

Variable Selection for Support Vector Machines in High Dimensions

Xiang Zhang and Yichao Wu

North Carolina State University, Raleigh, NC, USA.

Lan Wang

The University of Minnesota, Minneapolis, MN, USA.

Runze Li

The Pennsylvania State University, University Park, PA, USA.

Summary. The support vector machine (SVM) is a powerful binary classification tool with high accuracy and great flexibility. It has achieved great success, but its performance can be seriously impaired if many redundant covariates are included. Some efforts have been devoted to studying variable selection for SVMs, but asymptotic properties, such as variable selection consistency, are largely unknown when the number of predictors diverges to infinity. In this work, we establish a unified theory for a general class of nonconvex penalized SVMs in the high-dimensional setting. We prove that there exists a local minimizer of the nonconvex penalized SVM objective function possessing the oracle property even when the number of predictors grows exponentially with the sample size. Furthermore, we give sufficient conditions under which the proposed algorithm identifies the oracle solution among potentially multiple local minimizers. Numerical examples provide supportive evidence.

Keywords: Nonconvex penalty; oracle property; support vector machines; ultra-high dimension; variable selection

1. Introduction

Due to the recent advent of new technologies for data acquisition and storage, we have seen an explosive growth of data complexity in a variety of research areas such as genomics, imaging and finance. As a result, the number of predictors becomes huge. However there are only a moderate number of instances available for study (Donoho et al., 2000). For example, in tumor classification using genomic data, expression values of tens of thousands of genes are available, but the number of arrays is typically at the order of tens. Classification of high-dimensional data poses many statistical challenges and calls for new methods and theories. In this article we consider high-dimensional classification where the number of covariates diverges with the sample size and can be potentially much larger than the sample size.

Support vector machine (SVM, Vapnik, 1996) is a powerful binary classification tool with high accuracy and great flexibility. It has achieved success in many applications. However, one serious drawback of the standard SVM is that its performance can be adversely affected if many redundant variables are included in building the decision rule (Friedman et al., 2001), see the evidence in the numerical results of Section 5.1. Classification using all features has been shown to be as poor as random guessing due to noise accumulation in high-dimensional space (Fan and Fan, 2008). Many methods have been proposed to remedy this problem, such as the recursive feature elimination suggested by Guyon et al. (2002). In

particular, superior performance can be achieved with a unified method, namely achieving variable selection and prediction simultaneously (Fan and Li, 2001) by using an appropriate sparsity penalty. It is well known that the standard SVM can fit in the regularization framework of $loss + penalty$ using the hinge loss and L_2 penalty. Based on this, several attempts have been made to achieve variable selection for the SVM by replacing the L_2 penalty with other forms of penalty. Bradley and Mangasarian (1998), Zhu et al. (2004), and Wegkamp and Yuan (2011) considered the L_1 -penalized SVM; Zou and Yuan (2008) proposed to use the F_∞ -norm SVM to select groups of predictors; Wang et al. (2006) and Wang et al. (2007) suggested the elastic net penalty for the SVM; Zou (2007) proposed to penalize the SVM with the adaptive LASSO penalty; Zhang et al. (2006), Becker et al. (2011) and Park et al. (2012) studied the smoothly clipped absolute deviation (SCAD, Fan and Li, 2001)-penalized SVM. Recently Park et al. (2012) studied the oracle property of the SCAD-penalized SVM with a fixed number of predictors. Yet, to the best of our knowledge, the theory of variable selection consistency of sparse SVMs in high dimensions or ultra-high dimensions (Fan and Lv, 2008) has not been studied so far.

In this article, we study the variable selection consistency of sparse SVMs. Instead of using the L_2 penalty, we consider the penalized SVM with a general class of nonconvex penalties, such as the SCAD penalty or the minimax concave penalty (MCP, Zhang, 2010). Though the convex L_1 penalty can also induce sparsity, it is well known that its variable selection consistency in linear regression relies on the stringent “irrepresentable condition” on the design matrix. This condition, however, can easily be violated in practice, see examples in Zou (2006) and Meinshausen and Yu (2009). Moreover, the regularization parameter for model selection consistency in this case is not optimal for prediction accuracy (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2007). For the nonconvex penalty, Kim et al. (2008) investigated the oracle property of the SCAD-penalized least squares regression in the high dimensions. However, a different set of proving techniques are needed for the nonconvex penalized SVMs because the hinge loss in the SVM is not a smooth function. The Karush-Huhn-Tucker local optimality condition is generally not sufficient for the setup of a nonsmooth loss plus a nonconvex penalty. A new sufficient optimality condition based on subgradient calculation is used in the technical proof in this paper. We prove that under some general conditions, with probability tending to one, the oracle estimator is a local minimizer of the nonconvex penalized SVM objective function where the number of variables may grow exponentially with the sample size. By oracle estimator, we mean an estimator obtained by minimizing the empirical hinge loss with only relevant covariates. As one referee pointed out, with a finite sample, the empirical hinge loss may have multiple minimizers because the objective function is piecewise linear. This issue will vanish asymptotically because we assume that the population hinge loss has a unique minimizer. Such an assumption on the population hinge loss has been made in the existing literature (Koo et al., 2008).

