 41  
 42 42  
 43 43

1     REMARK 4. Condition (3.3) is an asymptotic model identifiability condition, 1  
 2 similar to that in [Chen and Chen \(2008\)](#). This condition states that if we consider 2  
 3 any model which contains at most  $K_n$  covariates, it cannot predict the response 3  
 4 variable as well as the true model does if it is not the true model. To give some 4  
 5 intuition of this condition, as in [Chen and Chen \(2008\)](#), one can show that for 5  
 6  $A \not\subseteq A_0$ , 6

$$7 \quad n^{-1} \|(\mathbf{I}_n - \mathbf{P}_A) \mathbf{X}_{A_0} \boldsymbol{\beta}_{A_0}^*\|^2 \geq \lambda_{\min}(n^{-1} \mathbf{X}_{A_0 \cup A}^T \mathbf{X}_{A_0 \cup A}) \|\boldsymbol{\beta}_{A_0 \cap A^c}^*\|^2 7$$

$$8 \quad \geq \lambda_{\min}(n^{-1} \mathbf{X}_{A_0 \cup A}^T \mathbf{X}_{A_0 \cup A}) \min_{\beta_j \neq 0} \beta_j^{*2}. 8$$

$$9 \quad 9$$

$$10 \quad 10$$

11 The theorem confirms that the BIC criterion for shrinkage parameter selection in- 11  
 12 vestigated in [Wang, Li and Tsai \(2007\)](#), [Wang, Li and Leng \(2009\)](#) and [Zhang, Li 12](#)  
 13 and [Tsai \(2010\)](#) can be modified and extended to ultra-high dimensionality. Care- 13  
 14 fully examining the proof, it is worth noting that the consistency of the HBIC only 14  
 15 requires a consistent solution path but does not rely on the particular method used 15  
 16 to construct the path. Hence, the proposed HBIC has the potential to be gener- 16  
 17 alized to other settings with ultra-high dimensionality. The sequence  $C_n$  should 17  
 18 diverge to  $\infty$  slowly, for example,  $C_n = \log(\log n)$ , which is used in our numerical 18  
 19 studies. 19

20 **3.3. Relaxing the conditions on the minimal signal.** Theorem 3.2, which is the 20  
 21 main result of the paper, implies that the oracle property of the calibrated CCCP 21  
 22 estimator requires the following lower bound on the magnitude of the smallest 22  
 23 nonzero regression coefficient 23

$$24 \quad (3.4) \quad d_* > \lambda > cq\sqrt{\log p/n}, 24$$

25 where  $a > b$  means  $\lim_{n \rightarrow \infty} a/b = \infty$ , and  $c$  is a constant that depends on the 25  
 26 design matrix  $\mathbf{X}$  and other unknown factors such as  $\sigma^2$ . When the true model di- 26  
 27 mension  $q$  is fixed, the lower bound for  $d_*$  is arbitrarily close to the optimal lower 27  
 28 bound  $c\sqrt{\log p/n}$  for nonconvex penalized approaches [e.g., [Zhang \(2010a\)](#)]. 28  
 29 However, when  $q$  is diverging, this bound is suboptimal. In general, there is a 29  
 30 tradeoff between the conditions on  $d_*$  and the conditions on the design matrix. 30  
 31 Comparing to the results in the literature, Theorem 3.2 imposes weak conditions 31  
 32 on the design matrix and the algorithm we investigate is transparent. In this section, 32  
 33 we will prove that the optimal lower bound of  $d_*$  can be achieved by the calibrated 33  
 34 CCCP procedure under a set of slightly stronger conditions on the design matrix. 34  
 35 35

36 Note that the calibrated CCCP estimator depends on  $\widehat{\boldsymbol{\beta}}^{(1)}$ , which is the 36  
 37 Lasso estimator obtained after the first iteration of the CCCP algorithm. In 37  
 38 fact, the lower bound of  $d_*$  is proportional to the  $l_\infty$  convergence rate of  $\widehat{\boldsymbol{\beta}}^{(1)}$  38  
 39 to  $\boldsymbol{\beta}^*$ , and condition (A4) only implies that  $\max_j |\widehat{\beta}_j^{(1)} - \beta_j^*|$  is proportional to 39  
 40  $O_p(q\sqrt{\log p/n/\tau})$ . If 40

$$41 \quad (3.5) \quad \max_j |\widehat{\beta}_j^{(1)} - \beta_j^*| = O_p(\sqrt{\log p/n/\tau}), 41$$

$$42 \quad 42$$

$$43 \quad 43$$

1 we can show that  $d_* > c\sqrt{\log p/n}/\tau$  for any  $\tau = o(1)$ , and hence we can achieve 1  
 2 almost the optimal lower bound for  $d_*$ . Now, the question is under what conditions 2  
 3 inequality (3.5) holds. Let  $v_{ij}$  be the  $(i, j)$  entry of  $\mathbf{X}^T \mathbf{X}$ . Lounici (2008) derived 3  
 4 the convergence rate (3.5) under the condition of mutual coherence: 4

$$(3.6) \quad \max_{i \neq j} |v_{ij}| > b/q$$

5  
 6  
 7 for some constant  $b > 0$ . However, the mutual coherence condition would be too 7  
 8 strong for practical purposes when  $q$  is diverging, since it requires that the pairwise 8  
 9 correlations between all possible pairs are sufficiently small. In this subsection, we 9  
 10 give an alternative condition for (3.5) based on the  $l_1$  operation norm of  $\mathbf{X}^T \mathbf{X}$ . 10  
 11

12 We replace condition (A4) with the slightly stronger condition (A4') below. 12  
 13 We also introduce an additional condition (A6) based on the matrix  $l_1$  operational 13  
 14 norm. For a given  $m \times m$  matrix  $\mathbf{A}$ , the  $l_1$  operational norm  $\|\mathbf{A}\|_1$  is defined by 14  
 15  $\|\mathbf{A}\|_1 = \max_{i=1, \dots, m} \sum_{j=1}^m |a_{ij}|$ , where  $a_{ij}$  is the  $(i, j)$ th entry of  $\mathbf{A}$ . Let 15

$$\zeta_{\max}(m) = \max_{|B| \leq m, A_0 \subset B} \|n^{-1} \mathbf{X}_B^T \mathbf{X}_B\|_1,$$

$$\zeta_{\min}(m) = \max_{|B| \leq m, A_0 \subset B} \|(n^{-1} \mathbf{X}_B^T \mathbf{X}_B)^{-1}\|_1.$$

16  
 17  
 18  
 19  
 20 Condition (A4'): There exist positive constants  $\alpha$  and  $\kappa_{\min}$  such that 20  
 21

$$(3.7) \quad \xi_{\min}((\alpha + 1)q) \geq \kappa_{\min}$$

22  
 23 and 23  
 24

$$(3.8) \quad \frac{\xi_{\max}(\alpha q)}{\alpha} \leq \frac{1}{576} \kappa_{\min} \left( 1 - 3 \sqrt{\frac{\xi_{\max}(\alpha q)}{\alpha \kappa_{\min}}} \right)^2,$$

25  
 26  
 27 where  $\xi_{\max}(m) = \max_{|B| \leq m, A_0 \subset B} \lambda_{\max}(n^{-1} \mathbf{X}_B^T \mathbf{X}_B)$ . 28

29 Condition (A6): Let  $u = \alpha + 1$ . There exist finite positive constants  $\eta_{\max}$  and 29  
 30  $\eta_{\min}$  such that 30

$$\limsup_{n \rightarrow \infty} \zeta_{\max}(uq) \leq \eta_{\max} < \infty$$

31  
 32 and 32  
 33

$$\limsup_{n \rightarrow \infty} \zeta_{\min}(uq) \leq \eta_{\min} < \infty.$$

34  
 35  
 36  
 37 REMARK 5. Similar conditions to condition (A4') were considered by 37  
 38 Meinshausen and Yu (2009) and Bickel, Ritov and Tsybakov (2009) for the  $l_2$  38  
 39 convergence of the Lasso estimator. However, (3.8) of condition (A4'), which es- 39  
 40 sentially assumes that  $\xi_{\max}(\alpha q)/\alpha$  is sufficiently small, is weaker, at least asymp- 40  
 41 totically, than the corresponding condition in Meinshausen and Yu (2009) and 41  
 42 Bickel, Ritov and Tsybakov (2009), which assumes that  $\xi_{\max}(q + \min\{n, p\})$  is 42  
 43 43

1 bounded. Zhang and Zhang (2012) proved that  $|\{j : \hat{\beta}_j \neq 0\} \cup A_0| \leq (\alpha + 1)q$  under 1  
 2 condition (A4'). In addition, condition (A4') implies condition (A4) [see Bickel, 2  
 3 Ritov and Tsybakov (2009)]. Condition (A6) is not too restrictive. Assume the  $\mathbf{x}_i$ 's 3  
 4 are randomly sampled from a distribution with mean  $\mathbf{0}$  and covariance matrix  $\Sigma$ . 4  
 5 If the  $l_1$  operational norm of  $\Sigma$  and  $\Sigma^{-1}$  are bounded, then we have  $\zeta_{\max}(uq) \leq$  5  
 6  $\max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B\|_1 + o_p(1)$  and  $\zeta_{\min}(uq) \leq \max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B^{-1}\|_1 + o_p(1)$  6  
 7 provided that  $q$  does not diverge too fast. Here  $\Sigma_B$  is the  $|B| \times |B|$  subma- 7  
 8 trix whose entries consist of  $\sigma_{jl}$ , the  $(j, l)$ th entry of  $\Sigma$ , for  $j \in B$  and  $l \in B$ . 8  
 9 See Proposition A.1 in the online supplementary material [Wang, Kim and Li 9  
 10 (2013)] of this paper. An example of  $\Sigma$  satisfying  $\max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B\|_1 < \infty$  10  
 11 and  $\max_{|B| \leq uq, A_0 \subset B} \|\Sigma_B^{-1}\|_1 < \infty$  is a block diagonal matrix where each block is 11  
 12 well posed and of finite dimension. Moreover, condition (A6) is almost necessary 12  
 13 for the  $l_\infty$  convergence of the Lasso estimator. Suppose that  $p$  is small and  $d_*$  is 13  
 14 large so that all coefficients of the Lasso coefficients are nonzero. Then, 14  
 15

$$16 \quad \hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{ls} + \tau \lambda (\mathbf{X}^T \mathbf{X} / n)^{-1} \boldsymbol{\delta},$$

17 where  $\hat{\boldsymbol{\beta}}^{ls}$  is the least square estimator, and  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)$  with  $\delta_j = \text{sign}(\hat{\beta}_j^{ls})$ . 17  
 18 Hence, for the sup norm between  $\hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{ls}$  to be the order of  $\tau \lambda$ , the  $l_1$  operational 18  
 19 norm of  $(\mathbf{X}^T \mathbf{X} / n)^{-1}$  should be bounded. 19  
 20

21 **THEOREM 3.6.** *Assume that conditions (A1)–(A3), (A4'), (A5) and (A6)* 21  
 22 *hold.* 22

