Ann Inst Stat Math (2013) 65:807–832
DOI 10.1007/s10463-012-0396-3

On constrained and regularized high-dimensional regression

Xiaotong Shen · Wei Pan · Yunzhang Zhu · Hui Zhou

Received: 2 July 2012 / Revised: 12 October 2012 / Published online: 12 January 2013

Abstract High-dimensional feature selection has become increasingly crucial for seeking parsimonious models in estimation. For selection consistency, we derive one necessary and sufficient condition formulated on the notion of degree of separation. The minimal degree of separation is necessary for any method to be selection consistent. At a level slightly higher than the minimal degree of separation, selection consistency is achieved by a constrained L_0-method and its computational surrogate—the constrained truncated L_1-method. This permits up to exponentially many features in the sample size. In other words, these methods are optimal in feature selection against any selection method. In contrast, their regularization counterparts—the L_0-regularization and truncated L_1-regularization methods enable so under slightly stronger assumptions. More importantly, sharper parameter estimation/prediction is realized through such selection, leading to minimax parameter estimation. This, otherwise, is impossible in the absence of a good selection method for high-dimensional analysis.

Keywords Constrained regression · Parameter and nonparametric models · Nonconvex regularization · Difference convex programming · (p, n) versus fixed p-asymptotics

Research supported in part by NSF grant DMS-0906616 and DMS-1207771, and NIH grants 1R01GM081535-01 and HL65462.

The authors thank the editors and the reviewers for helpful comments and suggestions.

X. Shen (✉) · Y. Zhu
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA
e-mail: xshen@stat.umn.edu

W. Pan · H. Zhou
Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

1 Introduction

Feature selection is one effective means for sparse modeling in knowledge discovery. Despite the progress in low-dimensional analysis, there remain many important issues. One such issue is to what extent informative features can be reconstructed given a limited amount of data at hand. Towards high-dimensional feature selection, we derive one necessary condition for feature selection, which is attainable by the constrained method and is nearly attained by the method of regularization. On this basis, we further explore these methods for parameter estimation as a result of such a selection.

Consider feature selection based on a random sample $(Y_i, \boldsymbol{x}_i)_{i=1}^n$ from:

$$Y_i = \mu_i + \epsilon_i; \quad \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^0; \quad \epsilon_i \sim N(0, \sigma^2); \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0) = (\boldsymbol{\beta}_{A_0}, \mathbf{0}_{A_0^c})^T$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ are p-dimensional vectors of regression coefficients and features (predictors), and $\boldsymbol{x}_i$ is independent of random error ϵ_i. In (1), feature selection estimates $A_0 = \{j : \beta_j^0 \neq 0\}$ of informative features, together with estimation of true coefficients $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_{A_0}, \mathbf{0}_{A_0^c})^T$, where $\mathbf{0}_{A_0^c}$ denotes a vector of 0s over its complement A_0^c of A_0, and representation $\mu = \boldsymbol{\beta}^T \boldsymbol{x}$ is generic, encompassing, for instance, linear regression and basis pursuit (Chen et al. 2001). Of particular interest is a high-dimensional situation in which p can be much larger than n, and A_0 may depend on (p, n) with $p_0 = |A_0|$, where $|A|$ denotes the size of set A. This describes parametric and nonparametric cases, with A_0 corresponding to a true model as in the parametric case when A_0 is independent of (p, n), and a best approximation of a true model as in basis pursuit otherwise.

Recently, considerable effort has been devoted to selection consistency under (1) to push feature selection into an ultra-high dimensional situation. In a situation as such, little is known about selection consistency for many methods in terms of (p, n)-asymptotics as $n, p \to \infty$, although some methods such as adaptive Lasso (Zou 2006; Zou and Li 2008) have been examined for fixed p-asymptotics as $n \to \infty$. For (p, n)-asymptotics, Bayesian information criterion (BIC; Schwarz 1978), which is derived under a fixed p-asymptotic approximation of the posterior model probability, needs to be modified to accommodate a higher dimension. In Chen and Chen (2008), it is showed that a modified BIC is selection consistent when p is of order of n^κ for some $\kappa > 0$; Liu and Yang (2010) proved that another modified BIC allows p to be an order of $\exp(cp_0 n)$ for some $c > 0$. It appears that exponentially many features are possible for some methods. For L_1-regularization, Tibshirani (1996), Meinshausen and Buhlmann (2006), Zhao and Yu (2006) and Wainwright (2009) proved that the Lasso is sign consistent and thus selection consistent, under a strong irrepresentable assumption that is nearly necessary. As pointed by Zhang (2010), this assumption is restrictive because of nonadaptiveness of the Lasso. For the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) regularization, Kim et al. (2008) and Lv and Fan (2009) showed that some consistent local minimizers exist for SCAD. More recently, Zhang (2010) proved that the minimum concavity penalty (MCP) is selection consistent under a sparse Riesz condition and an information requirement, where the sparse Riesz condition is weaker than the irrepresentable assumption;

Shen et al. (2012) showed that a global minimizer of the constrained L_0-method is selection consistent, under a "degree-of-separation" condition under the Hellinger distance. To understand how a method performs in a high-dimensional situation, it is imperative that we study necessary and sufficient conditions for selection consistency for feature selection, which is a nonconvex problem itself.

This paper establishes results with selection consistency. First, we characterize consistent feature selection for any method through one simple necessary condition in the L_2-metric, which is sufficient up to a constant factor. Now define a measure of the level of difficulty for feature selection: $C_{\min} = C_{\min}(\boldsymbol{\beta}^0, \boldsymbol{X}) \equiv \min_{\{\beta_A : A \neq A_0, |A| \leq p_0\}} \frac{1}{n \max(|A_0 \setminus A|, 1)} \|\boldsymbol{X}_{A_0} \boldsymbol{\beta}^0_{A_0} - \boldsymbol{X}_A \boldsymbol{\beta}_A\|^2$, $\boldsymbol{X}_A$ and $\boldsymbol{\beta}_A$ are the design matrix for subset A of predictors and the regression coefficient vector over A, and $\|\cdot\|$ is the usual Euclidean norm in $\mathcal{R}^n$. The measure $C_{\min}$ defines the degree of separation between A_0 and a least favorable candidate model for feature selection in the L_2-norm, which occurs among candidate models of sizes p_0 or less. As indicated in Theorem 1, roughly, a requirement for selection consistency is

$$C_{\min}(\boldsymbol{\beta}^0, \boldsymbol{X}) \geq d_1 \sigma^2 \frac{\log p}{n}, \tag{2}$$

for some positive constant $d_1 \leq 1/4$ that may depend on $\boldsymbol{X}$. In short, the minimal degree of separation is required for correct identification of informative features, translating to an upper bound on p that is in an order of $\exp(n \frac{C_{\min}}{d_1 \sigma^2})$, for any method and $(\boldsymbol{\beta}_0, \boldsymbol{X})$. This further sharpens the result of Shen et al. (2012) in (1). In view of (2), the Lasso does not achieve feature selection under (2), and it remains unknown if either the SCAD or MCP does.

This paper addresses an attainment issue of the necessary condition (2) with regard to (p_0, p, n). Specifically, we prove, in Theorems 2 and 3, selection consistency is achieved under (2) by global minimizers of the constrained L_0-method and its computational surrogate—the truncated L_1-method for some $d_1 > 0$, respectively defined in (8) and (13). Most importantly, as showed in Theorems 4 and 5, its regularization counterparts defined in (9) and (16) yield selection consistency under a stronger version of (2):

$$C^*_{\min} \geq d_1 \sigma^2 \frac{\log p}{n}, \text{ if } \alpha > 1, \quad C^*_{\min} \geq d_1 \sigma^2 \frac{p_0 \max\left(\log \frac{p}{p_0}, 1\right)}{n}, \text{ if } \alpha = 1, \tag{3}$$

for some $d_1 > 0$, where $C^*_{\min} \equiv \min_{\{\beta_A : A \neq A_0, |A| \leq \alpha p_0\}} \frac{1}{n \max(|A_0 \setminus A|, 1)} \|\boldsymbol{X}_{A_0} \boldsymbol{\beta}^0_{A_0} - \boldsymbol{X}_A \boldsymbol{\beta}_A\|^2$. This says that the L_0-regularization and truncated L_1-regularization methods are optimal when p_0 is independent of (p, n), as in the parametric case, but may be suboptimal when p_0 depends on (p, n). In this sense, the constrained method is more preferable because of its theoretical merits. Note that these two methods are not equivalent for a nonconvex problem, which is unlike an L_1 problem. Moreover, for these methods, selection consistency holds uniformly over $B_0(u, l) = \{\boldsymbol{\beta} : p_0 = \sum_{j=1}^p I(\beta_j \neq 0) \leq u, C_{\min}(\boldsymbol{\beta}, \boldsymbol{X}) \geq l\}$ with $l = d_1 \sigma^2 \frac{\log p}{n}$ and constant $d_1 > 0$,

which is called an L_0-band with upper and lower radii u and l ($u > l > 0$), and is a subset of an L_0-ball that is most relevant to feature selection.

This paper also addresses another issue—parameter estimation involving feature selection. In a low-dimensional situation, it is known that Akaike's information criterion (Akaike 1973) is optimal in parameter estimation/prediction even if it can be inconsistent in feature selection, c.f., Yang and Barron (1998). In other words, optimal parameter estimation can be achieved without feature selection. In a high-dimensional situation, it is no longer the case. In (1), the minimax rate of convergence in the L_2-norm over an L_0-ball $B_0(u, 0)$ is $\sqrt{\frac{u \log(p/u)}{n}}$ (Raskutti et al. 2009), which is optimal for parameter estimation without feature selection. As to be seen, sharper accuracy of parameter estimation can be achieved through removal of noninformative features by a good selection method. In particular, as showed in Theorems 2–6, a minimax rate $\sqrt{\frac{u}{n}}$ in the L_2-risk over an L_0-band $B_0(u, l)$ with some $u > l > 0$ is achieved by the constrained L_0-method as well as its regularization counterpart. Note that excluding a neighborhood of the origin for an L_0-band $B_0(u, l)$ is necessary to assure existence of a good selection method, as suggested in (2). Moreover, the corresponding estimators defined by these methods are asymptotic minimax over $B_0(u, l)$, recovering the optimal risk of the oracle estimator, defined as the least squares estimator given A_0. In short, sharper optimal parameter estimation is achieved by the constrained L_0-method and L_0-regularization method. This is impossible without removal of noninformative features (Raskutti et al. 2009). To our knowledge, it remains largely unknown if this property is shared by other methods.

Finally, for constrained truncated L_1-regression, we derive a constrained difference convex (DC) algorithm that is showed to be equivalent to its unconstrained DC algorithm of Shen et al. (2012) with respect to their solutions, although constrained L_0-regression and L_0-regularization methods are not generally equivalent with regard to their global minimizers. Importantly, we show that a local minimizer of the regularization criterion does share the desirable properties as a global minimizer under stronger assumptions, c.f., Theorem 6.

The paper is organized in five sections. Section 2 derives the necessary condition (2) for selection consistency. Section 3 constructs an optimal constrained method to address the attainment issue, in addition to optimal parameter estimation. Section 4 derives parallel results for its regularization counterpart. Section 5 establishes equivalence between a constrained DC algorithm and its unconstrained counterpart with regard to their solutions. The appendix contains technical proofs.

2 Necessary conditions

This section establishes the necessary condition (2) by estimating the minimal degree of separation required for selection consistency.

Selection consistency requires that $P(\hat{A} \neq A_0) \to 0$ as $n, p \to \infty$ under the true probability P, for an estimate $\hat{A} = \{j : \hat{\beta}_j \neq 0; j = 1, \ldots, p\}$ of $A_0 = \{j : \beta_j^0 \neq 0; j = 1, \ldots, p\}$. To derive a lower bound requirement for $C_{\min}(\boldsymbol{\beta}^0, \boldsymbol{X})$, we construct an approximate least favorable situation under P, over an L_0-band $B_0(u, l)$,

as defined in Sect. 1, to avoid superefficiency (Ibragimov and Has'minskii 1981). Then we estimate the smallest possible value of $l > 0$ under which selection consistency holds for $\hat{A}$ over $\boldsymbol{\beta}^0 \in B_0(u, l)$, that is,

$$\sup_{\{\boldsymbol{\beta}^0 \in B_0(u,l)\}} P(\hat{A} \neq A_0) \to 0, \quad \text{as } n, p \to 0.$$

Let $r(p_0, \boldsymbol{X}) = \frac{\max_{1\le j\le p} n^{-1}\|\boldsymbol{x}^{(j)}\|^2}{\min_{\boldsymbol{\beta}^0:|\beta_j^0|\ge 1; j\in A_0, |A_0|\le p_0} C_{\min}(\boldsymbol{\beta}^0, \boldsymbol{X})}$, where $A_0 = \{j : \beta_j^0 \neq 0\}$ and $\boldsymbol{x}^{(j)} = (x_{1j}, \ldots, x_{nj})^T$. Theorem 1 below gives a good estimate of l.

Theorem 1 (Necessity for selection consistency) *For any $\hat{A}$ and (u, l) with $u > l > 0$, we have*

$$\sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} P(\hat{A} \neq A_0) \to 0, \quad \textit{as } n, p \to \infty, \tag{4}$$

implying that $l > \frac{1}{4r(u,\boldsymbol{X})}\sigma^2 \frac{\log p}{n}$. Moreover, if $r(u, \boldsymbol{X}) \le \frac{1}{4d_1}$, where $d_1 > 0$ is a constant independent of (n, p), then $l > d_1\sigma^2 \frac{\log p}{n}$ with $d_1 \le 1/4$.

Theorem 1 says that (2) is necessary to achieve selection consistency indeed for any method, as characterized by (4), where the smallest possible l is $\frac{1}{2r(u,\boldsymbol{X})}\sigma^2 \frac{\log p}{n}$, depending on a design matrix $\boldsymbol{X}$ through $r(u, \boldsymbol{X})$. Given $\boldsymbol{X}$, an upper bound of $r(u, \boldsymbol{X})$ may be computed. A loose bound, for instance, can be $r(u, \boldsymbol{X}) \le \frac{\max_{1\le j\le p} n^{-1}\|\boldsymbol{x}^{(j)}\|^2}{\min_{|B|\le 2p_0, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B)}$ by Lemma 1, where $c_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix. Sufficiently, $r(u, \boldsymbol{X})$ is upper bounded by a constant independent of (u, n, p) when $\boldsymbol{x}^{(j)}$; $j = 1, \ldots, p$, are standardized, and $\min_{|B|\le 2p_0, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B)$ is bounded away from zero.