Even though the nonconvex penalized SVMs are shown to enjoy the aforementioned local oracle property, it is largely unknown whether numerical algorithms can identify this local minimizer since the involved objective function is nonconvex and typically multiple local minimizers exist. Existing methods rely heavily on conditions that guarantee the local minimizer to be unique. In general, when the convexity of the hinge loss function dominates the concavity of the penalty, the nonconvex penalized SVMs actually have a unique minimizer due to global convexity. Recently Kim and Kwon (2012) gave sufficient conditions for a unique minimizer of the nonconvex penalized least square regression when global convexity is not satisfied. However, for ultra-high-dimensional cases, it would be

unrealistic to assume the existence of a unique local minimizer. See Wang et al. (2013) for relevant discussion and a possible solution to nonconvex penalized regression.

In this article, we further extend our results by verifying that with probability tending to one, the local linear approximation (LLA) algorithm (Zou and Li, 2008) is guaranteed to yield an oracle estimator with desirable properties under appropriate conditions. This result extends the localizability condition for folded concave penalized estimation in Fan et al. (2014) by relaxing the differentiability condition of the loss function. We further provide sufficient conditions under which an appropriate initial estimator for the LLA algorithm is available. To our knowledge, this is the first asymptotic result of this type in the setting of a nonsmooth loss function with a nonconvex penalty. We demonstrate that the estimator from the algorithm simultaneously achieves variable selection consistency and accurate prediction even in the ultra-high-dimensional setting.

The rest of this paper is organized as follows. Section 2 introduces the methodology of nonconvex penalized SVMs. Section 3 contains the main results of the properties of nonconvex penalized SVMs. The implementation procedure is summarized in Section 4. Simulation studies and a real data example are provided in Section 5, followed by a discussion in Section 6. Technical proofs are presented in Section 7. A zip file containing R demonstration codes for one simulation example and the real data example is available at <http://www4.stat.ncsu.edu/~wu/soft/VarSelforSVMbyZhangWuWangLi.zip>.

2. Nonconvex penalized support vector machines

We begin with the basic setup and notation. In binary classification, we are typically given a random sample $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ from an unknown population distribution $P(\mathbf{X}, Y)$. Here $Y_i \in \{1, -1\}$ denotes the categorical label and $\mathbf{X}_i = (X_{i0}, X_{i1}, \dots, X_{ip})^T = (X_{i0}, (\mathbf{X}_i^*)^T)^T$ denotes the input covariates with $X_{i0} = 1$ corresponding to the intercept term. The goal is to estimate a classification rule that can be used to predict output labels for future observations with input covariates only. With potentially varying misclassification cost specified by weight $W_i = w$ if $Y_i = 1$ and $W_i = 1 - w$ if $Y_i = -1$ for some $0 < w < 1$, the linear-weighted support vector machine (WSVM, Lin et al., 2002) estimates the classification boundary by solving

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \lambda (\boldsymbol{\beta}^*)^T \boldsymbol{\beta}^*,$$

where $(1 - u)_+ = \max\{1 - u, 0\}$ denotes the hinge loss, $\lambda > 0$ is a regularization parameter, and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}^*)^T)^T$ with $\boldsymbol{\beta}^* = (\beta_1, \beta_2, \dots, \beta_p)^T$. The standard SVM is a special case of the WSVM with weight parameter $w = 0.5$. In this paper, we consider the WSVM for more generality. In general, the corresponding decision rule, $\text{sign}(\mathbf{X}^T \boldsymbol{\beta})$, uses all covariates and is not capable of selecting relevant covariates.