23 (1) *If  $\tau = o(1)$ , then for all  $n$  sufficiently large,* 23  
 24

$$25 \quad P(\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}^{(o)}) \geq 1 - 8p \exp[-n\tau^2\lambda^2 / (8\sigma^2)].$$

26 (2) *If  $\tau = o(1)$  and  $\log p = o(n\tau^2\lambda^2)$ , then* 26  
 27

$$28 \quad P(\hat{\boldsymbol{\beta}}(\lambda) = \hat{\boldsymbol{\beta}}^{(o)}) \rightarrow 1$$

29 *as  $n \rightarrow \infty$ .* 29  
 30

31 (3) *Assume that the conditions of (2) and (3.3) hold. Let  $\hat{\lambda}$  be the tuning param-* 31  
 32 *eter selected by HBIC. If  $C_n \rightarrow \infty$ ,  $qC_n \log(p) = o(n)$ ,  $K_n^2 \log(p) \log(n) = o(n)$ ,* 32  
 33 *then  $P(M_{\hat{\lambda}} = A_0) \rightarrow 1$ , as  $n, p \rightarrow \infty$ .* 33  
 34

35 **REMARK 6.** We only need  $\tau = o(1)$  in Theorem 3.6 for the probability bound 35  
 36 of the calibrated CCCP estimator, while Theorem 3.2 requires  $\tau\gamma^{-2}q = o(1)$ . Under 36  
 37 the conditions of Theorem 3.6, the oracle property of  $\hat{\boldsymbol{\beta}}(\lambda)$  holds when 37  
 38

$$39 \quad (3.9) \quad d_* > \lambda > \frac{1}{\tau} \sqrt{\log p / n}.$$

40 Since  $\tau$  can converge to 0 arbitrarily slowly (e.g.,  $\tau = 1/\log n$ ), the lower bound 40  
 41 of  $d_*$  given by (3.9),  $\sqrt{\log p / n} / \tau$ , is almost optimal. 41  
 42  
 43

#### 4. Numerical results.

4.1. *Monte Carlo studies.* We now investigate the sparsity recovery and estimation properties of the proposed estimator via numerical simulations. We compare the following estimators: the oracle estimator which assumes the availability of the knowledge of the true underlying model; the Lasso estimator (implemented using the R package `glmnet`); the adaptive Lasso estimator [denoted by ALasso, Zou (2006), Section 2.8 of Bühlmann and van de Geer (2011)], the hard-thresholded Lasso estimator [denoted by HLasso, Section 2.8, Bühlmann and van de Geer (2011)], the SCAD estimator from the original CCCP algorithm without calibration (denoted by SCAD); the MCP estimator with  $a = 1.5$  and 3. For Lasso and SCAD, 5-fold cross-validation is used to select the tuning parameter; for ALasso, sequential tuning as described in Chapter 2 of Bühlmann and van de Geer (2011) is applied. For HLasso, following a referee's suggestion, we first used  $\lambda$  as the tuning parameter to obtain the initial Lasso estimator, then thresholded the Lasso estimator using thresholding parameter  $\eta = c\lambda$  for some  $c > 0$  and refitted least squares regression. We denote the solution path of HLasso by  $\hat{\beta}^{\text{HL}}(\lambda, c\lambda)$ , and apply HBIC to select  $\lambda$ . We consider  $c = 2$  and set  $C_n = \log \log n$  in the HBIC as it is found they lead to overall good performance for HLasso. The MCP estimator is computed using the R package PLUS with the theoretical optimal tuning parameter value  $\lambda = \sigma \sqrt{(2/n) \log p}$ , where the standard deviation  $\sigma$  is taken to be known. For the proposed calibrated CCCP estimator (denoted by New), we take  $\tau = 1/\log n$  and set  $C_n = \log \log n$  in the HBIC. We observe that the new estimator performs similarly if we take  $\tau = \lambda$ . In the following, we report simulation results from two examples. Results of additional simulations can be found in the online supplemental file.

EXAMPLE 1. We generate a random sample  $\{y_i, \mathbf{x}_i\}$ ,  $i = 1, \dots, 100$  from the following linear regression model:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \varepsilon_i,$$

where  $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, \mathbf{0}_{p-5}^T)^T$  with  $\mathbf{0}_k$  denoting a  $k$ -dimensional vector of zeros, the  $p$ -dimensional vector  $\mathbf{x}_i$  has the  $N(\mathbf{0}_p, \boldsymbol{\Sigma})$  distribution with covariance matrix  $\boldsymbol{\Sigma}$ ,  $\varepsilon_i$  is independent of  $\mathbf{x}_i$  and has a normal distribution with mean zero and standard deviation  $\sigma = 2$ . This simulation setup was considered in Fan and Li (2001) for a small  $p$  case. In this example, we consider  $p = 3000$  and the following choices of  $\boldsymbol{\Sigma}$ : (1) Case 1a: the  $(i, j)$ th entry of  $\boldsymbol{\Sigma}$  is equal to  $0.5^{|i-j|}$ ,  $1 \leq i, j \leq p$ ; (2) Case 1b: the  $(i, j)$ th entry of  $\boldsymbol{\Sigma}$  is equal to  $0.8^{|i-j|}$ ,  $1 \leq i, j \leq p$ ; (3) Case 1c: the  $(i, j)$ th entry of  $\boldsymbol{\Sigma}$  equal to 1 if  $i = j$  and 0.5 if  $1 \leq i \neq j \leq p$ .

EXAMPLE 2. We consider a more challenging case by modifying Example 1 case 1a. We divide the  $p$  components of  $\boldsymbol{\beta}^*$  into continuous blocks of size 20. We randomly select 10 blocks and assign each block the value  $(3, 1.5, 0, 0, 2, \mathbf{0}_{15}^T)/1.5$ .

Hence, the number of nonzero coefficients is 30. The entries in other blocks are set to be zero. We consider  $\sigma = 1$ . Two different cases are investigated: (1) Case 2a:  $n = 200$  and  $p = 3000$ ; (2) Case 2b:  $n = 300$  and  $p = 4000$ .

In the two examples, based on 100 simulation runs we report the average number of nonzero coefficients correctly estimated to be nonzero (i.e., true positive, denoted by TP) and average number of zero coefficients incorrectly estimated to be nonzero (i.e., false positive, denoted by FP) and the proportion of times the true model is exactly identified (denoted by TM). These three quantities describe the ability of various estimators for sparsity recovery. To measure the estimation accuracy, we report the mean squared error (MSE), which is defined to be  $100^{-1} \sum_{m=1}^{100} \|\widehat{\beta}^{(m)} - \beta^*\|^2$ , where  $\widehat{\beta}^{(m)}$  is the estimator from the  $m$ th simulation run.

The results are summarized in Tables 1 and 2. It is not surprising that Lasso always overfits. Other procedures improve the performance of Lasso by reducing the

TABLE 1

*Example 1. We report TP (the average number of nonzero coefficients correctly estimated to be nonzero, i.e., true positive), FP (average number of zero coefficients incorrectly estimated to be nonzero, i.e., false positive), TM (the proportion of the true model being exactly identified) and MSE*

Case	Method	TP	FP	TM	MSE
1a	Oracle	3.00	0.00	1.00	0.146
	Lasso	3.00	28.99	0.00	1.101
	ALasso	3.00	11.47	0.01	1.327
	HLasso	3.00	0.49	0.79	0.383
	SCAD	3.00	10.12	0.08	1.496
	MCP ( $a = 1.5$ )	2.89	0.28	0.76	0.561
	MCP ( $a = 3$ )	2.91	0.42	0.68	1.292
	New	2.99	<b>0.09</b>	<b>0.91</b>	<b>0.222</b>
1b	Oracle	3.00	0.00	1.00	0.314
	Lasso	3.00	20.64	0.00	1.248
	ALasso	3.00	8.84	0.02	1.527
	HLasso	2.79	0.50	0.56	1.244
	SCAD	2.99	7.42	0.17	1.598
	MCP ( $a = 1.5$ )	2.02	0.51	0.06	5.118
	MCP ( $a = 3$ )	1.99	0.60	0.02	5.437
	New	2.77	<b>0.21</b>	<b>0.66</b>	<b>1.150</b>
1c	Oracle	3.00	0.00	1.00	0.195
	Lasso	2.99	28.22	0.00	2.987
	ALasso	2.96	10.09	0.02	2.433
	HLasso	2.84	0.77	0.56	1.361
	SCAD	2.96	18.09	0.01	3.428
	MCP ( $a = 1.5$ )	2.67	<b>0.17</b>	<b>0.72</b>	1.636
	MCP ( $a = 3$ )	2.77	0.22	0.68	1.677
	New	2.79	0.46	0.58	<b>1.244</b>



## NONCONVEX PENALIZED REGRESSION

17

TABLE 2  
*Example 2. Captions are the same as those in Table 1*

Case	Method	TP	FP	TM	MSE
2a	Oracle	30.00	0.00	1.00	0.223
	Lasso	30.00	143.14	0.00	3.365
	ALasso	29.98	7.50	0.00	0.393
	HLasso	29.97	1.09	0.74	0.312
	SCAD	29.98	46.15	0.00	2.495
	MCP ( $a = 3$ )	29.83	0.50	<b>0.92</b>	0.807
	New	29.99	<b>0.20</b>	0.89	<b>0.247</b>
2b	Oracle	30.00	0.00	1.00	0.137
	Lasso	30.00	133.65	0.00	1.089
	ALasso	30.00	1.32	0.29	0.165
	HLasso	30.00	<b>0.00</b>	<b>1.00</b>	0.137
	SCAD	30.00	21.83	0.00	0.599
	MCP ( $a = 3$ )	30.00	0.08	0.92	0.137
	New	30.00	<b>0.00</b>	0.99	<b>0.135</b>

false positive rate. The SCAD estimator from the original CCCP algorithm without calibration has no guarantee to find a good local minimum and has low probability of identifying the true model. The best overall performance is achieved by the calibrated new estimator: the probability of identifying the true model is high and the MSE is relatively small. The HLasso (with thresholding parameter selected by our proposed HBIC) and MCP (using PLUS algorithm and the theoretically optimal tuning parameter) also have overall fine performance. We do not report the results of the MCP with  $a = 1.5$  for Example 2 since the PLUS algorithm sometimes runs into convergence problems.

*4.2. Real data analysis.* To demonstrate the application, we analyze the gene expression data set of Scheetz et al. (2006), which contains expression values of 31,042 probe sets on 120 twelve-week-old male offspring of rats. We are interested in identifying genes whose expressions are related to that of gene TRIM32 (known to be associated with human diseases of the retina) corresponding to probe 1389163\_at. We first preprocess the data as described in Huang, Ma and Zhang (2008) to exclude genes that are either not expressed or lacking sufficient variation. This leaves 18,957 genes.