Lemma 1 below gives a connection between $C_{\min}$ and the true signal's resolution level $\gamma_{\min} = \gamma_{\min}(\boldsymbol{\beta}^0) \equiv \min\{|\beta_k^0| : k \in A_0\}$.

Lemma 1

$$\begin{aligned} C_{\min} &= \min_{A_1 \neq A_0, |A_1|\le p_0} n^{-1}\|(I - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}^0\|^2 \\ &\ge \min_{|A_1|\le p_0} c_{\min}(n^{-1}\boldsymbol{X}_{A_0\cap A_1^c}^T(I - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0\cap A_1^c})\gamma_{\min}^2 \\ &\ge \min_{|B|\le 2p_0, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B)\gamma_{\min}^2 \ge 0, \end{aligned} \tag{5}$$

where $\boldsymbol{P}_{A_1}$ is the projection matrix for $\boldsymbol{X}_{A_1}$ with $A_1 \subset \{1, \ldots, p_0\}$. In addition,

$$C_{\min} \le \max_{j\in A_0} n^{-1}\|\boldsymbol{x}^{(j)}\|^2\gamma_{\min}^2 \le c_{\max}(n^{-1}\boldsymbol{X}_{A_0}^T\boldsymbol{X}_{A_0})\gamma_{\min}^2, \tag{6}$$

where $\boldsymbol{x}^{(j)} = (x_{1j}, \ldots, x_{nj})^T$, $c_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix.

For verification of (2), it can be checked using a stronger but simpler condition according to Lemma 1. That is,

$$\gamma_{\min}^2 \min_{|B|\le 2p_0, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B) \ge d_1\sigma^2\frac{\log p}{n}. \tag{7}$$

One major difference between (7) and (2) is that (7) involves eigenvalues of $\boldsymbol{X}_B^T\boldsymbol{X}_B$ with $|B| \le 2p_0$ instead of those of $\boldsymbol{X}_B$ with $|B| \le p_0$ in (2). As a result, (7) may not be tight in that (7) is not satisfied but (2) is. This occurs, for instance, when $\min_{|B|\le 2p_0, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B) = 0$ but $C_{\min} > 0$. This is so when any p_0 features are linearly independent but a set of d features are linearly dependent for $d > p_0$.

Concerning necessary conditions for selection consistency in the literature, Theorem 1 requires less regularity conditions, which are attainable up to a factor d_1 as showed in Theorems 2 and 3. To our knowledge, the best available lower bound is roughly $\gamma_{\min}^2 \ge C_0\frac{\log(p-u)}{n}$ in Theorem 3 of Zhang (2010), under the sparse Riesz condition with a dimension restriction $M_2u + 1 \le d^* \le p$ for some $M_2 \ge 16$, and $\gamma_{\min}^2 \ge C_0\frac{\log(p-u)}{n}$. In particular, under the assumptions there, $C_{\min} \ge d_1^*\sigma^2\frac{\log(p-u)}{n}$ by Lemma 1, for some constant $d_1^* > 0$. Moreover, the assumptions of Theorem 1 may hold even when those of Theorem 3 of Zhang (2010) are not met, which occurs, for instance, in the presence of more than p_0 linearly independent noninformative features.

3 Constrained method

This section addresses the issue of attainment under the necessary condition (2). Specifically, we aim at reconstruction of the oracle estimator—the least squares estimate $\hat{\boldsymbol{\beta}}^{ol} = (\hat{\boldsymbol{\beta}}_{A_0}^{ol}, \mathbf{0}_{A_0^c})^T$ given A_0 by the constrained method, ultimately leading to reconstruction of A_0.

3.1 Constrained L_0-method

Consider constrained least squares regression with the L_0-constraint $\sum_{j=1}^p I(\beta_j \ne 0)$. The constrained least squares criterion is

$$S(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2 \text{ subject to } \sum_{j=1}^p I(\beta_j \ne 0) \le K, \tag{8}$$

where $K > 0$ is an integer-valued tuning parameter. Note that (8) is not equivalent to its unconstrained nonconvex counterpart—the L_0-regularization:

$$\frac{1}{2}\sum_{i=1}^n (y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2 + \lambda\sum_{j=1}^p I(|\beta_j| \ne 0), \tag{9}$$

where $\lambda > 0$ is a regularization parameter corresponding to K in (8).

Moreover, tuning involves a discrete parameter K in (8), which is easier than that for (9) with a continuous parameter $\lambda > 0$. This phenomenon has been also observed in Gu (1998) for spline estimation.

The next theorem says that a global minimizer of (8) $\hat{\boldsymbol{\beta}}^{L_0} = (\hat{\boldsymbol{\beta}}^{L_0}_{\hat{A}^{L_0}}, \mathbf{0})$ consistently reconstructs the oracle estimator at a degree of separation level that is slightly higher than the minimal in (2). Without loss of generality, assume that a global minimizer of (8) exists.

Theorem 2 (Error bound for a global minimizer of (8)) *Under (1), when $K = p_0$, we have, for any (p_0, p, n)*

$$P(\hat{\boldsymbol{\beta}}^{L_0} \neq \hat{\boldsymbol{\beta}}^{ol}) \leq \frac{e+1}{e-1} \exp\left(\frac{n}{18\sigma^2}\left(C_{\min} - 36\frac{\log p}{n}\sigma^2\right)\right). \tag{10}$$

Assume that $u < \min(p, n)$ and constant $d_1 > 36$. Let $l = d_1\sigma^2 \frac{\log p}{n}$. As $n, p \to \infty$, the following results hold:

(A) *Under (2), $\hat{\boldsymbol{\beta}}^{L_0}$ consistently reconstructs $\hat{\boldsymbol{\beta}}^{ol}$, implying selection consistency of $\hat{A}^{L_0}$ for A_0. Moreover,*

$$\sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} P(\hat{A}^{L_0} \neq A_0) \leq \sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} P(\hat{\boldsymbol{\beta}}^{L_0} \neq \hat{\boldsymbol{\beta}}^{ol}) \to 0, \tag{11}$$

which agrees with the lower bound (4) in (p_0, p, n) asymptotically, where $B_0(u, l) = \{\boldsymbol{\beta} : p_0 = \sum_{j=1}^{p} I(\beta_j \neq 0) \leq u, C_{\min}(\boldsymbol{\beta}, \boldsymbol{X}) \geq l\}$.

(B) *Under (2), $n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0} - \boldsymbol{\beta}^0)\|^2 = (1 + o(1))n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0)\|^2 = \sigma^2\frac{p_0}{n}$. In addition, $\hat{\boldsymbol{\beta}}^{L_0}$ is risk-minimax in that*

$$\begin{aligned}\sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0} - \boldsymbol{\beta}^0)\|^2 &= (1 + o(1))n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{ol} - \boldsymbol{\beta}^0)\|^2 = \sigma^2\frac{u}{n} \\ &= \inf_{T_n} \sup_{\boldsymbol{\beta} \in B_0(u,l)} n^{-1}E\|\boldsymbol{X}(T_n - \boldsymbol{\beta}^0)\|^2. \end{aligned} \tag{12}$$

Theorem 2 says that $\hat{\boldsymbol{\beta}}^{L_0}$ consistently reconstructs the oracle estimator $\hat{\boldsymbol{\beta}}^{ol}$, which suffices to establish the attainment of (2) and its uniform version (4) for selection consistency by $\hat{A}^{L_0}$ in (p_0, p, n) except a factor $d_1 > 0$. This permits exponentially many candidate predictors $p \leq p_0 \exp(n\frac{C_{\min}}{d_1\sigma^2})$ for reconstruction. Moreover, $\hat{\boldsymbol{\beta}}^{L_0}$ is risk-minimax optimal for parameter estimation. This is achieved through tuning K over integers ranging from 0 to $\min(n, p)$.

3.2 Constrained truncated L_1-method

We now examine an L_0 surrogate (the truncated L_1-constraint), which was suggested for the method of regularization (Shen et al. 2012). Here the surrogate function $J(|z|)$

is $\min(|z|, \tau)$, approximating the L_0-function as $\tau \to 0$. With this surrogate function, the corresponding constrained least squares criterion in (8) becomes:

$$S(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2 \text{ subject to } \frac{1}{\tau}\sum_{j=1}^{p}\min(|\beta_j|, \tau) \le K, \tag{13}$$

where K and τ are non-negative tuning parameters.

The next theorem presents a parallel result for a global minimizer of (13) $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}^T_{\hat{A}^T}, \mathbf{0})$ as in Theorem 2.

Theorem 3 (Error bound for a global minimizer of (13)) *Under* (1), *if* $K = p_0$ *and* $0 < \tau \le \sigma\sqrt{\frac{6}{(n+2)pc_{\max}(\boldsymbol{X}^T\boldsymbol{X})}}$, *then*

$$P(\hat{\boldsymbol{\beta}}^T \ne \hat{\boldsymbol{\beta}}^{ol}) \le \frac{e+1}{e-1}\exp\left(-\frac{n}{20\sigma^2}\left(C_{\min} - 40\sigma^2\frac{\log p}{n}\right)\right). \tag{14}$$

All the results for $\hat{\boldsymbol{\beta}}^{L_0}$ *in Theorem 2 continue to hold for* $\hat{\boldsymbol{\beta}}^T$ *when* $d_1 > 40$.

For parameter estimation in (B), it is known that the minimax rate of convergence in the L_2-norm is $\sqrt{\frac{u\log(p/u)}{n}}$ over $B_0(u, 0)$, c.f. Raskutti et al. (2009). Nevertheless, a sharper rate of $\sqrt{\frac{p_0}{n}}$ is achieved by the L_0-penalty and its computational surrogate under "degree-of-separation" condition, which can be made uniformly over an L_0-band $B_0(u, l)$ with $l > 0$. In other words, these methods are optimal with regard to parameter estimation, because they recover the optimal L_2-risk of the oracle estimator are asymptotic minimax.

4 Regularization nearly necessary condition

4.1 L_0-regularization

Now consider (9), where we assume, without loss of generality, that a global minimizer exists, because the cost function (9) is bounded by zero almost surely. Denote by $\hat{\boldsymbol{\beta}}^{l_0} = (\hat{\boldsymbol{\beta}}^{l_0}_{\hat{A}^{l_0}}, \mathbf{0})$ a global minimizer of (9).

Theorem 4 (Error bound for a global minimizer of (9)) *Under* (1) *and* $\alpha > 1$,

$$\begin{aligned} P(\hat{\boldsymbol{\beta}}^{l_0} \ne \hat{\boldsymbol{\beta}}^{ol}) \le\ & 4\exp\left(-\left(\frac{nC^*_{\min}}{18\sigma^2} - (\alpha+1)\log(p+1) - \frac{\lambda}{2\sigma^2}\right)\right) \\ & +4\exp\left(-\left(\frac{(\alpha-1)\lambda}{3\alpha\sigma^2} - \left(1+\frac{1}{\alpha}\right)\log(p+1) - \frac{2}{3}\right)\right). \end{aligned} \tag{15}$$

Moreover, if $\sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2 \le c_1 exp(c_2 p_0)$ *for some constant* c_j; $j = 1, 2$, *then all the results in Theorems 2 continue to hold under* (3) *with* $C_{\min}$ *replaced*

by $C^*_{\min}$, *when* $d_1 > \frac{9(\alpha^2+3\alpha+2)}{\alpha-1}$, *and* $\frac{\lambda}{n} \in (\frac{3(\alpha+1)\log(p+1)\sigma^2}{2(\alpha-1)n}, \frac{1}{9}C^*_{\min})$. *Similarly, for* $\alpha = 1$, *all the above results hold under* (3) *with* $C^*_{\min}$ *replaced by* $C_{\min}$, *when* $d_1 > 225$, *and* $\frac{\lambda}{n} \in (18\sigma^2 \frac{p_0 \max(\log \frac{p}{p_0}, 1)}{n}, \frac{1}{9}C_{\min})$.

Theorem 4 derives parallel results of the constrained method under a condition that is slightly stronger. This may be attributed to non-equivalence between these two methods in tuning. Note that the case of $\alpha = 1$ is suboptimal as compared to that of $\alpha > 1$. This is in contrast of the results in Theorems 2 and 3.

4.2 Truncated L_1-regularization

Next consider a global minimizer $\hat{\boldsymbol{\beta}}^{tl} = (\hat{\boldsymbol{\beta}}^{tl}_{\hat{A}^{tl}}, \mathbf{0})$ of the computational surrogate of the L_0-regularization:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2 + \frac{\lambda}{\tau}\sum_{j=1}^{p}\min(|\beta_j|, \tau). \tag{16}$$

For (16), we describe its global minimizer in a simple case to provide an insight into the truncated L_1 function as a computational surrogate of the L_0-function.

Proposition 1 *In the orthogonal design case, the truncated* L_1 *penalty (TLP) estimate defined by* (16) *becomes* $\hat{\beta}^{ol}_j I(|\hat{\beta}^{ol}_j| \geq \sqrt{2\lambda})$ *when* $\tau \leq \sqrt{\lambda/2}$, *which reduces to the thresholding rule defined by a global minimizer of the cost function of* L_0*-regularization* (9), *and is*

$$\begin{cases} \hat{\beta}^{ol}_j & \text{if } |\hat{\beta}^{ol}_j| \geq \frac{\lambda}{2\tau} + \tau; \\ (|\hat{\beta}^{ol}_j| - \frac{\lambda}{\tau})_+ sign(\hat{\beta}^{ol}_j) & \text{if } |\hat{\beta}^{ol}_j| \leq \frac{\lambda}{2\tau} + \tau \end{cases}$$

when $\tau > \sqrt{\lambda/2}$; $j = 1, \ldots, p$. *Here* $\hat{\beta}^{ol}_j$ *is the ordinary least squares estimate for* β_j. *Note that there are two distinct global minimizers if* $|\hat{\beta}^{ol}_j| = \frac{\lambda}{2\tau} + \tau$.