Towards variable selection for the linear WSVM, we consider the population linear-weighted hinge loss $\mathbb{E}\{W(1 - Y \mathbf{X}^T \boldsymbol{\beta})_+\}$. Let $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \dots, \beta_{0p})^T = (\beta_{00}, (\boldsymbol{\beta}_0^*)^T)^T$ denote the true parameter value, which is defined as the minimizer of the population-weighted hinge loss. Namely

$$\boldsymbol{\beta}_0 = \arg \min_{\boldsymbol{\beta}} \mathbb{E}\{W(1 - Y \mathbf{X}^T \boldsymbol{\beta})_+\}. \quad (1)$$

The number of covariates $p = p_n$ is allowed to increase with the sample size n . It is even possible that p_n is much larger than n . In this paper we assume the true parameter $\boldsymbol{\beta}_0$ to

be sparse. Let $A = \{1 \leq j \leq p_n; \beta_{0j} \neq 0\}$ be the index set of the nonzero coefficients. Let $q = q_n = |A|$ be the cardinality of set A , which is also allowed to increase with n . Without loss of generality, we assume that the last $p_n - q_n$ components of β_0 are zero. That is, $\beta_0^T = (\beta_{01}^T, \mathbf{0}^T)$. Correspondingly, we write $\mathbf{X}_i^T = (\mathbf{Z}_i^T, \mathbf{R}_i^T)$, where $\mathbf{Z}_i = (X_{i0}, X_{i1}, \dots, X_{iq})^T = (1, (\mathbf{Z}_i^*)^T)^T$ and $\mathbf{R}_i = (X_{i[q+1]}, \dots, X_{ip})^T$. Further we denote π_+ (resp. π_-) to be the marginal probability of the label $Y = +1$ (resp. -1).

To facilitate our theoretical analysis, we introduce the gradient vector and Hessian matrix of the population linear-weighted hinge loss. Let $L(\beta_1) = \mathbb{E}\{W(1 - Y\mathbf{Z}^T\beta_1)_+\}$ be the population linear-weighted hinge loss using only relevant covariates. Define $S(\beta_1) = (S(\beta_1)_j)$ to be the $(q_n + 1)$ -dimensional vector given by

$$S(\beta_1) = -\mathbb{E}\{I(1 - Y\mathbf{Z}^T\beta_1 \geq 0)WY\mathbf{Z}\},$$

where $I(\cdot)$ denotes the indicator function. Also define $H(\beta_1) = (H(\beta_1)_{jk})$ to be the $(q_n + 1) \times (q_n + 1)$ -dimensional matrix given by

$$H(\beta_1) = \mathbb{E}\{\delta(1 - Y\mathbf{Z}^T\beta_1)W\mathbf{Z}\mathbf{Z}^T\},$$

where $\delta(\cdot)$ denotes the Dirac delta function. It can be shown that if well-defined, $S(\beta_1)$ and $H(\beta_1)$ can be considered to be the gradient vector and Hessian matrix of $L(\beta_1)$, respectively. See Lemma 2 of Koo et al. (2008) for details.

2.1. Nonconvex penalized support vector machines

By acting as if the true sparsity structure is known in advance, the oracle estimator is defined as $\hat{\beta} = (\hat{\beta}_1^T, \mathbf{0}^T)^T$, where

$$\hat{\beta}_1 = \arg \min_{\beta_1} n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{Z}_i^T \beta_1)_+. \quad (2)$$

Here the objective function is piecewise linear. With a finite sample, it may have multiple minimizers. In that case, $\hat{\beta}_1$ can be chosen to be any minimizer. Our forthcoming theoretical results still hold. In the limit as $n \rightarrow \infty$, $\hat{\beta}_1$ minimizes the population version of the objective function $\mathbb{E}\{W(1 - Y\mathbf{Z}^T\beta_1)_+\}$. Lin (2002) showed that when the misclassification costs are equal, the minimizer of $\mathbb{E}\{(1 - Yf(\mathbf{Z}))_+\}$ over measurable $f(\mathbf{Z})$ is the Bayes rule $\text{sgn}(p(\mathbf{Z}) - 1/2)$, where $p(\mathbf{z}) = P(Y = +1 | \mathbf{Z} = \mathbf{z})$. This suggests that the oracle estimator is aiming at approximating the Bayes rule. In practice, achieving the oracle estimator is very challenging, because the sparsity structure of the true parameter β_0 is largely unknown. Later we will show that under some regularity conditions, our proposed algorithm can find the oracle estimator and claim convergence with high probability. Indeed, the numerical examples in Section 5.1 demonstrate that the estimator selected by our proposed algorithm has performance close to that of the Bayes rule. Note that the Bayes rule is unattainable here because we assume no knowledge on the high dimensional conditional density $P(\mathbf{X}|Y)$.