For the analysis, we select 3000 genes that display the largest variance in expression level. We further analyze the top  $p$  ( $p = 1000$  and  $2000$ ) genes that have the largest absolute value of marginal correlation with gene TRIM32. We randomly partition the 120 rats into the training data set (80 rats) and testing data set (40 rats). We use the training data set to fit the model and select the tuning parameter; and use the testing data set to evaluate the prediction performance. We

TABLE 3  
*Gene expression data analysis. The results are based on 100 random partitions of the original data set*

$p$	Method	ave model size	Prediction error
1000	Lasso	31.17	<b>0.586</b>
	ALasso	11.76	0.646
	HLasso	12.04	0.676
	SCAD	4.81	0.827
	MCP ( $a = 1.5$ )	11.79	0.668
	MCP ( $a = 3$ )	7.02	0.768
	New	8.50	0.689
2000	Lasso	32.01	<b>0.604</b>
	ALasso	11.01	0.661
	HLasso	10.82	0.689
	SCAD	4.57	0.850
	MCP ( $a = 1.5$ )	11.33	0.700
	MCP ( $a = 3$ )	6.78	0.788
	New	7.91	0.736

perform 1000 random partitions and report in Table 3 the average model sizes and the average prediction error on the testing data set for  $p = 1000$  and 2000. For the MCP estimators, the tuning parameters are selected by cross-validation since the standard deviation of the random error is not known. We observe that the Lasso procedure yields the smallest prediction error. However, this is achieved by fitting substantially more complex models. The calibrated CCCP algorithm as well as ALasso and HLasso result in much sparser models with still small prediction errors. The performance of the MCP procedure is satisfactory but its optimal performance depends on the parameter  $a$ . In screening or diagnostic applications, it is often important to develop an accurate diagnostic test using as few features as possible in order to control the cost. The same consideration also matters when selecting target genes in gene therapies.

We also applied the calibrated CCCP procedure directly to the 18,957 genes and evaluated the predicative performance based on 100 random partitions. The calibrated CCCP estimator has an average model size 8.1 and an average prediction error 0.58. Note that the model size and predictive performance are similar to what we obtain when we first select 1000 (or 2000) genes with the largest variance and marginal correlation. This demonstrates the stability of the calibrated CCCP estimator in ultra-high dimension.

When a probe is simultaneously identified by different variable selection procedures, we consider it as evidence for the strength of the signal. Probe 1368113\_at is identified by both Lasso and the calibrated CCCP estimator. This probe corresponds to gene *tff2*, which was found to up-regulate cell proliferation in developing

1 mice retina [Paunel-Görgülü et al. (2011)]. On the other hand, the probes identi- 1  
 2 fied by the calibrated CCCP but not by Lasso also merit further investigation. For 2  
 3 instance, probe 1371168\_at was identified by the calibrated CCCP estimator but 3  
 4 not by Lasso. This probe corresponds to gene mpp2, which was found to be related 4  
 5 to protein metabolism abnormalities in the development of retinopathy in diabetic 5  
 6 mice [Gao et al. (2009)]. 6  
 7

8 *4.3. Extension to penalized logistic regression.* Regularized logistic regres- 8  
 9 sion is known to automatically result in a sparse set of features for classification in 9  
 10 ultra-high dimension [van de Geer (2008), Kwon and Kim (2013)]. We consider 10  
 11 the representative two-class classification problem, where the response variable  $y_i$  11  
 12 takes two possible values 0 or 1, indicating the class membership. It is assumed 12  
 13 that 13

$$(4.1) \quad P(y_i = 1 | \mathbf{x}_i) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}.$$

14 The penalized logistic regression estimator minimizes 14  
 15

$$n^{-1} \sum_{i=1}^n [-(\mathbf{x}_i^T \boldsymbol{\beta}) y_i + \log\{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}] + \sum_{j=1}^p p_\lambda(|\beta_j|).$$

16  
 17  
 18  
 19  
 20  
 21 When a nonconvex penalty is adopted, it is easy to see that the CCCP algorithm can 21  
 22 be extended to this case without difficulty as the penalized log-likelihood naturally 22  
 23 possesses the convex-concave decomposition discussed in Section 2.2 of the main 23  
 24 paper, because of the convexity of the negative log-likelihood for the exponential 24  
 25 family. For easy implementation, the CCCP algorithm can be combined with the 25  
 26 iteratively reweighted least squares algorithm for ordinary logistic regression, thus 26  
 27 taking advantage of the CCCP algorithm for linear regression. Denote the noncon- 27  
 28 vex penalized logistic regression estimator by  $\hat{\boldsymbol{\beta}}$ , then for a new feature vector  $\mathbf{x}$ , 28  
 29 the predicted class membership is  $I(\exp(\mathbf{x}^T \hat{\boldsymbol{\beta}}) / (1 + \exp(\mathbf{x}^T \hat{\boldsymbol{\beta}})) > 0.5)$ . 29  
 30

31 We demonstrate the performance of nonconvex penalized logistic regression 31  
 32 for classification through the following example: we generate  $\mathbf{x}_i$  as in Example 1 32  
 33 of the main paper, and the response variable  $y_i$  is generated according to (4.1) 33  
 34 with  $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, \mathbf{0}_{p-50}^T)^T$ . We consider sample size  $n = 300$  and feature 34  
 35 dimension  $p = 2000$ . Furthermore, an independent test set of size 1000 is used to 35  
 36 evaluate the misclassification error. The simulation results are reported in Table 4. 36  
 37 The results demonstrate that the calibrated CCCP estimator is effective in both 37  
 38 accurate classification and identifying the relevant features. 38

39 We expect that the theory we derived for the linear regression case continues to 39  
 40 hold for the logistic regression under similar conditions due to the convexity of the 40  
 41 negative log-likelihood function and the fact that the Bernoulli random variables 41  
 42 automatically satisfies the sub-Gaussian tail assumption. The latter is essential for 42  
 43 obtaining the exponential bounds in deriving the theory. 43

TABLE 4  
*Simulations for classification in high dimension ( $n = 300$ ,  $p = 2000$ )*

Method	TP	FP	TM	Misclassification rate
Oracle	3.00	0.00	1.00	0.116
Lasso	<b>3.00</b>	46.48	0.00	0.134
SCAD	2.08	4.02	0.04	0.161
ALASSO	2.02	4.58	0.00	0.188
HLASSO	2.87	<b>0.00</b>	0.87	0.120
MCP ( $a = 3$ )	2.96	0.56	0.54	0.128
New	2.99	<b>0.00</b>	<b>0.99</b>	<b>0.116</b>

**5. Revisiting local minima of nonconvex penalized regression.** In the following, we shall revisit the issue of multiple local minima of nonconvex penalized regression. We derive an  $L_2$  bound of the distance between a sparse local minimum and the oracle estimator. The result indicates that a local minimum which is sufficiently sparse often enjoys fairly accurate estimation even when it is not the oracle estimator. This result, to our knowledge, is new in the literature on high-dimensional nonconvex penalized regression.

Our theory applies the necessary condition for the local minimizer as in [Tao and An \(1997\)](#) for convex differencing problems. Let

$$Q_n(\boldsymbol{\beta}) = (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p p_\lambda(|\beta_j|)$$

and

$$\nabla(\boldsymbol{\beta}) = \{\boldsymbol{\xi} \in \mathcal{R}^p : \xi_j = -n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda l_j\},$$

where  $l_j = \text{sign}(\beta_j)$  if  $\beta_j \neq 0$  and  $l_j \in [-1, 1]$  otherwise,  $1 \leq j \leq p$ . As  $Q_n(\boldsymbol{\beta})$  can be expressed as the difference of two convex functions, a necessary condition for  $\boldsymbol{\beta}$  to be a local minimizer of  $Q_n(\boldsymbol{\beta})$  is

$$(5.1) \quad \frac{\partial h_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \in \nabla(\boldsymbol{\beta}),$$

where  $h_n(\boldsymbol{\beta}) = \sum_{j=1}^p J_\lambda(|\beta_j|)$ , where  $J_\lambda(|\beta_j|)$  is defined in Section 2.2 for SCAD and MCP penalty functions.

To facilitate our study, we introduce below a new concept.

**DEFINITION 5.1.** The relaxed sparse Riesz condition (SRC) in an  $L_0$ -neighborhood of the true model is satisfied for a positive integer  $m$  ( $2q \leq m \leq n$ ) if

$$\xi_{\min}(m) \geq c_* \quad \text{for some } 0 < c_* < \infty,$$

where  $\xi_{\min}$  is defined in (2.5).

1     REMARK 7. The *relaxed SRC condition* is related to, but generally weaker 1  
2 than the *sparse Reisz condition* [Zhang and Huang (2008), Zhang (2010a)], the *re-* 2  
3 *stricted eigenvalue condition* of Bickel, Ritov and Tsybakov (2009) and the *partial* 3  
4 *orthogonality condition* of Huang, Ma and Zhang (2008). 4

5     The theorem below unveils that for a given sparse estimator which is a local 5  
6 minimum of (1.1), its  $L_2$  distance to the oracle estimator  $\hat{\beta}^{(o)}$  has an upper bound, 6  
7 which is determined by three key factors: tuning parameter  $\lambda$ , the sparsity size of 7  
8 the local solution, and the magnitude of the smallest sparse eigenvalue as charac- 8  
9 terized by the relaxed SRC condition. To this end, we consider any local minimum 9  
10  $\hat{\beta} = (\hat{\beta}_j, \dots, \hat{\beta}_j)^T$  corresponding to the tuning parameter  $\lambda$ . Assume that the spar- 10  
11 sity size of this local solution satisfies:  $\|\hat{\beta}\|_0 \leq qu_n$  for some  $u_n > 0$ . 11  
12 12

13     THEOREM 5.2 (Properties of the local minima of nonconvex penalized re- 13  
14 gression). Consider SCAD or MCP penalized least squares regression. Assume 14  
15 that conditions (A1) and (A2) hold, and that the relaxed SRC condition in an 15  
16  $L_0$ -neighborhood of the true model is satisfied for  $m = qu_n^*$  where  $u_n^* = u_n + 1$ . 16  
17 Then if  $\lambda = o(d_*)$ , then for all  $n$  sufficiently large, 17  
18 18

$$\begin{aligned}
 & P \left\{ \|\hat{\beta}(\lambda) - \hat{\beta}^{(o)}\| \leq 2\lambda \sqrt{qu_n^* \xi_{\min}^{-1}(qu_n^*)} \right\} \\
 (5.2) \quad & \geq 1 - 2q \exp[-C_1 n (d_* - a\lambda)^2 / (2\sigma^2)] \\
 & \quad - 2(p - q) \exp[-n\lambda^2 / (2\sigma^2)],
 \end{aligned}$$

24 where  $\xi_{\min}(m)$  is defined in (2.5) and the positive constant  $C_1$  is defined in (A1). 24  
25 25

26     COROLLARY 5.3. Under the conditions of Theorem 5.2, if we take  $\lambda =$  26  
27  $\sqrt{3 \log(p)/n}$ , then we have 27  
28 28

$$\begin{aligned}
 & P \left\{ \|\hat{\beta}(\lambda) - \hat{\beta}^{(o)}\|^2 \leq 12 \frac{qu_n^* \log(p)}{n \xi_{\min}^2(qu_n^*)} \right\} \\
 & \geq 1 - 2q \exp[-C_1 n (d_* - a\lambda)^2 / (2\sigma^2)] - 2(p - q) \exp[-n\lambda^2 / (2\sigma^2)].
 \end{aligned}$$