Proposition 1 suggests that the TLP function yields the thresholding rule of the L_0-regularization when the value of τ is small enough in that $\tau \leq \sqrt{\lambda/2}$.

Theorem 5 (Error bound for a global minimizer of (16)) *Under* (1), *if* $0 < \tau \leq \sqrt{\frac{2\lambda}{(n+1)c_{\max}(\boldsymbol{X}^T\boldsymbol{X})}}$ *and* $\alpha > 1$, *then* $P(\hat{\boldsymbol{\beta}}^{tl} \neq \hat{\boldsymbol{\beta}}^{ol})$ *is upper bounded by*

$$\min\left(\frac{\sqrt{2}|A_0|n^{1/2}\tau}{\sigma\sqrt{\pi}c^{-1/2}_{\min}(\frac{1}{n}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})}\exp\left(-\frac{n(\gamma_{\min}-\tau)^2}{2\sigma^2 c^{-1}_{\min}(\frac{1}{n}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})}\right), |A_0|\Phi\left(-\frac{n^{1/2}(\gamma_{\min}-\tau)}{\sigma c^{-1/2}_{\min}(\frac{1}{n}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})}\right)\right)$$

$$+4\exp\left(-\left(\frac{nC^*_{\min}}{20\sigma^2}-(\alpha+1)\log(p+1)-\frac{\lambda}{2\sigma^2}\right)\right)$$
$$+4\exp\left(-\left(\frac{(\alpha-1)\lambda}{3\alpha\sigma^2}-\left(1+\frac{1}{\alpha}\right)(\log(p+1)-\frac{5}{3})\right)\right), \tag{17}$$

where $\Phi(\cdot)$ *is the cumulative distribution function of* $N(0,1)$. *If* $sup_{\boldsymbol{\beta}^0\in B_0(u,l)}\frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2 \le c_1\exp(c_2 p_0)$ *for some constant* c_j; $j=1,2$, *then all the results in Theorem 2 continue to hold under* (3) *with* $C^*_{\min}$ *replaced by* $C_{\min}$, *when* $d_1 > \frac{10(\alpha^2+3\alpha+2)}{\alpha-1}$, *and* $\frac{\lambda}{n} \in (\frac{3(\alpha+1)\log(p+1)\sigma^2}{2(\alpha-1)n}, \frac{1}{10}C^*_{\min})$ *and* $\tau \le \sqrt{\frac{2\lambda}{(n+1)c_{\max}(\boldsymbol{X}^T\boldsymbol{X})}}$. *Similarly, if* $\alpha=1$, *Then all the results continue to hold under* (3) *with* $C_{\min}$ *replaced by* $C^*_{\min}$, *when* $d_1 > 225$, *and* $\frac{\lambda}{n} \in (18\sigma^2\frac{p_0\max(\log\frac{p}{p_0},1)}{n}, \frac{1}{10}C_{\min})$.

Theorem 5 says that the computational surrogate shares the desired statistical properties of the L_0-regularization. This occurs when τ is chosen to be sufficiently small, or $\tau \le \sqrt{\frac{2\lambda}{(n+1)c_{\max}(\boldsymbol{X}^T\boldsymbol{X})}}$. This result suggests that tuning should be concentrated more on λ whereas τ does not need a refined search. In practice, τ should not be too small.

5 Nonconvex minimization

To solve (16), we derive a constrained DC method by approximating the constraint function in (16) by a sequence of nonincreasing approximating functions through DC programming. This is a so-called prime approach for unconstrained regularization that is a dual problem of (16), namely,

$$\frac{1}{2}\sum_{i=1}^{n}(y_i-\boldsymbol{\beta}^T\boldsymbol{x}_i)^2+\frac{\lambda}{\tau}\sum_{j=1}^{p}\min(|\beta_j|,\tau), \tag{18}$$

where $\lambda \ge 0$ is a regularizer or Lagrange multiplier for (16).

To proceed, we first decompose the nonconvex constraint in (16) into a difference to two convex functions:

$$\frac{1}{\tau}\sum_{j=1}^{p}\min(|\beta_j|,\tau)=S_1(\boldsymbol{\beta})-S_2(\boldsymbol{\beta}), \tag{19}$$

where $S_1(\boldsymbol{\beta})=\frac{1}{\tau}\sum_{j=1}^{p}|\beta_j|$ and $S_2(\boldsymbol{\beta})=\frac{1}{\tau}\sum_{j=1}^{p}\max(|\beta_j|-\tau,0)$. Given (19), a sequence of upper approximations of the constraint function is constructed by successively replacing $S_2(\boldsymbol{\beta})$ by its minorization at iteration m:

$$S_1(\boldsymbol{\beta})-(S_2(\hat{\boldsymbol{\beta}}^{(m-1)})+(|\boldsymbol{\beta}|-|\hat{\boldsymbol{\beta}}^{(m-1)}|)^T\nabla S_2(|\hat{\boldsymbol{\beta}}^{(m-1)}|)), \tag{20}$$

where $\nabla S_2=\frac{1}{\tau}I(|\hat{\boldsymbol{\beta}}^{(m-1)}|>\tau)$ is a subgradient of S_2 in $|\boldsymbol{\beta}|$, and $|\cdot|$ is used for vectors, taking the absolute value in each component. At iteration m, the mth subproblem becomes

$$\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}), \quad \text{subject to } \frac{1}{\tau}\sum_{j=1}^{p}|\beta_j| I(|\hat{\beta}_j^{(m-1)}| \le \tau) \le K - \sum_{j=1}^{p} I(|\hat{\beta}_j^{(m-1)}| > \tau). \tag{21}$$

Minimizing (21) in $\boldsymbol{\beta}$ yields its minimizer $\hat{\boldsymbol{\beta}}^{(m)}$. The process continues until termination.

A constrained DC algorithm is summarized as follows:

Algorithm 1

Step 1. (Initialization) Supply a good initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$, say the Lasso estimate.

Step 2. (Iteration) At iteration m, compute $\hat{\boldsymbol{\beta}}^{(m)}$ by solving (21). This can be done through the constrained Lasso algorithm of Osborne et al. (2000), which is implemented in Lasso2 in the R-package.

Step 3. (Stopping rule) Terminate when $S(\hat{\boldsymbol{\beta}}^{(m-1)}) - S(\hat{\boldsymbol{\beta}}^{(m)}) \le 0$. Then the estimate $\hat{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}^{(m^*-1)}$, where m^* is the smallest index satisfying the termination criterion.

There is a connection between the prime approach and its dual approach in Shen et al. (2012), although nonconvex problems (16) and (18) are not equivalent, where (18) is solved through DC programming by approximating the cost function in (18) to minimize

$$S(\boldsymbol{\beta}) + \frac{\lambda}{\tau}\sum_{j=1}^{p}|\beta_j| I(|\hat{\beta}_j^{(m-1)}| \le \tau) \tag{22}$$

iteratively with respect to m. As to be shown in Lemma 2, the prime DC approach as implemented by Algorithm 1 is equivalent to the dual DC approach implemented through Algorithm 1 of Shen et al. (2012). The equivalence is established for their solutions, regardless of the modes of implementation, because a coordinate decent method breaks down for (16) but works for (18). Given the equivalence, no improvement of Algorithm 1 is expected over Algorithm 1 of Shen et al. (2012). We refer to Shen et al. (2012) for simulation comparisons of various methods with regard to accuracy of selection and predictive accuracy.

Lemma 2 (Equivalence) *The DC solution for* (16)*, computed through* (21) *in Algorithm 1 is equivalent to that for* (18)*, computed using Algorithm 1 of Shen et al. (2012). Specifically, given any* $\lambda, 0 \le \lambda < \infty$*, and initial value of Algorithm 1 for* (18)*, there exist a* K *and an initial value of Algorithm 1 of Shen et al. (2012) for* (21) *such that the DC solution of* (22) *is also a DC solution of* (21)*, and vice versa. Moreover, Algorithm 1 has the finite termination property and* $S(\hat{\boldsymbol{\beta}}^{(m)})$ *nonincreases in* m*, as its unconstrained counterpart.*

Now consider a local minimizer of (16) $\hat{\boldsymbol{\beta}}^{lo} = (\hat{\boldsymbol{\beta}}^{lo}_{\hat{A}^{lo}}, \mathbf{0})$ satisfying a local optimality condition of (16):

$$-(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda}{\tau}b_j = 0, \quad j = 1, \ldots, p \tag{23}$$

where $b_j = \text{sign}(\beta_j)$ if $0 < |\beta_j| < \tau$; $b_j \in [-1, 1]$ if $\beta_j = 0$; $b_j = 0$ if $|\beta_j| > \tau$; $b_j = \emptyset$ if $|\beta_j| = \tau$, is the regular subdifferential of $J_{T,\tau}(|\beta_j|)$ at β_j, and $\emptyset$ is the empty set. The reader may consult Rockafellar and Wets (2003) for optimal conditions of continuous but nondifferentiable functions.

Theorem 6 (Error bound for a local minimizer of (16)) *Under* (1)*, for* $\hat{\boldsymbol{\beta}}^{lo}$ *satisfying* (23)*, including the solution from Algorithm 1 of Shen et al. (2012), if* $\tau^2 \geq \frac{4\sqrt{2K^*}\lambda}{n \min_{|B|\leq 2K^*, A_0 \subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B^T \boldsymbol{X}_B)}$ *then,*

$$P(\hat{\boldsymbol{\beta}}^{lo} \neq \hat{\boldsymbol{\beta}}^{ol}) \leq \min\left(\frac{\sqrt{2}|A_0| n^{1/2}(3\tau/2)}{\sqrt{\pi}\sigma c_{\min}^{-1/2}(\frac{1}{n}\boldsymbol{X}_{A_0}^T \boldsymbol{X}_{A_0})} \exp\left(-\frac{n(\gamma_{\min} - 3\tau/2)^2}{2\sigma^2 c_{\min}^{-1}(\frac{1}{n}\boldsymbol{X}_{A_0}^T \boldsymbol{X}_{A_0})} \right), \right.$$
$$\left. |A_0|\Phi\left(-\frac{n^{1/2}(\gamma_{\min} - 3\tau/2)}{\sigma c_{\min}^{-1/2}(\frac{1}{n}\boldsymbol{X}_{A_0}^T \boldsymbol{X}_{A_0})} \right) \right) + (p - |A_0|)\Phi\left(-\frac{\lambda/\tau}{\sigma \max_{1\leq j\leq p} \|\boldsymbol{x}^{(j)}\|} \right), \quad (24)$$

where $K^\star$ *is the upper bound of the maximum number of non-zero predictors, with* $p_0 \leq K^\star \leq \min\{n/2, p\}$*. If* $c_{\max}(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}) p^2\tau^2 \leq c_1 \exp(c_2 p_0)$ *for some constant* $c_1 > 0$*, then all the results in Theorem 4 and 5 continue to hold if* $\tau \leq \frac{\gamma_{\min}}{2}$, $\frac{\log p_0}{n} \leq \frac{c_{\min}(n^{-1}\boldsymbol{X}_{A_0}\boldsymbol{X}_{A_0})\gamma_{\min}^2}{5\sigma^2}$, $\frac{\log p}{n} \leq \frac{\lambda^2}{2\tau^2\sigma^2 n \max_{1\leq j\leq p}\|\boldsymbol{x}^{(j)}\|^2}$*, sufficiently,*

$$\frac{\log p}{n} < \frac{(\min_{|B|\leq 2K^\star, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B\boldsymbol{X}_B))^2\gamma_{\min}^2}{256K^\star\sigma^2} \frac{n}{\max_{j\in A_0}\|x^{(j)}\|^2},$$

where $B_0(u, l)$ *is replaced by*

$$\left\{ \boldsymbol{\beta} \in \mathcal{R}^p : \sum_{j=1}^p I(\beta_j \neq 0) \leq u, \gamma_{\min}^2(\boldsymbol{\beta}) \min_{|B|\leq 2K^*, A_0\subseteq B} c_{\min}(n^{-1}\boldsymbol{X}_B\boldsymbol{X}_B) \geq l \right\},$$

with $l = 256\sigma^2 K^* \frac{\log p}{n} \frac{\max_{j\in A_0}\|\boldsymbol{x}^{(j)}\|^2}{n}$.

Theorem 6 says that a local minimizer of (16) achieves the objectives of a global minimizer of (16) under stronger assumptions.

Lemma 3 *Results in Theorems 1–6 continue to hold for fixed p with* $n \to \infty$ *with* (2) *replaced by* $\lim_{n\to\infty} nC_{\min} = \infty$.

6 Appendix

Proof of Theorem 1 Our proof constructs an approximated least favorable situation for feature selection and uses Fano's Lemma. According to Fano's Lemma (Ibragimov and Has'minskii 1981), for any mapping $T = T(Y_1, \ldots, Y_n)$ taking values in $\{1, \ldots, s\}$, $s^{-1}\sum_{j=1}^s P_j(T(Y_1, \ldots, Y_n) = j) \leq \sum_{1\leq j,k\leq s} n\frac{K(q_j,q_k)+\log 2}{s^2\log(s-1)}$, where

$K(q_j, q_k) = \int q_j \log(q_j/q_k)$ is the Kullback–Leibler information for densities q_j versus q_k corresponding P_j and P_k.

Let $S = \{\boldsymbol{\beta}_j\}_{j=0}^{p}$ be a collection of parameters with components equal to $\gamma_{\min}$ or 0 satisfying that for any $1 \le j, j' \le p+1$, $\|\boldsymbol{\beta}_{j'} - \boldsymbol{\beta}_j\|^2 \le 4\gamma_{\min}^2$. For example, we may choose $\boldsymbol{\beta}_0 = \sum_{k=1}^{p_0-1} \gamma_{\min}\boldsymbol{\delta}_k$, $\boldsymbol{\beta}_j = \boldsymbol{\beta}_0 - \gamma_{\min}\boldsymbol{\delta}_j$; $j = 1, \ldots, p_0 - 1$ and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_0 + \gamma_{\min} e_j$; $j = p_0, \ldots, p$, where $\boldsymbol{\delta}_k$ is a vector of length p with its kth element being 1 and 0 otherwise. Let q_j is the corresponding probability density defined by $\boldsymbol{\beta}_j$, $j = 0, \ldots, p$.

Then we have, for any $\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'} \in S$, $K(q_j, q_{j'}) = \frac{1}{2\sigma^2 n}\|\boldsymbol{X}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})\|^2 \le \frac{2\max_{1\le j\le p}\|\boldsymbol{x}^{(j)}\|^2\gamma_{\min}^2}{n\sigma^2} \le \frac{2r(p_0,\boldsymbol{X})C_{\min}(\boldsymbol{\beta}^0,\boldsymbol{X})}{\sigma^2}$ by Lemma 1. It follows from Fano's lemma with S and $s = p+1$ that $s^{-1}\sum_{j\in S} P_j(T = j) \le \frac{2nr(p_0,\boldsymbol{X})C_{\min}(\boldsymbol{\beta}^0,\boldsymbol{X})+\sigma^2\log 2}{\sigma^2\log p}$, implying that

$$\sup_{\{(\boldsymbol{\beta},\boldsymbol{X}):C_{\min}(\boldsymbol{\beta}^0,\boldsymbol{X})\le R^*(p_0,\boldsymbol{X})\}} P(\hat{A} \ne A_0) \ge 1 - \frac{2nr(p_0,\boldsymbol{X})C_{\min}(\boldsymbol{\beta}^0,\boldsymbol{X}) + \sigma^2\log 2}{\sigma^2\log p}, \tag{25}$$

which is bounded below by a constant $c_* > 0$ with $R^*(p_0,\boldsymbol{X}) = \frac{\sigma^2(1-c_*)\log p}{2nr(p_0,\boldsymbol{X})}$. For (4), if $\sup_{\boldsymbol{\beta}^0\in B_0(u,l)} P(\hat{A} \ne A_0) \to 0$, then it follows from (25) that $B_0(u,l)$ cannot interact with a L_0-ball $B_0(R^*(u,\boldsymbol{X}), 0)$, thus $l \ge R^*(u,\boldsymbol{X})$ with $l = \frac{1}{4r(u,\boldsymbol{X})}\sigma^2\frac{\log p}{n}$, and $d_0 = \frac{1}{4r(u,\boldsymbol{X})}$, for any $\boldsymbol{\beta}^0 \in B_0(u,l)$. By (6), $r(u,\boldsymbol{X}) \ge 1$. Hence, $d_1 \le 1/4$. This completes the proof. □

Proof of Lemma 1 The first inequality follows from Lemma 3 of Shen et al. (2012). For the second, note that

$$\begin{aligned} C_{\min} &= n^{-1}\min_{A\ne A_0, |A|\le p_0} \frac{1}{\max(|A_0 \setminus A|, 1)}\|(I - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \\ &\le \min_{j\in A_0} n^{-1}\|(I - \boldsymbol{P}_{A_0\setminus\{j\}})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \le \min_{j\in A_0}\left(n^{-1}\|\boldsymbol{x}^{(j)}\|^2\beta_j^2\right) \\ &\le \gamma_{\min}^2 \max_{j\in A_0} n^{-1}\|\boldsymbol{x}^{(j)}\|^2. \end{aligned}$$

This together with $\max_{j\in A_0} n^{-1}\|\boldsymbol{x}^{(j)}\|^2 \le c_{\max}(n^{-1}\boldsymbol{X}_{A_0}^T\boldsymbol{X}_{A_0})$ implies that the desired result. This completes the proof. □

Next we present a technical lemma to be used below.

Lemma 4 *Let $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$ be two projection matrices onto the column space of $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$, respectively. For any integer $r \ge 2$,*

$$Tr((\boldsymbol{P}_A - \boldsymbol{P}_B)^r) \le Tr((\boldsymbol{P}_A - \boldsymbol{P}_B)^2) \le |A| + |B| - 2|A \cap B|, \tag{26}$$

where Tr denotes the trace of a matrix.

Proof Before proceeding, we prove that $0 \le \lambda_{\max}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2) \le 1$. Note that $(\boldsymbol{P}_A - \boldsymbol{P}_B)^2$ is non-negative definite. Then, for any $\boldsymbol{x}$, $0 \le ((\boldsymbol{P}_A - \boldsymbol{P}_B)\boldsymbol{x})^T((\boldsymbol{P}_A - \boldsymbol{P}_B)\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{P}_A - \boldsymbol{P}_B)^2\boldsymbol{x}$, implying that $\lambda_{\max}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2) = \sup_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^T(\boldsymbol{P}_A - \boldsymbol{P}_B)^2\boldsymbol{x}}{\|\boldsymbol{x}\|^2} \ge 0$, where $|A|$ denotes size of set A, and $\|\cdot\|$ is the usual L_2-norm. Moreover, $\boldsymbol{x}^T(\boldsymbol{P}_A - \boldsymbol{P}_B)^2\boldsymbol{x} = ((\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{x})^T(\boldsymbol{P}_B\boldsymbol{x}) + ((\boldsymbol{I} - \boldsymbol{P}_B)\boldsymbol{x})^T(\boldsymbol{P}_A\boldsymbol{x})$. By inequality that $2ab \le a^2 + b^2$ for any real numbers a, b, and the fact that $(\boldsymbol{I} - \boldsymbol{P}_A)^2 = (\boldsymbol{I} - \boldsymbol{P}_A)$ and $\boldsymbol{P}_B^2 = \boldsymbol{P}_B$, $((\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{x})^T(\boldsymbol{P}_B\boldsymbol{x}) \le \frac{1}{2}(\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{P}_A)^2\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{P}_B^2\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{P}_B\boldsymbol{x})$. Thus, $\boldsymbol{x}^T(\boldsymbol{P}_A - \boldsymbol{P}_B)^2\boldsymbol{x} \le \frac{1}{2}(\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{P}_B\boldsymbol{x} + \boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{P}_B)\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{P}_A\boldsymbol{x}) = \|\boldsymbol{x}\|^2$. Hence, $\lambda_{\max}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2) \le 1$.

For the first inequality in (26), first consider the case of even r. In this case, $(\boldsymbol{P}_A - \boldsymbol{P}_B)^r$ is non-negative definite. By Lemma 6.5 of Zhou et al. (1998), $\mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^r) \le \mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2)(\lambda_{\max}(\boldsymbol{P}_A - \boldsymbol{P}_B)^2)^{r/2-1} \le \mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2)$, for any integer $r \ge 3$. Next consider the case of odd valued r. Now $\mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^r) \le \mathrm{Tr}(P_A(\boldsymbol{P}_A - \boldsymbol{P}_B)^{r-1}) \le \mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^{r-1})$, which reduces to the case of even valued r.

To prove the second inequality in (26), note that $\mathrm{Tr}((\boldsymbol{P}_A - \boldsymbol{P}_B)^2) = |A| + |B| - 2Tr(\boldsymbol{P}_A\boldsymbol{P}_B)$. If $A \cap B = \emptyset$, $\mathrm{Tr}(\boldsymbol{P}_A\boldsymbol{P}_B) \ge \mathrm{Tr}(\boldsymbol{P}_A)\lambda_{\min}(\boldsymbol{P}_B) = 0$ by Lemma 6.5 of Zhou et al. (1998), implying the second inequality in (26). If $A \cap B \neq \emptyset$, we write, without loss of generality, $\boldsymbol{X}_A = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{|A|-s}, \ldots, \boldsymbol{x}_{|A|})$ and $\boldsymbol{X}_B = (\boldsymbol{x}_{|A|-s+1}, \ldots, \boldsymbol{x}_{|A|}, \ldots, \boldsymbol{x}_{|A|+|B|-s})$ with $s \le |A| \le |B|$ and $s = |A \cap B|$. Now we construct an orthonormal basis for the column space of $\boldsymbol{X}_{A\cap B}$: $\boldsymbol{e}_{|A|-s+1}, \ldots, \boldsymbol{e}_{|A|}$, followed by two orthonormal bases that are orthogonal to it through the Gram–Schmidt orthogonalization. These are $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{|A|-s}$ and $\boldsymbol{e}_{|A|+1}, \ldots, \boldsymbol{e}_{|A|+|B|-s}$, in the column spaces of $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$, respectively. As a result of the construction, $\boldsymbol{P}_A = \sum_{i=1}^{|A|} \boldsymbol{e}_i\boldsymbol{e}_i^T$ and $\boldsymbol{P}_B = \sum_{j=|A|-s+1}^{|A|+|B|-s} \boldsymbol{e}_j\boldsymbol{e}_j^T$. Consequently,

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{P}_A\boldsymbol{P}_B) &= \sum_{i=1}^{|A|}\sum_{j=|A|-s+1}^{|A|+|B|-s} \mathrm{Tr}(\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{e}_j\boldsymbol{e}_j^T) = \sum_{i=1}^{|A|}\sum_{j=|A|-s+1}^{|A|+|B|-s} (\boldsymbol{e}_i^T\boldsymbol{e}_j)^2 \\ &= \sum_{j=|A|-s+1}^{|A|} (\boldsymbol{e}_j^T\boldsymbol{e}_j)^2 + \sum_{i=1}^{|A|-s}\sum_{j=|A|+1}^{|A|+|B|-s} (\boldsymbol{e}_i^T\boldsymbol{e}_j)^2 \ge \sum_{j=|A|-s+1}^{|A|} 1 = s,\end{aligned}$$

yielding the second inequality in (26). This completes the proof. □

Proof of Theorem 2 We bound the reconstruction error directly. Note that $|\hat{A}^{L_0}| \le p_0$ when $K = p_0$. If $\hat{A}^{L_0} = A_0$ then $\hat{\boldsymbol{\beta}}^{L_0} = \hat{\boldsymbol{\beta}}^{ol}$. Let $S(\boldsymbol{\beta}) \equiv \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}_A\boldsymbol{\beta}_A\|^2$. Note that $A \subset \{1, \ldots, p\}$ can be partitioned into $(A \setminus A_0) \cup (A_0 \cap A)$. Then

$$\begin{aligned} I \equiv P(\hat{\boldsymbol{\beta}}^{L_0} \neq \hat{\boldsymbol{\beta}}^{ol}) &\le \sum_{A\subset\{1,\ldots,p\}, A\neq A_0, |A|\le p_0} P(S(\hat{\boldsymbol{\beta}}^{L_0}_{\hat{A}^{L_0}}) - S(\hat{\boldsymbol{\beta}}^{ol}_{A_0}) \le 0, \hat{A}^{L_0} = A) \\ &\le \sum_{k=0}^{p_0-1}\sum_{j=0}^{p_0-k} \binom{p-p_0}{j}\binom{p_0}{k} P(S(\hat{\boldsymbol{\beta}}^{L_0}) - S(\hat{\boldsymbol{\beta}}^{ol}) \le 0, B_{kj}), \end{aligned} \tag{27}$$

where $B_{kj} = \{\hat{A}^{L_0} = A, |A_0 \cap A| = k, |A \setminus A_0| = j\}$, and $\binom{n}{k}$ is the binomial coefficient indexed by n and k. On event B_{kj}, $\|\boldsymbol{Y} - \boldsymbol{X}_A \hat{\boldsymbol{\beta}}^{L_0}\|^2 \geq \|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{Y}\|^2$. Hence,

$$\begin{aligned} 2(S(\hat{\boldsymbol{\beta}}^{L_0}) - S(\hat{\boldsymbol{\beta}}^{ol})) &\geq \|(\boldsymbol{I} - \boldsymbol{P}_A)(\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \boldsymbol{\epsilon})\|^2 - \|(\boldsymbol{I} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}\|^2 \\ &= 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \boldsymbol{\epsilon}^T\left(\boldsymbol{P}_A - \boldsymbol{P}_{A_0}\right)\boldsymbol{\epsilon}. \end{aligned}$$

For any δ with $0 < \delta < 1$, and any A with $|A_0 \cap A| = k$ and $|A \setminus A_0| = j; k = 0, \ldots, p_0 - 1$, $j = 1, \ldots, p_0 - k$, $P(S(\hat{\boldsymbol{\beta}}^{L_0}_A) - S(\hat{\boldsymbol{\beta}}^{ol}_{A_0}) \leq 0, B_{kj})$ is upper bounded by

$$\begin{aligned} &P(\delta\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 + 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} \leq 0) \\ &\quad + P((1-\delta)\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \boldsymbol{\epsilon}^T(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon} \leq 0) \equiv I_1(B_{kj}) + I_2(B_{kj}). \end{aligned}$$

Let $L_1(A) \equiv -2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}$ and $L_2(A) \equiv \boldsymbol{\epsilon}^T(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}$, which follow $N(0, 4\sigma^2\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2)$ and a weighted χ^2-distribution, respectively. Let $b(A) = \|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2$. An application of Markov's inequality with the normal moment generating function yields that

$$I_1(B_{kj}) \leq E\exp\left(\frac{t_1 L_1(A)}{\sigma^2}\right)\exp\left(-\frac{\delta t_1 b(A)}{\sigma^2}\right) \leq \exp\left(\frac{2t_1^2 - \delta t_1}{\sigma^2} ni C_{\min}\right),$$

for any $0 < t_1 < 1/2$, where $i \equiv p_0 - k$, and $nC_{\min} \leq \frac{b(A)}{\max(|A_0 \setminus A|, 1)} = \frac{b(A)}{i}$ has been used in the last inequality with $|A_0 \setminus A| = p_0 - |A_0 \cap A| = p_0 - k$. For $I_2(B_{kj})$, it follows from Lemma 4 that the moment generating function $M(t)$ of $\boldsymbol{\epsilon}^T(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}/\sigma^2$ satisfies: $\log M(t) = \sum_{r=1}^{\infty}(2^{r-1}t^r/r)\mathrm{Tr}(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})^r \leq t(|A| - |A_0|) + \mathrm{Tr}(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})^2\sum_{r=2}^{\infty}(2^{r-1}t^r/r) \leq t(|A| - |A_0|) + t^2/(1-2t)\mathrm{Tr}(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})^2 \leq t(|A| - |A_0| + |A| + |A_0| - 2|A \cap A_0|) = 2t|A \setminus A_0|$, for $0 < t < 1/2$. Similarly, for any $0 < t_1 < 1/2$,

$$\begin{aligned} I_2(B_{kj}) &\leq E\exp\left(\frac{t_1 L_2(A)}{\sigma^2}\right)\exp\left(-\frac{b(A)(1-\delta)t_1}{\sigma^2}\right) \\ &\leq \exp\left(-\frac{(1-\delta)t_2 ni C_{\min}}{\sigma^2} + 2t_1 j\right). \end{aligned}$$