33     The simple form in the above corollary suggests that if a local minimum is suf- 33  
34 ficiently sparse, in the sense that  $u_n$  diverge to  $\infty$  very slowly, this bound is never- 34  
35 theless quite tight as the rate  $q \log(p)/n$  is near-oracle. The factor  $u_n \xi_{\min}^{-2}(qu_n^*)$  is 35  
36 expected to go to infinity at a relatively slow rate if the local solution is sufficiently 36  
37 sparse. Our experience with existing algorithms for solving nonconvex penalized 37  
38 regression is that they often yield a sparse local minimum, which however has 38  
39 a low probability to be the oracle estimator itself. 39  
40 40

41     **6. Proofs.** We will provide here proofs for the main theoretical results in this 41  
42 paper. 42  
43 43

1 PROOF OF THEOREM 3.2. By definition,  $\widehat{\boldsymbol{\beta}}(\lambda) = \arg \min_{\boldsymbol{\beta}} Q_{\lambda}(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}^{(1)})$ , 1  
 2 where  $Q_{\lambda}(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}^{(1)}) = (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \nabla J_{\lambda}(|\widehat{\beta}_j^{(1)}|) \beta_j + \lambda \sum_{j=1}^p |\beta_j|$ . 2  
 3 Since  $Q_{\lambda}(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}^{(1)})$  is a convex function of  $\boldsymbol{\beta}$ , the KKT condition is necessary and 3  
 4 sufficient for characterizing the minimum. To verify that  $\widehat{\boldsymbol{\beta}}^{(o)}$  is the minimizer of 4  
 5  $Q_{\lambda}(\boldsymbol{\beta} \mid \widehat{\boldsymbol{\beta}}^{(1)})$ , it is sufficient to show that 5  
 6

$$7 \quad (6.1) \quad n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)}) + \nabla J_{\lambda}(|\widehat{\beta}_j^{(1)}|) + \lambda \operatorname{sign}(\widehat{\beta}_j^{(o)}) = 0, \quad j \in A_0 \quad 7$$

8 and 8

$$10 \quad (6.2) \quad |n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)}) + \nabla J_{\lambda}(|\widehat{\beta}_j^{(1)}|)| \leq \lambda, \quad j \notin A_0. \quad 10$$

11 We first verify (6.1). Note that with the initial value  $\mathbf{0}$ , we have  $\widehat{\boldsymbol{\beta}}^{(1)} =$  11  
 12  $\arg \min_{\boldsymbol{\beta}} \{(2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \tau \lambda \|\boldsymbol{\beta}\|_1\}$ . Let  $F_{n3} = \{\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_1 \leq 16\tau \lambda \gamma^{-2} q\}$ , 12  
 13 where  $\|\cdot\|_1$  denotes the  $L_1$  norm. By modifying the proof of Theorem 7.2 of 13  
 14 Bickel, Ritov and Tsybakov (2009), we can show that under the conditions of the 14  
 15 theorem, 15  
 16

$$17 \quad (6.3) \quad P(F_{n3}) \geq 1 - 2p \exp(-n\tau^2 \lambda^2 / (8\sigma^2)). \quad 17$$

18 By the assumption of the theorem, on the event  $F_{n3}$ ,  $\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*\|_1 \leq \lambda/2$  18  
 19 for all  $n$  sufficiently large. Furthermore, we consider the event  $F_{n1}$  defined 19  
 20 in Lemma 3.1 with  $b_1 = 1/2$ . By Lemma 3.1, we have  $P(\|\widehat{\boldsymbol{\beta}}^{(o)} - \boldsymbol{\beta}^*\|_{\infty} \leq$  20  
 21  $\lambda/2) \geq 1 - 2q \exp[-C_1 n \lambda^2 / (8\sigma^2)]$ . By the assumption  $\lambda = o(d_*)$ , for all  $n$  21  
 22 sufficiently large, on the event  $F_{n1} \cap F_{n3}$ , we have  $\operatorname{sign}(\widehat{\beta}_j^{(1)}) = \operatorname{sign}(\widehat{\beta}_j^{(o)})$ , for 22  
 23  $j \in A_0$  and  $\min_{j \in A_0} |\widehat{\beta}_j^{(1)}| > a\lambda$ . Hence, by condition (A3), on the event  $F_{n1} \cap$  23  
 24  $F_{n3}$ ,  $\nabla J_{\lambda}(|\widehat{\beta}_j^{(1)}|) = -\lambda \operatorname{sign}(\widehat{\beta}_j^{(1)}) = -\lambda \operatorname{sign}(\widehat{\beta}_j^{(o)})$ . Furthermore,  $n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} -$  24  
 25  $\mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)}) = 0$ , for  $j \in A_0$ , following the definition of the oracle estimator. Therefore, 25  
 26 (6.1) holds with probability at least  $1 - 2q \exp[-C_1 n \lambda^2 / (8\sigma^2)] - 2p \exp(-n\tau^2 \lambda^2 /$  26  
 27  $(8\sigma^2))$ . 27

28 Next, we verify (6.2). On the event  $F_{n3}$ , we have  $\max_{j \notin A_0} |\widehat{\beta}_j^{(1)}| \leq \lambda/2$ , for 28  
 29 all  $n$  sufficiently large. We consider the event  $F_{n2}$  defined in Lemma 3.1 with 29  
 30  $b_2 = 1/2$ . Lemma 3.1 implies that  $P(F_{n2}) \geq 1 - 2(p - q) \exp[-n\lambda^2 / (8\sigma^2)]$ . On 30  
 31 the event  $F_{n2}$  we have  $\max_{j \in A_0^c} |n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)})| \leq \lambda/2$ . By condition (A3), 31  
 32 on the event  $F_{n2} \cap F_{n3}$ , (6.2) holds, and this occurs with probability at least  $1 -$  32  
 33  $2(p - q) \exp[-n\lambda^2 / (8\sigma^2)] - 2p \exp(-n\tau^2 \lambda^2 / (8\sigma^2))$ . 33  
 34

35 The above two steps proves (1). The result in (2) follows immediately from (1). 35  
 36  $\square$  36

37 PROOF OF COROLLARIES 3.3 AND 3.4. The proof follows immediately from 37  
 38 Theorem 3.2.  $\square$  38

39 PROOF OF THEOREM 3.5. Recall that  $M_{\lambda} = \{j : \widehat{\beta}_j(\lambda) \neq 0\}$ . We define the 39  
 40 following three index sets:  $\Lambda_{n-} = \{\lambda > 0 : \lambda \in \Lambda_n, A_0 \not\subset M_{\lambda}\}$ ,  $\Lambda_{n0} = \{\lambda > 0 : \lambda \in$  40  
 41  $\Lambda_n, A_0 \subset M_{\lambda}\}$ ,  $\Lambda_{n+} = \{\lambda > 0 : \lambda \in \Lambda_n, A_0 \subset M_{\lambda}\}$ . 41  
 42 42  
 43 43

1  $\Lambda_n, A_0 = M_\lambda$ , and  $\Lambda_{n+} = \{\lambda > 0: \lambda \in \Lambda_n, A_0 \subset M_\lambda \text{ and } A_0 \neq M_\lambda\}$ . In other 1  
 2 words,  $\Lambda_{n-}, \Lambda_{n0}$  and  $\Lambda_{n+}$  denote the sets of  $\lambda$  values which lead to underfitted, ex- 2  
 3 actly fitted and overfitted models, respectively. For a given model (or equivalently 3  
 4 an index set)  $M$ , let  $\text{SSE}_M = \inf_{\beta_M \in \mathbb{R}^{|M|}} \|\mathbf{y} - \mathbf{X}_M \beta_M\|^2$ . That is,  $\text{SSE}_M$  is the sum 4  
 5 of squared residuals when the least squares method is used to estimate model  $M$ . 5  
 6 Also, let  $\hat{\sigma}_M^2 = n^{-1} \text{SSE}_M$ . From the definition, we always have  $\hat{\sigma}_\lambda^2 \geq \hat{\sigma}_{M_\lambda}^2$ . 6

7 Consider  $\lambda_n$  satisfying the conditions of Theorem 3.2(2). We have  $P(M_{\lambda_n} = 7$   
 8  $A_0) \rightarrow 1$ . We will prove that  $P(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0) \rightarrow 1$  and 8  
 9  $P(\inf_{\lambda \in \Lambda_{n+}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0) \rightarrow 1$ . 9

10 Case I. Consider an arbitrary  $\lambda \in \Lambda_{n-}$ , that is, the model corresponding to  $M_\lambda$  10  
 11 is underfitted. 11

$$\begin{aligned}
 & P\left(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0\right) \\
 &= P\left(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0, M_{\lambda_n} = A_0\right) \\
 &\quad + P\left(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0, M_{\lambda_n} \neq A_0\right) \\
 &\geq P\left(\inf_{\lambda \in \Lambda_{n-}} \left[\log(\hat{\sigma}_{M_\lambda}^2 / \hat{\sigma}_{A_0}^2) + (|M_\lambda| - q) \frac{C_n \log(p)}{n}\right] > 0\right) + o(1),
 \end{aligned}$$

12 where the inequality uses Theorem 3.2(2). Furthermore, we observe that 12  
 13

$$\log\left(\frac{\hat{\sigma}_{M_\lambda}^2}{\hat{\sigma}_{A_0}^2}\right) = \log\left(1 + \frac{n[\hat{\sigma}_{M_\lambda}^2 - \hat{\sigma}_{A_0}^2]}{\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}\right).$$

14 Applying the inequality  $\log(1 + x) \geq \min\{0.5x, \log(2)\}$ ,  $\forall x > 0$ , we have 14  
 15

$$\begin{aligned}
 & P\left(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0\right) \\
 &\geq P\left(\min\left\{\inf_{\lambda \in \Lambda_{n-}} \frac{n(\hat{\sigma}_{M_\lambda}^2 - \hat{\sigma}_{A_0}^2)}{2\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}, \log(2)\right\} - \frac{qC_n \log(p)}{n} > 0\right) + o(1).
 \end{aligned}$$

16 To evaluate  $\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}$ , we apply Corollary 1.3 of Mikosch (1990) with 16  
 17 their  $A_n = \mathbf{I}_n - \mathbf{P}_{A_0}$ ,  $B_n = 2\sigma^4(n - q)$ ,  $\mu_n = \sigma^2$  and  $y_n = (n - q)/(\log n)$ , we 17  
 18 have  $P(\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} \leq 2\sigma^2(n - q)) \rightarrow 1$  as  $n \rightarrow \infty$ . Thus 18  
 19

$$\begin{aligned}
 & P\left(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0\right) \\
 &\geq P\left(\min\left\{\frac{\inf_{\lambda \in \Lambda_{n-}} n(\hat{\sigma}_{M_\lambda}^2 - \hat{\sigma}_{A_0}^2)}{4(n - q)\sigma^2}, \log(2)\right\} - \frac{qC_n \log(p)}{n} > 0\right) + o(1).
 \end{aligned}$$

20 In what follows, we will prove that  $qC_n \log(p) = o(\inf_{\lambda \in \Lambda_{n-}} n(\hat{\sigma}_{M_\lambda}^2 - \hat{\sigma}_{A_0}^2))$ , 20  
 21 which combining with the assumption  $qC_n \log(p) = o(n)$  leads to the conclusion 21  
 22  $P(\inf_{\lambda \in \Lambda_{n-}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0) \rightarrow 1$ . 22  
 23

We have

$$\begin{aligned} n(\widehat{\sigma}_{M_\lambda}^2 - \widehat{\sigma}_{M_T}^2) &= \boldsymbol{\mu}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\mu} + 2\boldsymbol{\mu}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T \mathbf{P}_{M_\lambda} \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \mathbf{P}_{A_0} \boldsymbol{\varepsilon} \\ &= I_1 + I_2 - I_3 + I_4, \end{aligned}$$

where  $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}^*$ ,  $\mathbf{P}_{M_\lambda}$  is the projection matrix into the space spanned by the columns of  $\mathbf{X}_{M_\lambda}$ , and the definition of  $I_i$ ,  $i = 1, 2, 3, 4$ , should be clear from the context. Let  $M_- = \{j : j \notin M_\lambda, j \in M_T\}$ . Note that  $M_-$  is nonempty since  $M_\lambda$  underfits.