Consequently, from (27) and bounds for $I_1(B_{kj})$ and $I_2(B_{kj})$,

$$\begin{aligned} I \leq \sum_{k=0}^{p_0-1}\sum_{j=0}^{p_0-k}(I_1(B_{kj}) + I_2(B_{kj})) \leq \sum_{i=1}^{p_0}\sum_{j=0}^{i}\binom{p-p_0}{j}\binom{p_0}{p_0-i} \\ \left(\exp\left(\frac{2t_1^2 - \delta t_1}{\sigma^2} ni C_{\min}\right) + \exp\left(-\frac{(1-\delta)t_2}{\sigma^2} ni C_{\min} + 2t_2 j\right)\right). \end{aligned}$$

For simplification, choose $t_1 = \frac{1}{3}$ and $\delta = \frac{2t_1+1}{2} = \frac{5}{6}$ such that $\delta t_1 - 2t_1^2 = (1-\delta)t_1 = \frac{1}{18}$. Note that $\binom{a}{b} \le a^b$ and $\log(p-p_0)+\log p_0 \le \log(\frac{p^2}{4}) \le 2\log p - 1$. Then

$$
\begin{aligned}
I &\le 2\sum_{i=1}^{p_0}\sum_{j=0}^{i}(p-p_0)^j p_0^i \exp\left(-\frac{i}{18\sigma^2}nC_{\min}+\frac{2}{3}j\right)\\
&= 2\sum_{i=1}^{p_0}\exp\left(-i\left(\frac{nC_{\min}}{18\sigma^2}-\log p_0\right)\right)\sum_{j=0}^{i}\exp\left(j\left(\frac{2}{3}+\log(p-p_0)\right)\right)\\
&\le \frac{2}{1-e^{-1}}R\left(\exp\left(-\frac{n}{18\sigma^2}\left(C_{\min}-36\frac{\log p}{n}\sigma^2\right)\right)\right),
\end{aligned}
$$

where $R(x) = x/(1-x)$ is the exponentiated logit function. Using the fact that $I \le 1$, we obtain that $I \le (\frac{2}{1-e^{-1}}+1)\exp(-\frac{n}{18\sigma^2}(C_{\min}-36\frac{\log p}{n}\sigma^2))$, leading to (10). Finally an application of the pointwise bound in (10) to $\boldsymbol{\beta}^0 \in B_0(u,l)$ yields (11), implying consistency by $P(\hat{A}^{L_0} \ne A_0) \le P(\hat{\boldsymbol{\beta}}^{L_0} \ne \hat{\boldsymbol{\beta}}^{ol})$. The result in (A) is established.

For (B), we note that $n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{ol}-\boldsymbol{\beta}^0)\|^2 = \frac{p_0}{2n}$. Let $D = 25\sigma^2$ and $G = \{\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2 \ge D\}$. Then

$$
\frac{1}{n}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 = \frac{1}{n}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2(I(G)+I(G^c)) \equiv T_1+T_2.
$$

For T_1, note that $\frac{1}{4n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 - \frac{1}{2n}\|\boldsymbol{\epsilon}\|^2 \le \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{L_0}\|^2 \le \frac{1}{2n}\|\boldsymbol{\epsilon}\|^2$, and $T_1 = DP(\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 \ge D) + \int_D^\infty P(\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 \ge x)\,\mathrm{d}x$. For any $x > 0$, by Markov's inequality with $t = \frac{1}{3}$,

$$
\begin{aligned}
&\int_D^\infty P\left(\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 \ge x\right)\mathrm{d}x\\
&\le \int_D^\infty P\left(\frac{1}{n}\|\boldsymbol{\epsilon}\|^2 \ge \frac{x}{4}\right)\mathrm{d}x \le \int_D^\infty E\exp\left(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\right)\exp\left(-nt\frac{x}{4\sigma^2}\right)\mathrm{d}x\\
&\le \int_D^\infty \exp\left(-\frac{nt}{12\sigma^2}(x-24\sigma^2)\right)\mathrm{d}x = \frac{12\sigma^2}{nt}\exp\left(-\frac{n}{12}\right) = o\left(\frac{p_0}{2n}\right).
\end{aligned}
$$

Similarly, $DP(\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{L_0}-\boldsymbol{\beta}^0)\|^2 \ge D) \le 25\sigma^2\exp(-\frac{nt}{12\sigma^2}(D-24\sigma^2)) = o(\frac{p_0}{2n})$. Hence, $T_1 = o(\frac{p_0}{2n})$. For T_2, note that

$$
\begin{aligned}
T_2 &\le DP(\hat{\boldsymbol{\beta}}^{L_0} \ne \hat{\boldsymbol{\beta}}^{ol}) + \frac{1}{n}E\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\\
&= 25\sigma^2 P(\hat{\boldsymbol{\beta}}^{L_0} \ne \hat{\boldsymbol{\beta}}^{ol}) + \frac{p_0}{2n} = (o(1)+1)\frac{p_0}{2n},
\end{aligned}
$$

implying the risk result.

For minimaxity, note that

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^0 \in B_0(u,l)} n^{-1} E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|^2 \geq \inf_{\hat{\boldsymbol{\beta}}_{A_0}} \sup_{\boldsymbol{\beta}^0_{A_0} \in \mathcal{B}} n^{-1} E\|\boldsymbol{X}_{A_0}(\hat{\boldsymbol{\beta}}_{A_0} - \boldsymbol{\beta}^0_{A_0})\|^2,$$

where $\mathcal{B} = \{\boldsymbol{\beta}_{A_0} : |A_0| = u, n^{-1}\|\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} - \boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \geq l\}$. The result follows from the same argument as that for the least squares estimate to be minimax, c.f., Judge and Bock (1978). This completes the proof. □

Proof of Theorem 3 Our strategy is similar to that in the Proof of Theorem 2. Let $S(\hat{\boldsymbol{\beta}}^T) \equiv \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^T_{A_1} - \boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^T_{A_2}\|^2$, $A = A_1 \cup A_2$, $A_1 = \{j \in A : |\hat{\beta}^T_j| > \tau\}$, $A_2 = \{j \in A : |\hat{\beta}^T_j| \leq \tau\}$ and $\|\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^T_{A_2}\|^2 \leq c_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau \sum_{j\in A_2}|\hat{\beta}^T_j|$. Note that $|A_1| + \frac{1}{\tau}\sum_{j\in|A_2|}|\hat{\beta}^T_j| \leq p_0$. Thus, if $A_1 = A_0$ then $\hat{\beta}^T_j = 0$ for all $j \in A_2$, implying that $\hat{\boldsymbol{\beta}}^T = \hat{\boldsymbol{\beta}}^{ol}$. Therefore, we only consider the case of $A_1 \neq A_0$.

Similarly, let $B_{kj} = \{\hat{A} = A : |A_0 \cap A_1| = k, |A_1 \setminus A_0| = j\}$, then $I \equiv P(\hat{\boldsymbol{\beta}}^T_{\hat{A}} \neq \hat{\boldsymbol{\beta}}^{ol}_{A_0}) \leq \sum_{k=0}^{p_0-1}\sum_{j=0}^{p_0-k} P(S(\hat{\boldsymbol{\beta}}^T_A) - S(\hat{\boldsymbol{\beta}}^{ol}_{A_0}) \leq 0, B_{kj})$. On B_{kj}, we simplify $S(\hat{\boldsymbol{\beta}}^T_A) - S(\hat{\boldsymbol{\beta}}^{ol}_{A_0})$. An application of inequality $\|\boldsymbol{U} - \boldsymbol{V}\|^2 \geq \frac{a-1}{a}\|\boldsymbol{U}\|^2 - (a-1)\|\boldsymbol{V}\|^2$ for $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^p$ and some $a > 1$, together with the fact that $\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^T_{A_1}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^{ol}_{A_1}\|^2$ yields that

$$\begin{aligned}
S(\hat{\boldsymbol{\beta}}^T) &\geq \frac{a-1}{2a}\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^T_{A_1}\|^2 - \frac{a-1}{2}\|\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^T_{A_2}\|^2 \\
&\geq \frac{a-1}{2a}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} + (\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{\epsilon}\|^2 - \frac{a-1}{2}pc_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau^2 \\
&\geq \frac{a-1}{a}\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} + \frac{a-1}{2a}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 \\
&\quad + \frac{a-1}{2a}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{\epsilon}\|^2 - \frac{a-1}{2}pc_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau^2.
\end{aligned}$$

Let $\lambda = \frac{a-1}{2}pc_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau^2$. Then

$$\begin{aligned}
&2(S(\hat{\boldsymbol{\beta}}^T) - S(\hat{\boldsymbol{\beta}}^{ol})) \geq 2(S(\hat{\boldsymbol{\beta}}^T) - \frac{1}{2}\|(I - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}\|^2) \\
&= 2((a-1)/a)\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} + ((a-1)/a)\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 \\
&\quad -\boldsymbol{\epsilon}^T(\boldsymbol{I} + (a-1)\boldsymbol{P}_{A_1} - a\boldsymbol{P}_{A_0})\boldsymbol{\epsilon}/a - 2\lambda \\
&= -\frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}) \\
&\quad +(a-1)\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \boldsymbol{\epsilon}^T(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon} - 2\lambda
\end{aligned}$$

For any $0 < \delta < 1$, let $b_1(A_1) = (a - 1 - \delta)\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2$, $b_2(A_1) = \delta\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - 2\lambda$, $L_1(A_1) = \frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})$

$(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})$, $L_2(A_1) = \boldsymbol{\epsilon}^T g(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}$. Note that $aL_1(A_1)$ follows $\sigma^2\chi^2_k$, where the non-central χ^2_k distribution has degrees of freedom $n - \min(r(A_1), n)$ with $r(A_1) \le |A_1|$ being the rank of A_1, and a non-central parameter $(a-1)^2\sigma^{-2}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2$. Hence,

$$\begin{aligned} &P(S(\hat{\boldsymbol{\beta}}^T_A) - S(\hat{\boldsymbol{\beta}}^{ol}_{A_0}) \le 0, B_{kj}) \\ &\quad \le P(L_1(A_1) \ge b_1(A_1)) + P(L_2(A_1) \ge b_2(A_1)) \equiv I_1(B_{kj}) + I_2(B_{kj}), \end{aligned}$$

where

$$\begin{aligned} I_1(B_{kj}) &\le E\exp\left(\frac{t_1}{\sigma^2}L_1(A_1)\right)\exp\left(-\frac{t_1}{\sigma^2}b_1(A_1)\right) \\ &= \frac{1}{(1-2t_1/a)^{\frac{n-r(A_1)}{2}}}\exp\left(-\frac{t_1(1-2t_1-\delta)}{\sigma^2(1+(1-2t_1)/(a-1))}niC_{\min}\right) \\ I_2(B_{kj}) &\le E\exp\left(\frac{t_1}{\sigma^2}L_2(A_1)\right)\exp\left(-\frac{t_1}{\sigma^2}b_2(A_1)\right) \\ &\le \exp\left(-\frac{\delta t_1}{\sigma^2}niC_{\min} + 2t_1 j + 2t_1\lambda/\sigma^2\right), \end{aligned}$$

for any $0 < t_1 < 1/2$, where the last inequality uses $nC_{\min} \le \frac{b(A)}{|A_0 \setminus A|}$ with $|A_0 \setminus A| = p_0 - |A_0 \cap A| = p_0 - k \equiv i$. Consequently,

$$\begin{aligned} I \le \sum_{i=1}^{p_0}\sum_{j=0}^{i}\binom{p-p_0}{j}\binom{p_0}{p_0-i}\Bigg(&\exp\left(-\frac{\delta t_1}{\sigma^2}niC_{\min} + 2t_1 j + t_1\lambda/\sigma^2\right) \\ &+ \frac{1}{(1-2t_1/a)^{\frac{n-r(A_1)}{2}}}\exp\left(-\frac{t_1(1-2t_1-\delta)}{\sigma^2(1+(1-2t_1)/(a-1))}niC_{\min}\right)\Bigg). \end{aligned}$$

To simplify this bound, choose $t_1 = \frac{1}{3}$, $\delta = \frac{1}{6}$, $a = n+1$ and $\lambda \le \sigma^2$. Similarly,

$$\begin{aligned} I &\le 2\sum_{i=1}^{p_0}\sum_{j=0}^{i}(p-p_0)^j p_0^i \exp\left(-\frac{i}{20\sigma^2}nC_{\min} + \frac{2}{3}j + \frac{1}{3}\right) \\ &= 2\sum_{i=1}^{p_0}\exp\left(-i\left(\frac{nC_{\min}}{20\sigma^2} - \log p_0\right)\right)\sum_{j=0}^{i}\exp(j(1+\log(p-p_0))) \\ &\le \left(\frac{2}{e-1}+1\right)R\left(\exp\left(-\frac{n}{20\sigma^2}\left(C_{\min} - 40\frac{\log p}{n}\sigma^2\right)\right)\right), \end{aligned}$$

yielding (14). The rest of the results follow similarly as in the Proof of Theorem 2. This completes the proof. □

Proof of Lemma 2 The finite termination property of Algorithm 1 follows from non-increasingness of $S^{(m)}(\hat{\boldsymbol{\beta}}^{(m)})$ in m, as in the Proof of Theorem 1 of Shen et al. (2012).

Now consider the DC solution of (22) $\hat{\boldsymbol{\beta}}^{(m)}$ at iteration m for given $K > 0$. Let the termination index be m^*. Then Karush–Kuhn–Tucker conditions imply that there exists a Lagrange multiplier $\lambda \geq 0$ such that the DC solution of (22) $\hat{\boldsymbol{\beta}}^{(m^*)}$ minimizes the Lagrange function $L(\boldsymbol{\beta}, \lambda) = S(\boldsymbol{\beta}) - \lambda(K - \sum_{j=1}^{p} |\beta_j| I(\hat{\beta}_j^{(m^*-1)}| > \tau))$, or equivalently. $\bar{S}(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| I(\hat{\beta}_j^{(m^*-1)}| > \tau)$, with respect to $\boldsymbol{\beta}$. By Theorem 1 of Shen et al. (2012), $\hat{\boldsymbol{\beta}}^{(m^*)} = \hat{\boldsymbol{\beta}}^{(m^*-1)}$ at termination. Consequently, $\bar{S}(\hat{\boldsymbol{\beta}}^{(m^*)}) = \bar{S}(\hat{\boldsymbol{\beta}}^{(m^*-1)})$. This means that if Algorithm 1 of Shen et al. (2012) is initialized with $\hat{\boldsymbol{\beta}}^{(m^*)}$ then $\hat{\boldsymbol{\beta}}^{(m^*)}$ is also a DC solution of (21) with respect to λ.

Conversely, for the solution of (22), the case of $\lambda = 0$ is trivial and is thus omitted. Now for given $\lambda > 0$ and a DC solution of (22) $\hat{\boldsymbol{\beta}}^{(m_0)}$ at iteration m, define $K^{(m_0)} = \frac{1}{\tau} \sum_{j=1}^{p} |\hat{\beta}_j^{(m_0)}| I(|\hat{\beta}_j^{(m_0-1)}| \leq \tau)$, where m_0 is the termination index of the unconstrained Algorithm 1 of Shen et al. (2012), which is assured by Theorem 1 of Shen et al. (2012). Hence, a DC solution $\hat{\boldsymbol{\beta}}^{(m_0)}$ of (22) is also a solution of (21) by checking Karush–Kuhn–Tucker conditions for the constrained problem with $K^{(m_0)}$. Similarly, if Algorithm 1 is initialized by $\hat{\boldsymbol{\beta}}^{(m_0)}$, then $\hat{\boldsymbol{\beta}}^{(m_0)}$ is also a DC solution of (21). This is because $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m_0)}$ for $m \geq m_0$. This completes the proof.

Proof of Proposition 1 It suffices to minimize componentwisely: $\hat{\beta}_j = \arg\min_{\beta_j} f_j(\beta_j)$, with $f(\beta_j) = \frac{1}{2}(\beta_j - \hat{\beta}_j^{ol})^2 + \lambda \min(\frac{|\beta_j|}{\tau}, 1)$; $j = 1, \ldots, p$. If $|\beta_j| \leq \tau$, $\hat{\beta}_j = (|\hat{\beta}_j^{ol}| - \frac{\lambda}{\tau})_+ \text{sign}(\hat{\beta}_j^{ol})$, otherwise, $\min_{|\beta_j|>\tau} f(\beta_j) = \lambda$ if $\beta_j = \hat{\beta}_j^{ol}$. Moreover, $\min_{\{\beta_j : |\beta_j| \leq \tau\}} f(\beta_j)$ is

$$\begin{cases} f(0) = \frac{1}{2}(\hat{\beta}_j^{ol})^2 & \text{when} \quad |\hat{\beta}_j^{ol}| \leq \frac{\lambda}{\tau}; \\ f(\text{sign}(\hat{\beta}_j^{ol})\tau) = \frac{1}{2}(\tau - |\hat{\beta}_j^{ol}|)^2 + \lambda & \text{when} \quad |\hat{\beta}_j^{ol}| \geq \frac{\lambda}{\tau} + \tau; \\ f\left((|\hat{\beta}_j^{ol}| - \frac{\lambda}{\tau})\text{sign}(\hat{\beta}_j^{ol})\right) = \frac{\lambda}{\tau}|\hat{\beta}_j^{ol}| - \frac{\lambda^2}{2\tau^2} & \text{when} \quad \frac{\lambda}{\tau} < |\hat{\beta}_j^{ol}| < \frac{\lambda}{\tau} + \tau; \end{cases}$$

Then comparing f at 0, τ, $(|\hat{\beta}_j^{ol}| - \frac{\lambda}{\tau})_+ \text{sign}(\hat{\beta}_j^{ol})$ against f at λ, the TLP estimate is

$$\begin{cases} \hat{\beta}_j^{ol} & \text{if} \quad |\hat{\beta}_j^{ol}| \geq \max(\frac{\lambda}{2\tau} + \tau, \frac{\lambda}{\tau}), \text{ or } \frac{\lambda}{\tau} \geq |\hat{\beta}_j^{ol}| \geq \max(\sqrt{2\lambda}, \tau); \\ (|\hat{\beta}_j^{ol}| - \frac{\lambda}{\tau})\text{sign}(\hat{\beta}_j^{ol}) & \text{if} \quad \frac{\lambda}{2\tau} + \tau \geq |\hat{\beta}_j^{ol}| \geq \max(\frac{\lambda}{\tau}, \tau), \text{ or } \tau \geq |\hat{\beta}_j^{ol}| \geq \frac{\lambda}{\tau}; \\ 0 & \text{if} \quad \min(\frac{\lambda}{\tau}, \tau) \geq |\hat{\beta}_j^{ol}|, \text{ or } \min(\sqrt{2\lambda}, \frac{\lambda}{\tau}) \geq |\hat{\beta}_j^{ol}| \geq \tau; \end{cases}$$

$j = 1, \ldots, p$, leading to the desired result. This completes the proof. □

Proof of Theorem 5 We only present the proof for the case where $\alpha > 1$. The proof for the case $\alpha = 1$ is similar, thus omitted. Write $\hat{A}^{tl} = \hat{A}_1 \cup \hat{A}_2$, $\hat{A}_1 = \{j \in \hat{A}^{tl} : |\hat{\beta}_j^{tl}| > \tau\}$ and $\hat{A}_2 = \{j \in \hat{A}^{tl} : |\hat{\beta}_j^{tl}| \leq \tau\}$. Then $P(\hat{\boldsymbol{\beta}}^{tl} \neq \hat{\boldsymbol{\beta}}^{ol}) \leq I_1 + I_2 + P(\hat{\boldsymbol{\beta}}^{ol}$ is not a solution of (23)), where the last term in this inequality is bounded by I_6 in the Proof of

Theorem 4. Thus, it suffices to bound $I_1 = P(\cup_{A_1 \subseteq \{1,\ldots,p\}: A_1 \neq A_0}(S(\hat{\boldsymbol{\beta}}^{tl}) - S(\tilde{\boldsymbol{\beta}}^{ol}) \leq 0, \hat{A}_1 = A_1))$; $I_2 = P(\hat{A}_1 = A_0, \hat{\boldsymbol{\beta}}^{tl} \neq \hat{\boldsymbol{\beta}}^{ol}, \hat{\boldsymbol{\beta}}^{ol}$ is a solution of (23)).

For I_1, on $\hat{A}_1 = A_1$ with $A_1 \neq A_0$, write $S(\hat{\boldsymbol{\beta}}^{tl})$ as $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^{tl}_{A_1} - \boldsymbol{X}_{\hat{A}_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2}\|^2 + \frac{\lambda}{\tau}\sum_{j\in\hat{A}_2}|\hat{\beta}^{tl}_j| + \lambda|A_1|$. Note that $\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^{tl}_{A_1}\|^2 \geq \|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{Y}\|^2$, and $\|\boldsymbol{X}_{\hat{A}_2}\hat{\boldsymbol{\beta}}^{tl}_{\hat{A}_2}\|^2 \leq c_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau\sum_{j\in\hat{A}_2}|\hat{\beta}^{tl}_j|$, and $\frac{\lambda}{\tau} - \frac{a}{2}c_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau \geq 0$ (by assumption) with real $a > 1$ to be chosen. Using $\|U - V\|^2 \geq \frac{a-1}{a}\|U\|^2 - (a-1)\|V\|^2$ for any vectors $U, V \in \mathcal{R}^n$,

$$
\begin{aligned}
S(\hat{\boldsymbol{\beta}}^{tl}) - \lambda|A_1| &\geq \frac{a-1}{2a}\|\boldsymbol{Y} - \boldsymbol{X}_{A_1}\hat{\boldsymbol{\beta}}^{tl}_{A_1}\|^2 - \frac{a-1}{2}\|\boldsymbol{X}_{\hat{A}_2}\hat{\boldsymbol{\beta}}^{tl}_{\hat{A}_2}\|^2 + \frac{\lambda}{\tau}\sum_{j\in\hat{A}_2}|\hat{\beta}^{tl}_j| \\
&\geq \left(\frac{a-1}{2a}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{Y}\|^2\right) + \left(\frac{\lambda}{\tau} - \frac{a-1}{2}c_{\max}(\boldsymbol{X}^T\boldsymbol{X})\tau\right)\sum_{j\in\hat{A}_2}|\hat{\beta}^{tl}_j| \\
&\geq \left(\frac{a-1}{2a}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + (\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{\epsilon}\|^2\right).
\end{aligned}
$$

So $2(S(\hat{\boldsymbol{\beta}}^{tl}) - S(\hat{\boldsymbol{\beta}}^{ol})) \geq -\frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}) + (a-1)\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \boldsymbol{\epsilon}^T(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon} + 2\lambda(|A_1| - p_0)$.

Note that $I_1 \leq \sum_{k=0}^{p_0-1}\sum_{j=0}^{p-k}\binom{p_0}{k}\binom{p-p_0}{j}P(S(\hat{\boldsymbol{\beta}}^{tl}) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, B_{kj})$, where $B_{kj} = \{\hat{A}_1 = A_1 \neq A_0 : |A_1 \setminus A_0| = j, |A_1 \cap A_0| = k\}$. For any $0 < \delta < 1$, let $b^1_{A_1} = (a - 1 - \delta)\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 + \lambda(|A_1| - p_0)$, $b^2_{A_1} = \delta\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 + \lambda(|A_1| - p_0)$, $L^1_{A_1} = \frac{1}{a}(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})^T(\boldsymbol{I} - \boldsymbol{P}_{A_1})(\boldsymbol{\epsilon} - (a-1)(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0})$, $L^2_{A_1} = \boldsymbol{\epsilon}^T(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}$. Note that $aL^1_{A_1}$ follows $\sigma^2\chi^2_k$, where the non-central χ^2_k distribution has degrees of freedom $n - \min(r(A_1), n)$ with $r(A_1) \leq |A_1|$ being the rank of A_1, and a non-central parameter $(a-1)^2\sigma^{-2}\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2$. For $L^2_{A_1}$, it follows from Lemma 4 that the moment generating function $M(t)$ of $L^2_{A_1}$ satisfies: $\log M(t) = \sum_{r=1}^{\infty}(2^{r-1}t^r/r)\mathrm{Tr}(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})^r \leq t(|A_1| - |A_0|) + \mathrm{Tr}(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})^2\sum_{r=2}^{\infty}(2^{r-1}t^r/r) \leq t(|A_1| - |A_0|) + t^2/(1-2t)\mathrm{Tr}(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})^2 \leq t(|A| - |A_0| + |A_1| + |A_0| - 2|A_1 \cap A_0|) = 2t|A_1 \setminus A_0|$, for $0 < t < 1/2$. Let $I^1_{kj} = P(L^1_{A_1} \geq b^1_{A_1})$ and $I^2_{kj} = P(L^2_{A_1} \geq b^2_{A_1})$. Hence, $P(S(\hat{\boldsymbol{\beta}}^{tl}) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, B_{kj}) \leq I^1_{kj} + I^2_{kj}$.

For $I^l_{kj}; l = 1, 2$, note that $\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \geq niC_{\min}$ if $j \leq [\alpha i] \equiv \alpha(p_0 - k)$ by definition of $C_{\min}$ or if $|A_1 \setminus A_0| \leq \alpha|A_0 \setminus A_1|(|A_1| + (\alpha - 1)|A_1 \cap A_0| \leq \alpha p_0)$ with $|A_0 \setminus A_1| = p_0 - |A_0 \cap A_1| = i$; or 0 if $[\alpha i] < j \leq p$. By Markov's inequality,

$$
\begin{aligned}
I^1_{kj} &\leq E\exp\left(\frac{t_1}{\sigma^2}L_1(A_1)\right)\exp\left(-\frac{t_1}{\sigma^2}b_1(A_1)\right) \\
&= \frac{1}{(1-2t_1/a)^{\frac{n-r(A_1)}{2}}}\exp\left(-\frac{t_1(1-2t_1-\delta)}{\sigma^2(1 + (1-2t_1)/(a-1))}niC_{\min} + t_1\lambda(i-j)/\sigma^2\right)
\end{aligned}
$$

$$I_{kj}^2 \le E\exp\left(\frac{t_1}{\sigma^2}L_2(A_1)\right)\exp\left(-\frac{t_1}{\sigma^2}b_2(A_1)\right)$$
$$\le \exp\left(-\frac{\delta t_1}{\sigma^2}niC_{\min}+2t_1 j+t_1\lambda(i-j)/\sigma^2\right),$$

for any $0 < t_1 < 1/2$. Therefore, $I_1 \le \sum_{k=0}^{p_0-1}\sum_{j=0}^{p-k}\binom{p-p_0}{j}\binom{p_0}{k}(I_{kj}^1+I_{kj}^2)$, which is bounded by

$$\sum_{i=1}^{p_0}\sum_{j=0}^{[\alpha i]}\binom{p-p_0}{j}\binom{p_0}{p_0-i}\left(\exp\left(-\frac{\delta t_1}{\sigma^2}niC_{\min}+2t_1 j+t_1\lambda(i-j)/\sigma^2\right)\right.$$
$$\left.+\frac{1}{(1-2t_1/a)^{\frac{n-r(A_1)}{2}}}\exp\left(-\frac{t_1(1-2t_1-\delta)}{\sigma^2(1+(1-2t_1)/(a-1))}niC_{\min}+t_1\lambda(i-j)/\sigma^2\right)\right)$$
$$+\sum_{i=1}^{p_0}\sum_{j=[\alpha i]+1}^{p}\binom{p-p_0}{j}\binom{p_0}{p_0-i}\left(\exp\left(2t_1 j+t_1\lambda(i-j)/\sigma^2\right)\right.$$
$$\left.+\frac{1}{(1-2t_1/a)^{\frac{n-r(A_1)}{2}}}\exp\left(t_1\lambda(i-j)/\sigma^2\right)\right).$$