By assumption (3.3),  $|I_1| \geq \kappa n$ , for all  $n$  sufficiently large. To evaluate  $I_2$ , we have

$$I_2 = 2\sqrt{\boldsymbol{\mu}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\mu}} Z(M_\lambda) = 2\sqrt{I_1} Z(M_\lambda),$$

where  $Z(M_\lambda) = \mathbf{a}_n^T \boldsymbol{\varepsilon}$  with  $\mathbf{a}_n^T = (\boldsymbol{\mu}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\mu})^{-1/2} \boldsymbol{\mu}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda})$ . Note that  $\|\mathbf{a}_n\|^2 = 1$  and  $|\Lambda_-| \leq \sum_{t=0}^{K_n} \binom{p}{t} \leq \sum_{t=0}^{K_n} p^t = \frac{p^{K_n+1}-1}{p-1} \leq 2p^{K_n}$ . Applying the sub-Gaussian tail property in (3.1), we have

$$\begin{aligned} P\left(\sup_{\eta \in \Lambda_{n-}} |Z(M_\lambda)| > \sqrt{n/\log(n)}\right) &\leq 4p^{K_n} \exp(-n/(2\sigma^2 \log(n))) \\ &= 4 \exp(K_n \log(p) - n/(2\sigma^2 \log(n))) \rightarrow 0 \end{aligned}$$

as  $K_n \log(p) \log(n) = o(n)$ . Hence,  $\sup_{\eta \in \Lambda_{n-}} |I_2| = o(I_1)$ . To evaluate  $I_3$ , let  $r(\lambda) = \text{Trace}(\mathbf{P}_{M_\lambda})$ . It follows from Proposition 3 of Zhang (2010a) that for the sub-Gaussian random variables  $\varepsilon_i$ ,  $\forall t > 0$ ,

$$\begin{aligned} P\left\{\frac{\boldsymbol{\varepsilon}^T \mathbf{P}_{M_\lambda} \boldsymbol{\varepsilon}}{r(\lambda)\sigma^2} \geq \frac{1+t}{[1-2/(e^{t/2}\sqrt{1+t}-1)]_+^2}\right\} &\leq \exp\left(-\frac{r(\lambda)t}{2}\right) (1+t)^{(r(\lambda))/2}. \end{aligned} \tag{6.4}$$

We take  $t = n/(2\sigma^2 K_n \log(n)) - 1$  in the above inequality. Then  $t \rightarrow \infty$  by the assumptions of the theorem. Thus for all  $n$  sufficiently large,

$$\begin{aligned} P\left(\sup_{\lambda \in \Lambda_{n-}} |\boldsymbol{\varepsilon}^T \mathbf{P}_{M_\lambda} \boldsymbol{\varepsilon}| > \frac{n}{\log(n)}\right) &\leq P\left(\sup_{\lambda \in \Lambda_{n-}} \left|\frac{\boldsymbol{\varepsilon}^T \mathbf{P}_{M_\lambda} \boldsymbol{\varepsilon}}{r(\lambda)\sigma^2}\right| > \frac{n}{\sigma^2 K_n \log(n)}\right) \\ &\leq P\left(\sup_{\lambda \in \Lambda_{n-}} \left|\frac{\boldsymbol{\varepsilon}^T \mathbf{P}_{M_\lambda} \boldsymbol{\varepsilon}}{r(\lambda)\sigma^2}\right| > \frac{1+t}{[1-2/(e^{t/2}\sqrt{1+t}-1)]_+^2}\right) \end{aligned}$$



$$\begin{aligned}
&\leq 2p^{K_n} \exp(-n/(8\sigma^2 K_n \log(n)))(n/(2\sigma^2 K_n \log(n)))^{K_n/2} \\
&\leq 2 \exp(K_n \log(p) - n/(8\sigma^2 K_n \log(n)) + K_n \log(n/(2\sigma^2 K_n \log(n)))) \\
&\rightarrow 0,
\end{aligned}$$

since  $K_n^2 \log(p) \log(n) = o(n)$ . Finally,  $\boldsymbol{\varepsilon}^T \mathbf{P}_{A_0} \boldsymbol{\varepsilon}$  does not depend on  $\lambda$ . Similarly as above,  $P(\sup_{\lambda \in \Lambda_{n-}} |I_4| \geq n/\log(n)) \rightarrow 0$  by the sub-Gaussian tail condition.

Therefore, with probability approaching one,  $n(\hat{\sigma}_{M_\lambda}^2 - \hat{\sigma}_{A_0}^2)$  is dominated by  $I_1$ . This finishes the proof for the first case as  $qC_n \log(p) = o(n)$ .

Case II. Consider an arbitrary  $\lambda \in \Lambda_{n+}$ , that is, the model corresponding to  $M_\lambda$  is overfitted. In this case, we have  $\mathbf{y}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \mathbf{y} = \boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\varepsilon}$ . Therefore,  $n(\hat{\sigma}_{A_0}^2 - \hat{\sigma}_{M_\lambda}^2) = \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}$ . Let  $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}$ , then

$$\log\left(\frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_{M_\lambda}^2}\right) = \log\left(1 + \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T (\mathbf{I}_n - \mathbf{P}_{M_\lambda}) \boldsymbol{\varepsilon}}\right) \leq \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}$$

by the fact  $\log(1+x) \leq x, \forall x \geq 0$ .

Similarly as in case I,

$$\begin{aligned}
&P\left(\inf_{\lambda \in \Lambda_{n+}} [\text{HBIC}(\lambda) - \text{HBIC}(\lambda_n)] > 0\right) \\
&= P\left(\inf_{\lambda \in \Lambda_{n+}} \left[-\log\left(\frac{\hat{\sigma}_{A_0}^2}{\hat{\sigma}_{M_\lambda}^2}\right) + (|M_\lambda| - q) \frac{C_n \log(p)}{n}\right] > 0\right) + o(1) \\
&\geq P\left(\inf_{\lambda \in \Lambda_{n+}} \left[(|M_\lambda| - q) \frac{C_n \log(p)}{n} - \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}\right] > 0\right) \\
&\quad + o(1) \\
&= P\left(\inf_{\lambda \in \Lambda_{n+}} \left\{(|M_\lambda| - q) \left[\frac{C_n \log(p)}{n} - \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}\right]\right\}\right) \\
&\quad + o(1).
\end{aligned}$$

It suffices to show that

$$P\left(\inf_{\lambda \in \Lambda_{n+}} \left[\frac{C_n \log(p)}{n} - \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}\right] > 0\right) \rightarrow 1,$$

which is implied by

$$P\left(\frac{C_n \log(p)}{n} - \frac{\sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}}{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon}} > 0\right) \rightarrow 1.$$

Note that  $E(\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}) = \text{Var}(\varepsilon_i) \text{Trace}(\mathbf{I}_n - \mathbf{P}_{A_0}) \leq (n-q)\sigma^2$ , hence  $\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} = O_p(n)$ . Similarly as in case I, we can show that  $P(\sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0}) \boldsymbol{\varepsilon} > n/\log(n)) \rightarrow 0$ , since  $K_n^2 \log(p) \log(n) = o(n)$ . Thus,  $\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} - \sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} -$

1  $\mathbf{P}_{A_0})\boldsymbol{\varepsilon} = O_p(n)$ . Furthermore, applying (6.4) by letting  $t = 8 \log(p) - 1$ , we have 1  
 2 for all  $n$  sufficiently large, 2

$$\begin{aligned}
 & P\left(\sup_{\lambda \in \Lambda_{n+}} \frac{\boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}}{|M_\lambda| - q} > 16\sigma^2 \log(p)\right) \\
 & \leq \sum_{|M_\lambda|=q+1}^p \binom{p-q}{|M_\lambda|-q} \exp\left(-\frac{(|M_\lambda|-q)t}{2}\right) (1+t)^{(|M_\lambda|-q)/2} \\
 & = \sum_{k=1}^{p-q} \binom{p-q}{k} \exp(-2k \log(p)) (8 \log(p))^{k/2} \\
 & = \sum_{k=1}^{p-q} \binom{p-q}{k} \left(\frac{\sqrt{8 \log(p)}}{p_n^2}\right)^k \leq \left(1 + \frac{\sqrt{8 \log(p)}}{p^2}\right)^{p-q} - 1 \rightarrow 0.
 \end{aligned}$$

16 Thus with probability approaching one, for all  $n$  sufficiently large, 16

$$\begin{aligned}
 & \frac{C_n \log(p)}{n} - \frac{\sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0})\boldsymbol{\varepsilon} / (|M_\lambda| - q)}{\widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}} - \sup_{\lambda \in \Lambda_{n+}} \boldsymbol{\varepsilon}^T (\mathbf{P}_{M_\lambda} - \mathbf{P}_{A_0})\boldsymbol{\varepsilon}} \\
 & > n^{-1} C_n \log(p) - n^{-1} O(\log(p)) > 0,
 \end{aligned}$$

22 since  $C_n \rightarrow \infty$ . This finishes the proof.  $\square$  22

24 **PROOF OF THEOREM 3.6.** We will first prove that there exists a constant 24  
 25  $C > 0$  such that for  $F_{n4} = \{\max_j |\widehat{\beta}_j^{(1)} - \beta_j^*| \leq C\tau\lambda\}$ , we have 25

$$(6.5) \quad P(F_{n4}) \geq 1 - 2p \exp\left(\frac{-n\tau^2\lambda^2}{8\sigma^2}\right).$$

30 Let  $F_{n5} = \{|S_j(\boldsymbol{\beta}^*)| \leq \tau\lambda/2 \text{ for all } j\}$ . Since 30

$$P(F_{n5}^c) \leq \sum_{j=1}^p P(|\mathbf{x}_{(j)}^T \boldsymbol{\varepsilon} / n| > \tau\lambda/2) \leq 2p \exp\left(\frac{-n\tau^2\lambda^2}{8\sigma^2}\right),$$

35 we have 35

$$P(F_{n5}) \geq 1 - 2p \exp\left(\frac{-n\tau^2\lambda^2}{8\sigma^2}\right).$$

39 Hence to prove (6.5), it suffices to show that  $F_{n5} \subset F_{n4}$ . 39

40 Let 40

$$\theta = \inf \left\{ \frac{q \|\mathbf{X}^T \mathbf{X} \mathbf{u}\|_\infty}{n \|\mathbf{u}\|_1} : \|\mathbf{u}_{A_0^c}\|_1 \leq 3 \|\mathbf{u}_{A_0}\|_1 \right\}.$$

43

1 Corollary 2 of Zhang and Zhang (2012) proves that on the event  $F_{n5}$ ,  $|A \cup A_0| \leq$  1  
 2  $(\alpha + 1)q$ , where  $A = \{j : \widehat{\beta}_j^{(1)} \neq 0\}$ , provided 2  
 3

$$4 \frac{\xi_{\max}(\alpha q)}{\alpha} \leq \frac{1}{36} \theta. \quad 4$$

5 Since  $\theta \geq \gamma^2/16$  [see (7) of Zhang and Zhang (2012)], where  $\gamma$  is defined in (A4) 5  
 6 and 6  
 7

$$8 \gamma \geq \sqrt{\kappa_{\min}} \left( 1 - 3 \sqrt{\frac{\xi_{\max}(\alpha q)}{\alpha \kappa_{\min}}} \right) \quad 8$$

9 [see Bickel, Ritov and Tsybakov (2009)], condition (A4') implies that 9  
 10

$$11 (6.6) \quad F_{n5} \subset \{|A \cup A_0| \leq (\alpha + 1)q\}. \quad 11$$

12 Let  $C(\boldsymbol{\beta}) = (2n)^{-1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \tau\lambda \sum_{j=1}^p |\beta_j|$ . Then we have 12  
 13

$$14 \begin{aligned} C(\boldsymbol{\beta}) - C(\boldsymbol{\beta}^*) &= \sum_{j=1}^p (\beta_j - \beta_j^*) S_j(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) / (2n) \\ &\quad + \tau\lambda \sum_{j=1}^p (|\beta_j| - |\beta_j^*|). \end{aligned} \quad 14$$

15 Let  $\widehat{\mathbf{X}}\boldsymbol{\beta}^*$  be the projection of  $\mathbf{X}\boldsymbol{\beta}^*$  onto  $\text{span}(\mathbf{X}_A)$ , the linear subspace spanned 15  
 16 by the column vectors of  $\mathbf{X}_A$ . We define the  $p$ -dimensional vector  $\boldsymbol{\gamma}^*$  such that 16  
 17  $\widehat{\mathbf{X}}\boldsymbol{\beta}^* = \mathbf{X}_A \boldsymbol{\gamma}_A^*$  and  $\gamma_j^* = 0$  for  $j \in A^c$ . We have 17  
 18

$$19 \begin{aligned} &(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*)^T \mathbf{X}^T \mathbf{X} (\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^*) \\ &= (\widehat{\boldsymbol{\beta}}_A^{(1)} - \boldsymbol{\gamma}_A^*)^T \mathbf{X}_A^T \mathbf{X}_A (\widehat{\boldsymbol{\beta}}_A^{(1)} - \boldsymbol{\gamma}_A^*) + \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}_A \boldsymbol{\gamma}_A^*\|^2. \end{aligned} \quad 19$$

20 Therefore, we can write 20  
 21

$$22 \begin{aligned} \widehat{\boldsymbol{\beta}}^{(1)} &= \arg \min_{\boldsymbol{\beta} : \boldsymbol{\beta}_{A^c} = \mathbf{0}} \left\{ \sum_{j \in A} \beta_j S_j(\boldsymbol{\beta}^*) \right. \\ &\quad \left. + (\boldsymbol{\beta}_A - \boldsymbol{\gamma}_A^*)^T \mathbf{X}_A^T \mathbf{X}_A (\boldsymbol{\beta}_A - \boldsymbol{\gamma}_A^*) / 2n + \tau\lambda \sum_{j \in A} |\beta_j| \right\}. \end{aligned} \quad 22$$

23 Hence  $\widehat{\boldsymbol{\beta}}_A^{(1)} - \boldsymbol{\gamma}_A^* = (\mathbf{X}_A^T \mathbf{X}_A / n)^{-1} \boldsymbol{\theta}_A$ , where  $\boldsymbol{\theta} \in R^p$  such that  $\theta_j = 0$  for  $j \in A^c$  23  
 24 and  $\theta_j = -S_j(\boldsymbol{\beta}^*) - \text{sign}(\widehat{\beta}_j) \tau\lambda$  for  $j \in A$ . On  $F_{n5}$ ,  $\max_j |\theta_j| \leq 3\tau\lambda/2$ . Therefore, 24  
 25 condition (A6) with (6.6) implies that on the event  $F_{n5}$ , 25  
 26

$$27 (6.7) \quad \max_{j \in A} |\widehat{\beta}_j^{(1)} - \gamma_j^*| \leq \eta_{\min} 3\tau\lambda/2. \quad 27$$

28 It follows from (6.7) that inequality (6.5) holds if we show that  $A_0 \subset A$ , in which 28  
 29 case  $\boldsymbol{\gamma}_A^* = \boldsymbol{\beta}_A^*$ . We will prove this by contradiction. Assume  $A^{(-)} = A_0 \cap A^c$  is 29  
 30

1 nonempty. Let  $\widehat{\mathbf{x}}_{(j)}$  be the projection of  $\mathbf{x}_{(j)}$  onto  $\text{span}(\mathbf{X}_A)$  and let  $\tilde{\mathbf{x}}_{(j)} = \mathbf{x}_{(j)} -$  1  
 2  $\widehat{\mathbf{x}}_{(j)}$ ,  $j \in A^{(-)}$ . Then, we can write 2

$$3 \quad \mathbf{X}\boldsymbol{\beta}^* = \mathbf{X}_A\boldsymbol{\gamma}_A^* + \sum_{j \in A^-} \tilde{\mathbf{x}}_{(j)}\beta_j^*. \quad 3$$

4 Let  $\tilde{\mathbf{y}} = \sum_{j \in A^-} \tilde{\mathbf{x}}_{(j)}\beta_j^*$ . By Lemma 6.1 below, there exists  $l \in A^-$  such that 4  
 5

$$6 \quad (6.8) \quad |\mathbf{x}_{(l)}^T \tilde{\mathbf{y}}/n| \geq \kappa_{\min} d_*. \quad 6$$

7 By the KKT condition, we have  $|\mathbf{x}_{(l)}^T (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(1)})/n + S_l(\boldsymbol{\beta}^*)| \leq \tau\lambda$ . However 7  
 8 we can write  $\mathbf{x}_{(l)}^T (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(1)})/n = \mathbf{x}_{(l)}^T \mathbf{X}_A(\boldsymbol{\gamma}_A^* - \widehat{\boldsymbol{\beta}}_A^{(1)})/n + \mathbf{x}_{(l)}^T \tilde{\mathbf{y}}/n$ . The inequal- 8  
 9 ities (6.8) and (6.7) with condition (A6) imply that on  $F_{n5}$  9

$$10 \quad \begin{aligned} & |\mathbf{x}_{(l)}^T (\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(1)})/n + S_l(\boldsymbol{\beta}^*)| \\ & \geq |\mathbf{x}_{(l)}^T \tilde{\mathbf{y}}/n| - |\mathbf{x}_{(l)}^T \mathbf{X}_A(\boldsymbol{\gamma}_A^* - \widehat{\boldsymbol{\beta}}_A^{(1)})/n| - |S_l(\boldsymbol{\beta}^*)| \\ & \geq |\mathbf{x}_{(l)}^T \tilde{\mathbf{y}}/n| - \|\mathbf{X}_{A \cup A_0}^T \mathbf{X}_{A \cup A_0}\|_1 \|\boldsymbol{\gamma}_A^* - \widehat{\boldsymbol{\beta}}_A^{(1)}\|_\infty - |S_l(\boldsymbol{\beta}^*)| \\ & \geq \kappa_{\min} d_* - \eta_{\max} \eta_{\min} 3\tau\lambda/2 - \tau\lambda/2 > \tau\lambda \end{aligned} \quad 10$$

11 if  $d_* > 3\tau\lambda(\eta_{\max}\eta_{\min} + 1)/(2\kappa_{\min})$ , which contradicts the KKT condition. Hence, 11  
 12 we eventually have  $A_0 \subset A$  on  $F_{n5}$  and this proves (6.5). 12

13 We now slightly modify the proof of (1) of Theorem 3.2. More specifically, 13  
 14 replacing  $F_{n3}$  by  $F_{n4}$ , we can show that  $F_{n1} \cap F_{n2} \cap F_{n4} \subset \{\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}\}$ , and this 14  
 15 proves (1). The result in (2) follows immediately from (1). The proof of (3) can be 15  
 16 done similarly to that of Theorem 3.5.  $\square$  16

17 In the proof of Theorem 3.6, we have used the following lemma, whose proof 17  
 18 is given in the online supplementary material [Wang, Kim and Li (2013)]. 18

19 LEMMA 6.1. *There exists  $l \in A^-$  which satisfies (6.8).* 19

20 PROOF OF THEOREM 5.2. By (5.1), a local minimizer  $\boldsymbol{\beta}$  necessarily satisfies: 20

$$21 \quad (6.9) \quad -n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \xi_j = 0, \quad j = 1, \dots, p, \quad 21$$

22 where  $\xi_j = \lambda l_j - \frac{\partial h_n(\boldsymbol{\beta})}{\partial \beta_j}$ , with  $l_j = \text{sign}(\beta_j)$  if  $\beta_j \neq 0$  and  $l_j \in [-1, 1]$  otherwise, 22  
 23  $1 \leq j \leq p$ . It is easy to see that  $|\xi_j| \leq \lambda$ ,  $1 \leq j \leq p$ . Although the objective func- 23  
 24 tion is nonconvex, abusing the notation a little, we refer to the collection of all 24  
 25 vectors in the form of the left-hand side of (6.9) as the subdifferential  $\partial Q_n(\boldsymbol{\beta})$  and 25  
 26 refer to a specific element of this set a subgradient. Then the necessary condition 26  
 27 stated above can be considered as an extension of the classical KKT condition. 27

28 Alternatively, minimizing  $Q_n(\boldsymbol{\beta})$  can be expressed as a constrained smooth 28  
 29 minimization problem [e.g., Kim, Choi and Oh (2008)]. By the corresponding 29  
 30 30  
 31 31  
 32 32  
 33 33  
 34 34  
 35 35  
 36 36  
 37 37  
 38 38  
 39 39  
 40 40  
 41 41  
 42 42  
 43 43

1 second-order sufficiency of KKT condition [e.g., Bertsekas (1999), page 320],  $\widehat{\boldsymbol{\beta}}$  1  
 2 is a local minimizer of  $Q_n(\boldsymbol{\beta})$  if 2

$$3 \quad n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \text{sgn}(\widehat{\beta}_j) \dot{p}_\lambda(\widehat{\beta}_j), \quad \widehat{\beta}_j \neq 0, \quad 3$$

$$4 \quad n^{-1} |\mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})| \leq \lambda, \quad \widehat{\beta}_j = 0. \quad 4$$