To simplify this bound, choose $t_1=\frac{1}{3}$, $\delta=\frac{1}{6}$, $a=n+1$. Note that $\sum_{j=0}^{b}\binom{a}{j}\le(a+1)^b$, and $\binom{a}{b}\le a^b$, for any integers $a, b > 0$. Then

$$I_1 \le 2\sum_{i=1}^{p_0}p_0^i\sum_{j=0}^{[\alpha i]}\binom{p-p_0}{j}\exp\left(-\frac{inC_{\min}}{20\sigma^2}+\frac{2j}{3}+\frac{(i-j)\lambda}{2\sigma^2}\right)$$
$$+2\sum_{j=[\alpha i]+1}^{p}(p-p_0)^j\exp\left(\frac{-(\alpha-1)j\lambda}{3\alpha\sigma^2}+\frac{2j}{3}\right)\sum_{i=0}^{[j/\alpha]}\binom{p_0}{i}$$
$$\le 2\sum_{i=1}^{p_0}\exp\left(-i\left(\frac{nC_{\min}}{20\sigma^2}-\log p_0-\alpha\log(p-p_0+1)-\frac{\lambda}{2\sigma^2}\right)\right)$$
$$+2\sum_{j=[\alpha i]+1}^{p}\exp\left(-j\left(\frac{(\alpha-1)\lambda}{3\alpha\sigma^2}-\log(p-p_0)-\frac{1}{\alpha}\log(p_0+1)-\frac{2}{3}\right)\right).$$

Using the fact that $I_1 \le 1$, $\log p_0+\alpha\log(p-p_0+1)\le(\alpha+1)(\log(p+1)-\log(\alpha+1)+\frac{\alpha}{\alpha+1}\log\alpha)$ and $\frac{1}{\alpha}\log(p_0+1)+\log(p-p_0)\le(1+\frac{1}{\alpha})(\log(p+1)-\frac{1}{\alpha+1}\log(\alpha+1)-\frac{\alpha}{\alpha+1}\log(1+\frac{1}{\alpha}))$, we obtain the second and third terms in the bound of (17).

For I_2, let $E=\{\min_{j\in A_0}|\hat\beta_j^{ol}|>\tau\}$. As in the Proof of Theorem 6, $P(E^c)\le|A_0|(\Phi(-\frac{n^{1/2}(\gamma_{\min}-\tau)}{\sigma c_{\min}^{-1/2}(\frac{1}{n}\boldsymbol{X}_{A_0}^T\boldsymbol{X}_{A_0})})-\Phi(-\frac{n^{1/2}(\gamma_{\min}+\tau)}{\sigma c_{\min}^{-1/2}(\frac{1}{n}\boldsymbol{X}_{A_0}^T\boldsymbol{X}_{A_0})}))$. On event E, $\hat{\boldsymbol{\beta}}_A^{ol}$ and $\hat{\boldsymbol{\beta}}_A^{tl}$ must be local minimizers of $\min_{\boldsymbol{\beta}_A}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}_A\boldsymbol{\beta}_A\|^2+\frac{\lambda}{\tau}\sum_{j\in\hat A_2}|\beta_j|$. Note that for any local solution $\hat{\boldsymbol{\beta}}_A$ satisfying $|\hat{\boldsymbol{\beta}}_{\bar A_0}|>\tau$ with $A_0\subset\bar A_0\subset A$ and $|\hat{\boldsymbol{\beta}}_{A_2}|\le\tau$ with $A_2=A\backslash\bar A_0$,

the local optimality condition for $\boldsymbol{\beta}^{tl}_{\bar{A}_0}$ is $(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}_{\bar{A}_0}\hat{\boldsymbol{\beta}}^{tl}_{\bar{A}_0}-\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2})=0$ for $j\in\bar{A}_0$, implying $\hat{\boldsymbol{\beta}}^{tl}_{\bar{A}_0}=(\boldsymbol{X}^T_{\bar{A}_0}\boldsymbol{X}_{\bar{A}_0})^{-1}\boldsymbol{X}^T_{\bar{A}_0}(\boldsymbol{Y}-\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2})$. This together with that for $\boldsymbol{\beta}_{A_2}$: $-(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}_{\bar{A}_0}\hat{\boldsymbol{\beta}}^{tl}_{\bar{A}_0}-\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2})+\frac{\lambda}{\tau}\text{sign}(\hat{\beta}^{tl}_j)=0$ for $j\in A_2$, implies that $-\boldsymbol{X}^T_{A_2}(\boldsymbol{I}-\boldsymbol{P}_{\bar{A}_0})(\boldsymbol{Y}-\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2})+\frac{\lambda}{\tau}\text{sign}(\hat{\boldsymbol{\beta}}^{tl}_{A_2})=0$ that is the local optimality for (28). Hence, both $\hat{\boldsymbol{\beta}}^{ol}_{A_2}$ and $\hat{\boldsymbol{\beta}}^{tl}_{A_2}$ are local minimizers of

$$\min_{\beta_{A_2}}\frac{1}{2}\|(\boldsymbol{I}-\boldsymbol{P}_{\bar{A}_0})\boldsymbol{Y}-(\boldsymbol{I}-\boldsymbol{P}_{\bar{A}_0})\boldsymbol{X}_{A_2}\boldsymbol{\beta}_{A_2}\|^2+\frac{\lambda}{\tau}\sum_{j\in A_2}|\beta_j|. \tag{28}$$

By Rainaldo (2007), $(\boldsymbol{I}-\boldsymbol{P}_{\bar{A}_0})\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{tl}_{A_2}=(\boldsymbol{I}-\boldsymbol{P}_{\bar{A}_0})\boldsymbol{X}_{A_2}\hat{\boldsymbol{\beta}}^{ol}_{A_2}$ and $\|\hat{\boldsymbol{\beta}}^{tl}_{A_2}\|_1=\|\hat{\boldsymbol{\beta}}^{ol}_{A_2}\|_1=0$. Thus, $\hat{\boldsymbol{\beta}}^{tl}=\hat{\boldsymbol{\beta}}^{ol}$ on E, implying that $I_2\le P(E^c)$. Combining the above bounds yields (17).

For (B), let $D=2C_{\min}+4\sigma^2$ and $G=\{\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\ge D\}$. Then

$$\frac{1}{n}E\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2=\frac{1}{n}E\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2(I(G)+I(G^c))\equiv T_1+T_2.$$

For T_1, note that $\frac{1}{4n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2-\frac{1}{2n}\|\boldsymbol{\epsilon}\|^2\le\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}\|^2\le\frac{1}{2n}\|\boldsymbol{\epsilon}\|^2+\frac{\lambda}{n}p_0$. Then for any $x>0$, $\{\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\ge x\}\subseteq\{\frac{x}{4}-\frac{\lambda}{n}p_0\le\frac{1}{n}\|\boldsymbol{\epsilon}\|^2\}\subseteq\{\frac{x}{4}-\frac{C_{\min}}{8}\le\frac{1}{n}\|\boldsymbol{\epsilon}\|^2\}$. By Markov's inequality with $t=\frac{1}{3}$, $T_1=DP(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\ge D)+\int_D^\infty P(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\ge x)\,\mathrm{d}x$. Note that the second term is upper bounded by

$$\begin{aligned}\int_D^\infty P\left(\frac{1}{n}\|\boldsymbol{\epsilon}\|^2\ge\frac{x}{4}-\frac{C_{\min}}{8}\right)\mathrm{d}x&\le\int_D^\infty E\exp\left(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\right)\exp\left(-nt\frac{(x-2C_{\min})}{8\sigma^2}\right)\mathrm{d}x\\&\le\int_D^\infty\exp\left(-\frac{nt}{\sigma^2}(x-2C_{\min}-\frac{\sigma^2}{1-2t})\right)\mathrm{d}x=o(\frac{p_0}{n}\sigma^2),\end{aligned}$$

so is $DP(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{tl}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2\ge D)$, implying that $T_1=o(\frac{p_0}{n}\sigma^2)$. For T_2, note that $C_{\min}\le\frac{1}{n}\|\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2$. Then $T_2\le DP(\hat{\boldsymbol{\beta}}^{tl}\ne\hat{\boldsymbol{\beta}}^{ol})+\frac{1}{n}E\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol}-\boldsymbol{X}\boldsymbol{\beta}^0\|^2$.

$$=\left(\frac{2}{n}\|\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2+4\sigma^2\right)P(\hat{\boldsymbol{\beta}}^{tl}\ne\hat{\boldsymbol{\beta}}^{ol})+\frac{p_0}{n}\sigma^2=(o(1)+1)\frac{p_0}{n}\sigma^2.$$

The desired result follows from the assumption on $\frac{1}{n}\|\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2$, (17) and (3).

For minimaxity, note that

$$\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}^0\in B_0(u,l)}n^{-1}E\|\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|^2\ge\inf_{\hat{\boldsymbol{\beta}}_{A_0}}\sup_{\boldsymbol{\beta}^0_{A_0}\in\mathcal{B}}n^{-1}E\|\boldsymbol{X}_{A_0}(\hat{\boldsymbol{\beta}}_{A_0}-\boldsymbol{\beta}^0_{A_0})\|^2,$$

where $\mathcal{B} = \{\boldsymbol{\beta}_{A_0} : |A_0| = u, n^{-1}\|\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} - \boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \geq l\}$. The result follows from the same argument as that for the least squares estimate to be minimax, c.f., Judge and Bock (1978). The proof for the case when $\alpha = 1$ is similar, thus omitted. □

Proof of Theorem 4 The proof is similar to that of Theorem 5 with some minor modifications. In the present case, no decomposition of $\hat{A}$ is necessary.

Note that $S(\hat{\boldsymbol{\beta}}^{l_0}) - \lambda|A| \geq \boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \frac{1}{2}\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 + \frac{1}{2}\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{\epsilon}\|^2$. So $2(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol})) \geq 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0} + \|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 - \boldsymbol{\epsilon}^T(\boldsymbol{P}_{A_1} - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon} + 2\lambda(|A| - p_0)$. Let $b^1_A = \delta\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \lambda(|A| - p_0)$, $b^2_A = (1-\delta)\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \lambda(|A| - p_0)$, $L^1_A = -2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}$ and $L^2_A = \boldsymbol{\epsilon}^T(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon}$. Note that L^1_A follows $N(0, 4\sigma^2\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2)$. Hence, for any δ with $0 < \delta < 1$,

$$
\begin{aligned}
&P(S(\hat{\boldsymbol{\beta}}^{l_0}) - S(\hat{\boldsymbol{\beta}}^{ol}) \leq 0, B_{kj}) \\
&\quad \leq P(\delta\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 + 2\boldsymbol{\epsilon}^T(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0} + \lambda(|A| - p_0) \leq 0) \\
&\qquad + P((1-\delta)\|(\boldsymbol{I} - \boldsymbol{P}_A)\boldsymbol{X}_{A_0}\boldsymbol{\beta}_{A_0}\|^2 - \boldsymbol{\epsilon}^T(\boldsymbol{P}_A - \boldsymbol{P}_{A_0})\boldsymbol{\epsilon} + \lambda(|A| - p_0) \leq 0) \equiv I^1_{kj} + I^2_{kj},
\end{aligned}
$$

where $I^l_{kj} \leq E\exp(\frac{t_l}{\sigma^2}L^l_A)\exp(-\frac{t_l}{\sigma^2}b^l_A)$; $0 < t_l < 1/2, l = 1, 2$. Note that $\|(\boldsymbol{I} - \boldsymbol{P}_{A_1})\boldsymbol{X}_{A_0}\boldsymbol{\beta}^0_{A_0}\|^2 \geq niC_{\min}$ if $j \leq 2i \equiv 2(p_0 - k)$ by definition of $C_{\min}$ or if $|A_1| + |A_1 \cap A_0| \leq 2p_0$ with $|A_0 \setminus A_1| = p_0 - |A_0 \cap A_1| = i$; or ≥ 0 if $\alpha i < j \leq p$. Then

$$
\begin{aligned}
I_1 \leq \sum_{i=1}^{p_0}\sum_{j=0}^{[\alpha i]}\binom{p-p_0}{j}\binom{p_0}{p_0-i}&\left(\exp\left(\frac{2t_1^2 - \delta t_1}{\sigma^2}niC_{\min} + \frac{t_1}{\sigma^2}\lambda(i-j)\right)\right. \\
&\left. + \exp\left(-\frac{(1-\delta)t_2}{\sigma^2}niC_{\min} + \frac{t_2}{\sigma^2}\lambda(i-j) + 2t_2 j\right)\right) \\
&+ \sum_{i=1}^{p_0}\sum_{j=[\alpha i]+1}^{p}\binom{p-p_0}{j}\binom{p_0}{p_0-i}\left(\exp\left(\frac{t_2\lambda}{\sigma^2}(i-j)\right) + \exp\left(\frac{t_4\lambda}{\sigma^2}(i-j)\right)\right).
\end{aligned}
$$