5 Consider the event  $F_n = F_{n2} \cap F_{n6}$ , where  $F_{n2}$  is defined in Lemma 3.1 with 5  
 6  $b_2 = 1$ , and  $F_{n6} = \{\min_{j \in A_0} |\widehat{\beta}_j^{(o)}| \geq a\lambda\}$ . Since  $|\widehat{\beta}_j^{(o)}| \geq |\beta_j^*| - |\widehat{\beta}_j^{(o)} - \beta_j^*|$  and 6  
 7  $\lambda = o(d_*)$ , similarly as in the proof for Lemma 3.1, we can show that for all  $n$  7  
 8 sufficiently large,  $P(F_{n6}) \geq 1 - 2q \exp[-C_1 n (d_* - a\lambda)^2 / (2\sigma^2)]$ . By Lemma 3.1, 8  
 9 for all  $n$  sufficiently large,  $P(F_n) \geq 1 - 2q \exp[-C_1 n (d_* - a\lambda)^2 / (2\sigma^2)] - 2(p - 9  
 10  $q) \exp[-n\lambda^2 / (2\sigma^2)]$ . It is apparent that on the event  $F_n$ , the oracle estimator  $\widehat{\boldsymbol{\beta}}^{(o)}$  10  
 11 satisfies the above sufficient condition. Therefore, by (6.9), there exist  $|\xi_j^{(o)}| \leq \lambda$ , 11  
 12  $1 \leq j \leq p$ , such that 12$

$$13 \quad -n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(o)}) + \xi_j^{(o)} = 0. \quad 13$$

14 Abusing notation a little, we denote this zero vector by  $\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\widehat{\boldsymbol{\beta}}^{(o)})$ . 14

15 Now for any local minimizer  $\widehat{\boldsymbol{\beta}}$  which satisfies the sparsity constraint  $\|\widehat{\boldsymbol{\beta}}\|_0 \leq 15$

16  $qu_n$ , we will prove by contradiction that under the conditions of the theorem we 16  
 17 must have  $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(o)}\| \leq 2\lambda \sqrt{qu_n^*} \xi_{\min}^{-1}(qu_n^*)$ , where  $u_n^* = u_n + 1$ . More specifically, 17  
 18 we will derive a contradiction by showing that none of the subgradients of  $Q_n(\boldsymbol{\beta})$  18  
 19 can be zero at  $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ . 19  
 20 Assume instead that  $\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(o)}\| > 2\lambda \sqrt{qu_n^*} \xi_{\min}^{-1}(qu_n^*)$ . Let  $A^* = \{j : \widehat{\beta}_j \neq 0 \text{ or}$  20  
 21  $\widehat{\beta}_j^{(o)} \neq 0\}$ , then  $\|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\| > 2\lambda \sqrt{qu_n^*} \xi_{\min}^{-1}(qu_n^*)$ . Let  $\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\widehat{\boldsymbol{\beta}}) = -n^{-1} \mathbf{x}_{(j)}^T (\mathbf{y} -$  21  
 22  $\mathbf{X}\widehat{\boldsymbol{\beta}}) + \eta_j$  be an arbitrary subgradient in the subdifferential  $\partial Q_n(\widehat{\boldsymbol{\beta}})$ . Let  $\boldsymbol{\eta} =$  22  
 23  $(\eta_1, \dots, \eta_p)^T$ , then  $\eta_j$  satisfies  $|\eta_j| \leq \lambda$ ,  $1 \leq j \leq p$ . We use  $\frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}})$  to denote 23  
 24 the size- $|A^*|$  subvector of  $\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\widehat{\boldsymbol{\beta}})$ , that is,  $\frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}}) = (\frac{\partial}{\partial \beta_j} Q_n(\widehat{\boldsymbol{\beta}}) : j \in A^*)^T$ . 24  
 25 And  $\frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}}^{(o)})$  is defined similarly. We have 25  
 26

$$26 \quad \left| \left( \frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}}) \right)^T \frac{(\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)})}{\|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\|} \right| \quad 26$$

$$27 \quad = \left| \left( \frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}}) - \frac{\partial}{\partial \boldsymbol{\beta}_{A^*}} Q_n(\widehat{\boldsymbol{\beta}}^{(o)}) \right)^T \frac{(\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)})}{\|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\|} \right| \quad 27$$

$$28 \quad = |n^{-1} (\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)})^T \mathbf{X}_{A^*}^T \mathbf{X}_{A^*} (\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}) / \|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\| \quad 28$$

$$29 \quad \quad + (\boldsymbol{\eta}_{A^*} - \boldsymbol{\xi}_{A^*}^{(o)})^T (\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}) / \|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\| \quad 29$$

$$30 \quad \geq \phi_{\min}(n^{-1} \mathbf{X}_{A^*}^T \mathbf{X}_{A^*}) \|\widehat{\boldsymbol{\beta}}_{A^*} - \widehat{\boldsymbol{\beta}}_{A^*}^{(o)}\| - 2\lambda \sqrt{qu_n^*} \quad 30$$

$$31 \quad > \xi_{\min}(qu_n^*) 2\lambda \sqrt{qu_n^*} \xi_{\min}^{-1}(qu_n^*) - 2\lambda \sqrt{qu_n^*} = 0, \quad 31$$

1 where the second equality follows from the expression of subgradient, the second  
 2 last inequality applies the Cauchy–Schwarz inequality, and the last inequality fol-  
 3 lows from the relaxed SRC condition in an  $L_0$ -neighborhood of the true model.  
 4 Thus, this contradicts with the fact that at least one of the subgradients is zero if  $\hat{\beta}$   
 5 is a local minimizer and the theorem is proved.  $\square$

6  
 7 PROOF OF COROLLARY 5.3. It follows directly from Theorem 5.2.  $\square$

## 8 SUPPLEMENTARY MATERIAL

9  
 10 **Supplement to “Calibrating nonconvex penalized regression in ultra-high**  
 11 **dimension”** (DOI: [10.1214/13-AOS1159SUPP](https://doi.org/10.1214/13-AOS1159SUPP); .pdf). This supplemental material  
 12 includes the proofs of Lemmas 3.1 and 6.1, and some additional numerical results.  
 13

## 14 REFERENCES

- 15  
 16 BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA. 16 <auto>  
 17 BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and 17  
 18 Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469](#) 18 <mr>  
 19 BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods,* 19  
 20 *Theory and Applications*. Springer, Heidelberg. [MR2807761](#) 20 <mr>  
 21 CAI, T. and ZHOU, H. (2013). Minimax estimation of large covariance matrices under  $l_1$  norm. 21 <unstr>  
 22 *Statist. Sinica*. To appear. 22  
 23 CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with 23 <mr>  
 24 large model spaces. *Biometrika* **95** 759–771. [MR2443189](#) 24  
 25 FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle 25 <mr>  
 26 properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581](#) 26  
 27 FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans.* 27 <mr>  
 28 *Inform. Theory* **57** 5467–5484. [MR2849368](#) 28  
 29 FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of param- 29 <mr>  
 30 eters. *Ann. Statist.* **32** 928–961. [MR2065194](#) 30 <auto>  
 31 FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometric regression tools 31  
 32 (with discussion). *Technometrics* **35** 109–148. 32  
 33 GAO, B.-B., PHIPPS, J. A., BURSELL, D., CLERMONT, A. C. and FEENER, E. P. (2009). An- 33  
 34 giotensin AT1 receptor antagonism ameliorates murine retinal proteome changes induced by diabe- 34 <mr>  
 35 tes. *J. Proteome Res.* **8** 5541–5549. 35  
 36 HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regres- 36 <mr>  
 37 sion models. *Statist. Sinica* **18** 1603–1618. [MR2469326](#) 37  
 38 HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617– 38 <mr>  
 39 1642. [MR2166557](#) 39  
 40 KIM, Y., CHOI, H. and OH, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. 40  
 41 *J. Amer. Statist. Assoc.* **103** 1665–1673. [MR2510294](#) 41 <mr>  
 42 KIM, Y. and KWON, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika* **99** 42  
 43 315–325. [MR2931256](#) 43 <mr>  
 44 KIM, Y., KWON, S. and CHOI, H. (2012). Consistent model selection criteria on high dimensions. 44  
 45 *J. Mach. Learn. Res.* **13** 1037–1057. [MR2930632](#) 45 <mr>  
 46 KWON, S. and KIM, Y. (2013). Large sample properties of the smoothly clipped absolute deviation 46  
 47 penalized maximum likelihood estimation on high dimensions. *Statistica Sinica*. To appear. 47 <unstr>

- 1 LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and  
2 Dantzig estimators. *Electron. J. Stat.* **2** 90–102. [MR2386087](#) 2 <mr>
- <uncited> 3 MAZUMDER, R., FRIEDMAN, J. and HASTIE, T. (2011). SparseNet: Coordinate descent with non-  
4 convex penalties. *J. Amer. Statist. Assoc.* **106** 1125–1138. 3 <auto>
- 5 MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-  
6 dimensional data. *Ann. Statist.* **37** 246–270. [MR2488351](#) 4 <auto>
- 7 MIKOSCH, T. (1990). Estimates for tail probabilities of quadratic and bilinear forms in sub-Gaussian  
8 random variables with applications to the law of the iterated logarithm. *Probab. Math. Statist.* **11**  
9 169–178. [MR1125746](#) 5 <mr>
- 10 PAUNEL-GÖRGÜLÜ, A. N., FRANKE, A. G., PAULSEN, F. P. and DÜNKER, N. (2011). Trefoil  
11 factor family peptide 2 acts pro-proliferative and pro-apoptotic in the murine retina. *Histochem.*  
12 *Cell Biol.* **135** 461–473. 6 <mr>
- <uncited> 13 RINALDO, A. (2007). A note on the uniqueness of the Lasso solution. Technical report, Dept. Statis-  
14 tics, Univ. Carneige Mellon. 7 <mr>
- 15 SCHEETZ, T. E., KIM, K. Y. A., SWIDERSKI, R. E., PHILP, A. R., BRAUN, T. A.,  
16 KNUDTSON, K. L., DORRANCE, A. M., DiBONA, G. F., HUANG, J., CASAVANT, T. L.,  
17 SHEFFIELD, V. C. and STONE, E. M. (2006). Regulation of gene expression in the mammalian  
18 eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **103** 14429–14434. 8 <mr>
- 19 TAO, P. D. and AN, L. T. H. (1997). Convex analysis approach to d.c. programming: Theory, algo-  
20 rithms and applications. *Acta Math. Vietnam.* **22** 289–355. [MR1479751](#) 9 <pbm>
- 21 TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*  
22 *Stat. Methodol.* **58** 267–288. [MR1379242](#) 10 <unstr>
- 23 VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann.*  
24 *Statist.* **36** 614–645. [MR2396809](#) 11 <unstr>
- 25 VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso  
26 for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.* **5** 688–  
27 749. [MR2820636](#) 12 <unstr>
- 28 WANG, H., LI, B. and LENG, C. (2009). Shrinkage tuning parameter selection with a diverging  
29 number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 671–683. [MR2749913](#) 13 <unstr>
- 30 WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped  
31 absolute deviation method. *Biometrika* **94** 553–568. [MR2410008](#) 14 <unstr>
- 32 WANG, L., KIM, Y. and LI, R. (2013). Supplement to “Calibrating nonconvex penalized regression  
33 in ultra-high dimension.” DOI:[10.1214/13-AOS1159SUPP](#). 15 <unstr>
- 34 YUILLE, A. L. and RANGARAJAN, A. (2003). The concave–convex procedure. *Neural Comput.* **15**  
35 915–936. 16 <unstr>
- 36 ZHANG, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann.*  
37 *Statist.* **38** 894–942. [MR2604701](#) 17 <unstr>
- <uncited> 38 ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach.*  
39 *Learn. Res.* **11** 1081–1107. [MR2629825](#) 18 <unstr>
- 40 ZHANG, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli*. To appear. 19 <unstr>
- 41 ZHANG, C. H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-  
42 dimensional regression. *Ann. Statist.* **36** 156–594. 20 <unstr>
- 43 ZHANG, Y., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized  
information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. [MR2656055](#) 21 <unstr>
- 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](#) 22 <unstr>
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**  
2541–2563. [MR2274449](#) 23 <unstr>
- ZHOU, S. H. (2010). Thresholded Lasso for high dimensional variable selection and statistical esti-  
mation. Available at [arXiv:1002.1583](#). 24 <unstr>

- 1 ZHOU, S. H., VAN DE GEER, S. A. and BÜHLMANN, P. (2009). Adaptive Lasso for high dimen- 1  
 2 sional regression and Gaussian graphical modeling. Available at [arXiv:0903.2515](https://arxiv.org/abs/0903.2515). 2 <unstr>  
 3 ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. 3 <mr>  
 4 [MR2279469](https://doi.org/10.1198/016214506000000333) 4  
 5 ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. 5 <mr>  
 6 *Ann. Statist.* **36** 1509–1533. [MR2435443](https://doi.org/10.1214/07-AN117) 6
- |   |  |  |
|---|--|--|
| <p>7 L. WANG<br/>       8 SCHOOL OF STATISTICS<br/>       9 UNIVERSITY OF MINNESOTA<br/>       10 MINNEAPOLIS, MINNESOTA 55455<br/>       11 USA<br/>       12 E-MAIL: <a href="mailto:wangx346@umn.edu">wangx346@umn.edu</a></p> | <p>Y. KIM<br/>       DEPARTMENT OF STATISTICS<br/>       SEOUL NATIONAL UNIVERSITY<br/>       SEOUL, KOREA<br/>       E-MAIL: <a href="mailto:ydkim0903@gmail.com">ydkim0903@gmail.com</a></p> | <p>7<br/>8<br/>9<br/>10<br/>11<br/>12<br/>13<br/>14<br/>15<br/>16<br/>17<br/>18<br/>19<br/>20<br/>21<br/>22<br/>23<br/>24<br/>25<br/>26<br/>27<br/>28<br/>29<br/>30<br/>31<br/>32<br/>33<br/>34<br/>35<br/>36<br/>37<br/>38<br/>39<br/>40<br/>41<br/>42<br/>43</p> |
|---|--|--|
- 
- |  |  |
|--|--|
| <p>12 R. LI<br/>       13 DEPARTMENT OF STATISTICS<br/>       14 AND THE METHODOLOGY CENTER<br/>       15 PENNSYLVANIA STATE UNIVERSITY<br/>       16 UNIVERSITY PARK, PENNSYLVANIA 16802<br/>       17 USA<br/>       18 E-MAIL: <a href="mailto:rzli@psu.edu">rzli@psu.edu</a></p> | <p>12<br/>13<br/>14<br/>15<br/>16<br/>17<br/>18<br/>19<br/>20<br/>21<br/>22<br/>23<br/>24<br/>25<br/>26<br/>27<br/>28<br/>29<br/>30<br/>31<br/>32<br/>33<br/>34<br/>35<br/>36<br/>37<br/>38<br/>39<br/>40<br/>41<br/>42<br/>43</p> |
|--|--|



## THE ORIGINAL REFERENCE LIST

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43
- The list of entries below corresponds to the original Reference section of your article. The bibliography section on previous page was retrieved from MathSciNet applying an automated procedure.
- Please check both lists and indicate those entries which lead to mistaken sources in automatically generated Reference list.
- Bertsekas, D. P. (1999) *Nonlinear programming* (2nd edition). Athena Scientific, Belmont, Mass.
- Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **37**, 1705-1732.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Cai, T. and Zhou, H. (2011). Minimax estimation of large covariance matrices under  $l_1$  norm. To appear in *Statistica Sinica*.
- Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759-771.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.
- Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Transaction on Information Theory*, **57**, 5467-5484.
- Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, **32**, 928 - 961.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometric regression tools (with discussion). *Technometrics*, **35**, 109-148.
- Gao, B. B., Phipps, J. A., Bursell, D., Clermont, A. C. and Feener, E. P. (2009). Angiotensin AT1 receptor antagonism ameliorates murine retinal proteome changes induced by diabetes. *Journal of Proteome Research*, **8**, 5541-5549.**
- Huang, J., Ma, S.G. and Zhang, C-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.
- Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics*, **33**, 1617 - 1642.
- Kim, Y., Choi, H. and Oh, H-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, **103**, 1665-1673.
- Kim, Y. and Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315-325.
- Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, **13**, 1037-1057.
- Kwon, S. and Kim, Y. (2011). Large sample properties of the smoothly clipped absolute deviation penalized maximum likelihood estimation on high dimensions. Accepted by *Statistica Sinica*.
- Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electric Journal of Statistics*, **2**, 90-102.
- Mazumder, R., Friedman, J. and Hastie, T. (2011). SparseNet: coordinate descent with non-convex penalties. *Journal of the American Statistical Association*, **106**, 1125-1138.
- Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, **37**, 246-270.
- Mikosch, T. (1991). Estimates for tail probabilities of quadratic and bilinear forms in subgaussian random variables. *Probability and Mathematical Statistics*, **11**, 169-178.
- Paunel-Görgülü, A. N., Franke, A.G., Paulsen, F. P. and Dünker, N. (2011). Trefoil factor family peptide 2 acts pro-proliferative and pro-apoptotic in the murine retina. *Histochemistry and Cell Biology*, **135**, 461-473.**
- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43

- 1 Rinaldo, A. (2007). A note on the uniqueness of the Lasso solution. Technical Report, Department  
2 of Statistics, Carnegie Mellon University. 2
- 3 Scheetz, T.E., Kim, K.-Y. A., Swiderski, R.E., Philp, A.R., Braun, T.A., Knudtson, K.L., Dorrance,  
4 A.M., DiBona, G.F., Huang, J., Casavant, T.L., Sheffield, V.C. and Stone, E.M. (2006). Regulation  
5 of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the  
6 National Academy of Sciences*, **103**, 14429-14434. 6
- 7 **Tao, P. D. and An, L.T.H. (1997). Convex analysis approach to D.C. programming: theory,  
8 algorithms and applications. *Acta Mathematica Vietnamica*, **22**, 289-355. 7**
- 9 Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical  
10 Society, Series B*, **58**, 267–288. 8
- 11 van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of  
12 Statistics*, **36**, 614-645. 9
- 13 van de Geer, S. A., Bühlmann, P. and Zhou, S. H. (2011). The adaptive and the thresholded Lasso for  
14 potentially misspecified models (and a lower bound for the Lasso) *Electronic Journal of Statistics*,  
15 **5**, 688-749. 11
- 16 Wang, H., Li, B., and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number  
17 of parameters, *Journal of Royal Statistical Society, Series B*, **71**, 671–683. 14
- 18 Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute  
19 deviation method. *Biometrika*, **94**, 553-568. 15
- 20 Wang, L., Kim, Y. and Li, R. (2013). Supplement to “Calibrating non-convex penalized regression in  
21 ultra-high dimension”. 16
- 22 Yuille, A. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, **15**, 915-  
23 936. 17
- 24 Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of  
25 Statistics*, **38**, 894-942. 18
- 26 Zhang, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *Journal of  
27 Machine Learning Research*, **11**, 1080-1107. 19
- 28 Zhang, T. (2012). Multi-stage Convex Relaxation for Feature Selection. *Bernoulli*. To appear 20
- 29 Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-  
30 dimensional regression. *Annals of Statistics*, **36**, 156-594. 21
- 31 Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized infor-  
32 mation criterion. *Journal of American Statistical Association*, **105**, 312-323. 22
- 33 Zhang, C.-H. and Zhang, T. (2012) A general theory of concave regularization for high dimensional  
34 sparse estimation problems. *Statistical Science*, **27** 576-593. 23
- 35 Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning  
36 Reserach*, **7**, 2541–2563. 24
- 37 Zhou, S. H. (2010) Thresholded Lasso for high dimensional variable selection and statistical estima-  
38 tion. arxiv 1002.1583. 25
- 39 Zhou, S. H., van de Geer, S. A. and Bühlmann, P. (2009) Adaptive Lasso for high dimensional  
40 regression and Gaussian graphical modeling. arxiv:0903.2515. 26
- 41 Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical  
42 Association*, **101**, 1418–1429. 27
- 43 Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models  
(with discussion). *Annals of Statistics*, **36**, 1509 - 1533. 28

1 META DATA IN THE PDF FILE 1

2 Following information will be included as pdf file Document Properties: 2

3  
4 **Title** : Calibrating nonconvex penalized regression in ultra-high 4  
dimension 5

5 **Author** : Lan Wang, Yongdai Kim, Runze Li 5

6 **Subject** : The Annals of Statistics, 2013, Vol. 0, No. 00, 1-32 6

7 **Keywords**: 62J05, 62J07, High-dimensional regression, LASSO, MCP, SCAD, 7  
8 variable selection, penalized least squares 8

9  
10 THE LIST OF URI ADDRESSES 10

11  
12  
13 Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are 13  
14 indicated as ERROR. Please check and update the list where necessary. The e-mail addresses 14  
15 are not checked – they are listed just for your information. More information can be found in 15  
16 the support page: 16

17 <http://www.e-publications.org/ims/support/urihelp.html>. 17

18 200 <http://www.imstat.org/aos/> [2:pp.1,1] OK 18

19 200 <http://www.imstat.org> [2:pp.1,1] OK 19

20 200 <http://www.ams.org/mathscinet/msc/msc2010.html> [2:pp.1,1] OK 20

21 404 <http://dx.doi.org/10.1214/13-AOS1159SUPP> [4:pp.30,30,35,35] Not Found 21

22 200 <http://arxiv.org/abs/1002.1583> [2:pp.37,37] OK 22

23 200 <http://arxiv.org/abs/0903.2515> [2:pp.37,37] OK 23

24 --- mailto:wangx346@umn.edu [2:pp.38,38] Check skip 24

25 --- mailto:ydkim0903@gmail.com [2:pp.38,38] Check skip 25

26 --- mailto:rzli@psu.edu [2:pp.38,38] Check skip 26

27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43