To simplify this bound, choose $t_1 = \frac{1}{3}$ and $\delta = \frac{2t_1+1}{2} = \frac{5}{6}$ such that $\delta t_1 - 2t_1^2 = (1-\delta)t_1 = \frac{1}{18}$. Note that $\binom{a}{b} \leq a^b$ and $\log(p - p_0) + \log p_0 \leq \log(\frac{p^2}{4}) \leq 2\log p - 1$. Note that $\sum_{j=0}^{b}\binom{a}{j} \leq (a+1)^b$, and $\binom{a}{b} \leq a^b$, for any integers $a, b > 0$. Then

$$
\begin{aligned}
I_1 \leq 2\sum_{i=1}^{p_0} p_0^i \sum_{j=0}^{[\alpha i]}\binom{p-p_0}{j}\exp\left(-\frac{inC_{\min}}{28\sigma^2} + \frac{2j}{3} + \frac{(i-j)\lambda}{2\sigma^2}\right) \\
+ 2\sum_{j=[\alpha i]+1}^{p}(p-p_0)^j\exp\left(\frac{-(\alpha-1)j\lambda}{3\alpha\sigma^2} + \frac{2j}{3}\right)\sum_{i=0}^{[j/\alpha]}\binom{p_0}{i}
\end{aligned}
$$

$$\le 2\sum_{i=1}^{p_0}\exp\left(-i\left(\frac{nC_{\min}}{18\sigma^2}-\log p_0-\alpha\log(p-p_0+1)-\frac{\lambda}{2\sigma^2}\right)\right)$$
$$+2\sum_{j=|\alpha i|+1}^{p}\exp\left(-j\left(\frac{(\alpha-1)\lambda}{3\alpha\sigma^2}-\log(p-p_0)-\frac{1}{\alpha}\log(p_0+1)-\frac{2}{3}\right)\right).$$

This, together with the fact that $I_1 \le 1$, $\log p_0+\alpha\log(p-p_0+1) \le (\alpha+1)(\log(p+1)-\log(\alpha+1)+\frac{\alpha}{\alpha+1}\log\alpha)$ and $\frac{1}{\alpha}\log(p_0+1)+\log(p-p_0) \le (1+\frac{1}{\alpha})(\log(p+1)-\frac{1}{\alpha+1}\log(\alpha+1)-\frac{\alpha}{\alpha+1}\log(1+\frac{1}{\alpha}))$, leads to (15). The risk and minimaxity results follow similarly as in the Proof of Theorem 5. This completes the proof. □

Proof of Theorem 6 Let $H = \{\min_{j\in A_0}|\hat\beta_j^{ol}| > 3\tau/2\} \cap \{\max_{j\notin A_0}|(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol})| \le \frac{\lambda}{\tau}\}$. Rewrite (23), for any subset A of non-zero coefficients and $\boldsymbol{\beta}$,

$$\begin{cases} -(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\frac{\lambda}{\tau}\text{sign}(\beta_j)I(|\beta_j|<\tau)=0, & j\in A,\\ |(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})| \le \frac{\lambda}{\tau}, & j\notin A. \end{cases} \tag{29}$$

Next we prove that $\hat{\boldsymbol{\beta}}^{ol}$ satisfies (29) on H. Note that the first event in H implies that $\nabla_j S_2(\hat{\boldsymbol{\beta}}^{ol})-\frac{\lambda}{\tau}\text{sign}(\hat\beta_j^{ol})=0$; $j=1,\ldots,p_0$. This, together with the property that $(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol})=0$; $j=1,\ldots,p_0$ yields the first equation of (29). The second event in H implies the second equation of (29) by $\hat{\boldsymbol{\beta}}^{ol}$.

For a unique minimum of (29) on H, suppose $\hat{\boldsymbol{\beta}}^{lo}_{\hat A} \ne \hat{\boldsymbol{\beta}}^{ol}_{A_0}$. Let $A* = \hat A\cup A_0$. Define $g(\boldsymbol{\beta}_{A*}) = S(\boldsymbol{\beta})$, where $\boldsymbol{\beta}=(\boldsymbol{\beta}_{A*},\boldsymbol{0}_{A*^c})$ and $\boldsymbol{\beta}_{A*}=(\beta_1,\ldots,\beta_{|A*|})^T$. Then

$$\left|\left(\frac{\partial}{\partial\boldsymbol{\beta}_{A*}}g(\hat{\boldsymbol{\beta}}^{lo}_{A*})-\frac{\partial}{\partial\boldsymbol{\beta}_{A*}}g(\hat{\boldsymbol{\beta}}^{ol}_{A*})\right)^T\frac{(\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*})}{\|\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*}\|}\right| = \left|\left(X^T_{A*}X_{A*}(\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*})\right.\right.$$
$$\left.\left.+\frac{\lambda}{\tau}\text{sign}(\hat{\boldsymbol{\beta}}^{lo}_{A*})I(|\hat{\boldsymbol{\beta}}^{lo}_{A*}|\le\tau)-\frac{\lambda}{\tau}\text{sign}(\hat{\boldsymbol{\beta}}^{ol}_{A*})I(|\hat{\boldsymbol{\beta}}^{ol}_{A*}|\le\tau)\right)^T\frac{(\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*})}{\|\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*}\|}\right|,$$

where $\hat{\boldsymbol{\beta}}^{ol}_{A*}$ and $\hat{\boldsymbol{\beta}}^{lo}_{A*}$ cannot attain at nondifferentiable concave points of the penalty by Lemma 1. Without loss of generality, assume that $\|\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*}\| \ge \tau/2$. Otherwise, for any $j\in A_0|\hat\beta_j^{lo}|>\tau$, and for $j\in A_0^c\cap A*|\hat\beta_j^{lo}|\le\tau$, implying $\hat{\boldsymbol{\beta}}^{lo}_{A*}=\hat{\boldsymbol{\beta}}^{ol}_{A*}$ on H, as shown from (28), which is impossible by assumption that $\hat{\boldsymbol{\beta}}^{lo}_{A*}\ne\hat{\boldsymbol{\beta}}^{ol}_{A*}$. By the Cauchy–Schwarz inequality $|(\frac{\lambda}{\tau}\text{sign}(\hat{\boldsymbol{\beta}}^{lo}_{A*})I(|\hat{\boldsymbol{\beta}}^{lo}_{A*}|\le\tau)-\frac{\lambda}{\tau}\text{sign}(\hat{\boldsymbol{\beta}}^{ol}_{A*})I(|\hat{\boldsymbol{\beta}}^{ol}_{A*}|\le\tau))^T(\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*})| \le \frac{2\lambda}{\tau}\sqrt{K^*}\|\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*}\|$ that is bounded below by $\min_{|B|\le 2K^*,A_0\subseteq B} nc_{\min}(n^{-1}\boldsymbol{X}_B^T\boldsymbol{X}_B)\frac{\tau}{2}-\frac{2\lambda}{\tau}\sqrt{K^*}>0$, contradicting to the fact that $\boldsymbol{0}\in(\frac{\partial}{\partial\boldsymbol{\beta}_{A*}}g(\hat{\boldsymbol{\beta}}^{lo}_{A*})-\frac{\partial}{\partial\boldsymbol{\beta}^{lo}_{A*}}g(\hat{\boldsymbol{\beta}}^{ol}_{A*}))^T\frac{(\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*})}{\|\hat{\boldsymbol{\beta}}^{lo}_{A*}-\hat{\boldsymbol{\beta}}^{ol}_{A*}\|}$ on H if $\hat{\boldsymbol{\beta}}^{lo}\ne\hat{\boldsymbol{\beta}}^{ol}$ is a local

minimizer of $S(\cdot)$ thus $g(\cdot)$. Hence, $g(\boldsymbol{\beta}^{lo}_{A*})$ has a unique local minimizer on H, implying $\hat{\boldsymbol{\beta}}^{ol} = \hat{\boldsymbol{\beta}}^{lo}$.

Note that $(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^{ol}) \sim N(0, \sigma^2\|(\boldsymbol{I} - \boldsymbol{P}_{A_0})\boldsymbol{x}^{(j)}\|^2)$, $\|(\boldsymbol{I} - \boldsymbol{P}_{A_0})\boldsymbol{x}^{(j)}\|^2 \le \|\boldsymbol{x}^{(j)}\|^2$, $\hat{\beta}^{ol}_j \sim N(\beta^0_j, Var(\hat{\beta}^{ol}_j))$, and $Var(\hat{\beta}^{ol}_j) \ge c^{-1}_{\min}(n^{-1}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})\sigma^2/n$. Then $P(\hat{\boldsymbol{\beta}}^{lo}_{\hat{A}^{lo}} \ne \hat{\boldsymbol{\beta}}^{ol}_{A_0}) \le P(H^c) \le \sum_{j\in A_0} P(|\hat{\beta}^{ol}_j| \le 3\tau/2) + \sum_{j\notin A_0} P(|(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^{ol})| > \frac{\lambda}{\tau}) \equiv I_6 + I_7$, where $I_6 \le |A_0|(\Phi(-\frac{n^{1/2}(\gamma_{\min}-3\tau/2)}{\sigma c^{-1/2}_{\min}(\frac{1}{n}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})}) - \Phi(-\frac{n^{1/2}(\gamma_{\min}+3\tau/2)}{\sigma c^{-1/2}_{\min}(\frac{1}{n}\boldsymbol{X}^T_{A_0}\boldsymbol{X}_{A_0})}))$, and $I_7 \le (p - |A_0|)\Phi(-\frac{\lambda/\tau}{\sigma \max_{1\le j\le p}\|\boldsymbol{x}^{(j)}\|})$. This yields (24).

For the risk property, let $\hat{A} = \{j : |\hat{\beta}^{lo}_j| \ge \tau\}$. By (29), $\hat{\boldsymbol{\beta}}^{lo}_{\hat{A}} = (\boldsymbol{X}^T_{\hat{A}}\boldsymbol{X}_{\hat{A}})^{-1}\boldsymbol{X}^T_{\hat{A}}(\boldsymbol{Y} - \boldsymbol{X}_{\hat{A}^c}\hat{\boldsymbol{\beta}}_{\hat{A}^c})$. As in the Proof of Theorem 5 for the global minimizer, we rewrite the risk as the sum of T_1 and T_2. For $T_1 = \int_C^\infty P(\frac{1}{n}\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{lo} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 \ge x)\,\mathrm{d}x$, by the triangular inequality, $\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{lo} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 = \|(\boldsymbol{I} - \boldsymbol{P}_{\hat{A}})(\boldsymbol{X}_{\hat{A}^c}\hat{\boldsymbol{\beta}}^{lo}_{\hat{A}^c} + \boldsymbol{X}\boldsymbol{\beta}^0) + \boldsymbol{P}_{\hat{A}}\boldsymbol{\epsilon}\|^2 \le 4(c_{\max}(\boldsymbol{X}^T\boldsymbol{X})p^2\tau^2 + \|\boldsymbol{X}\boldsymbol{\beta}^0\|^2) + 2\|\boldsymbol{\epsilon}\|^2$. Let $C = 7\sigma^2 + 2c_{\max}(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n})p^2\tau^2 + \frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2$ and $t = 1/3$. By Markov's inequality,

$$
\begin{aligned}
T_1 &\le \int_C^\infty P\left(\|\boldsymbol{\epsilon}\|^2 \ge \frac{xn}{2} - 2c_{\max}(\boldsymbol{X}^T\boldsymbol{X})p^2\tau^2 - 2\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2\right)\mathrm{d}x \\
&\le \int_C^\infty E(\exp(t\|\boldsymbol{\epsilon}\|^2/\sigma^2))\exp\left(-nt\frac{x - 2c_{\max}\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)p^2\tau^2 - \frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2}{2\sigma^2}\right)\mathrm{d}x \\
&\le \int_C^\infty \exp\left(-nt\frac{x - 6\sigma^2 - 2c_{\max}\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)p^2\tau^2 - \frac{4}{n}\|\boldsymbol{X}\boldsymbol{\beta}^0\|^2}{2\sigma^2}\right)\mathrm{d}x = o\left(\frac{p_0}{n}\sigma^2\right).
\end{aligned}
$$

For T_2, by the probability error bound, $T_2 \le CP(\hat{\boldsymbol{\beta}}^{lo} \ne \hat{\boldsymbol{\beta}}^{ol}) + \frac{1}{n}E\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{ol} - \boldsymbol{X}\boldsymbol{\beta}^0\|^2 = (1 + o(1))\frac{p_0}{n}\sigma^2$, leading to the desired result.

Finally, it remains to show that $\hat{\boldsymbol{\beta}}^{lo}$ satisfies (23). Note that the local optimality (21) is satisfied by $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(m)}$: $-(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda}{\tau}\mathrm{sign}(\beta_j)I(|\boldsymbol{\beta}^{(m-1)}| < \tau) = 0 j = 1, \ldots, p$. By construction, $\hat{\beta}^{(m)}_j = \hat{\beta}^{(m^*-1)}_j \ne \pm\tau$; for $m \ge m^*$ and $j = 1, \ldots, p$, implying (23). This completes the proof. □

References

Akaike, H. (1973). Information theory and the maximum likelihood principle. In V. Petrov, F. Csáki (Eds.), *International symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiádo.

Chen, J., Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, *95*, 759–771.

Chen, S.S., Donoho, D., Saunders, M.A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*, 129–159.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Gu, C. (1998). Model indexing and smoothing parameter in nonparametric function estimation. *Statistica Sinica*, *8*, 607–646.
Ibragimov, I., Has'minskii, R. (1981). *Statistical estimation*. New York: Springer.
Judge, G.G., Bock, M.E. (1978). *The statistical implications of pretest and Stein-rule estimators in econometrics*. Amsterdam: North-Holland.
Kim, Y., Choi, H., Oh, H.-S. (2008). Smoothly clipped absolute deviation of high dimensions. *Journal of the American Statistical Association*, *103*, 1665–1673.
Liu, W., Yang, Y.Y. (2010). *Consistency for BIC selection* (manuscript)
Lv, J., Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, *37*, 3498–3528.
Meinshausen, N., Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*, 1436–1462.
Osborne, M.R., Presnell, B., Turlach, B.A. (2000). On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, *9*, 319–337.
Rainaldo, A. (2007). *A note on the uniqueness of the Lasso solution*. Technical Report, Department of Statistics, Carnegie Mellon University.
Raskutti, G., Wainwright, M., Yu, B. (2009). *Minimax rates of estimation for high-dimensional linear regression over l_q balls*. Technical Report, UC Berkeley.
Rockafellar, R.T., Wets, R.J.B. (2011). Variational analysis. vol 317. New York: Springer.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Shen, X., Pan, W., Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, *107*, 223–232.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.
Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery of sparsity. *IEEE Transactions on Information Theory*, *55*, 2183–2202.
Yang, Y., Barron, A. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, *44*, 95–116.
Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zhao, P., Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.
Zhou, S., Shen, X., Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, *26*, 1760–1782.
Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, *36*, 1509–1533.