In this paper, we consider the nonconvex penalized hinge loss objective function:

$$Q(\beta) = n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \beta)_+ + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|), \quad (3)$$

where $p_{\lambda_n}(\cdot)$ is a symmetric penalty function with tuning parameter λ_n . Let $p'_{\lambda_n}(t)$ be the derivative of $p_{\lambda_n}(t)$ with respect to t . We consider a general class of nonconvex penalties that satisfy the following conditions.

(Condition 1) The symmetric penalty $p_{\lambda_n}(t)$ is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $p'_{\lambda_n}(t)$ on $(0, +\infty)$ and $p_{\lambda_n}(0) = 0$.

(Condition 2) There exists $a > 1$ such that $\lim_{t \rightarrow 0^+} p'_{\lambda_n}(t) = \lambda_n$, $p'_{\lambda_n}(t) \geq \lambda_n - t/a$ for $0 < t < a\lambda$ and $p'_{\lambda_n}(t) = 0$ for $t \geq a\lambda$.

The motivation for such a nonconvex penalty is that the convex L_1 penalty lacks the oracle property due to the over-penalization of large coefficients in the selected model. Consequently it is undesirable to use the L_1 penalty when the purpose of the data analysis is to select the relevant covariates among potentially high-dimensional candidates in classification. Note that p , q , λ and other related quantities are allowed to depend on n , and we suppress the subscript n whenever there is no confusion.

Two commonly used nonconvex penalties that satisfy Conditions 1 and 2 are the SCAD and MCP penalties. The SCAD penalty (Fan and Li, 2001) is defined by

$$p_{\lambda}(|\beta|) = \lambda|\beta|I(0 \leq |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} \times I(\lambda \leq |\beta| \leq a\lambda) + \frac{(a+1)\lambda^2}{2}I(|\beta| > a\lambda)$$

for some $a > 2$. The MCP (Zhang, 2010) is defined by

$$p_{\lambda}(|\beta|) = \lambda(|\beta| - \frac{\beta^2}{2a\lambda})I(0 \leq |\beta| < a\lambda) + \frac{a\lambda^2}{2}I(|\beta| \geq a\lambda) \text{ for some } a > 1.$$

3. Oracle property

3.1. Regularity conditions

To facilitate our technical proofs, we impose the following regularity conditions:

(A1) The densities of \mathbf{Z}^* given $Y = +1$ and $Y = -1$ are continuous and have common support in \mathcal{R}^q .

(A2) $\mathbb{E}[X_j^2] < \infty$ for $1 \leq j \leq q$.

(A3) The true parameter β_0 is unique and a nonzero vector.

(A4) $q_n = O(n^{c_1})$, namely $\lim_{n \rightarrow \infty} q_n/n^{c_1} < \infty$, for some $0 \leq c_1 < 1/2$.

(A5) There exists a constant $M_1 > 0$ such that $\lambda_{max}(n^{-1}\mathbf{X}_A^T\mathbf{X}_A) \leq M_1$, where \mathbf{X}_A is the first q_n+1 columns of the design matrix and λ_{max} denotes the largest eigenvalue. It is further assumed that $\max_{1 \leq i \leq n} \|\mathbf{Z}_i\| = O_p(\sqrt{q_n} \log(n))$, (\mathbf{Z}_i, Y_i) are in general position (Koenker, 2005, sect. 2.2), X_{ij} are sub-Gaussian random variables for $1 \leq i \leq n, q_n+1 \leq j \leq p_n$.

(A6) $\lambda_{min}(H(\beta_{01})) \geq M_2$ for some constant $M_2 > 0$, where λ_{min} denotes the smallest eigenvalue.

(A7) $n^{(1-c_2)/2} \min_{1 \leq j \leq q_n} |\beta_{0j}| \geq M_3$ for some constant $M_3 > 0$ and $2c_1 < c_2 \leq 1$.

(A8) Denote the conditional density of $\mathbf{Z}^T\beta_{01}$ given $Y = +1$ and $Y = -1$ as f and g , respectively. It is assumed that f is uniformly bounded away from 0 and ∞ in a neighborhood of 1 and g is uniformly bounded away from 0 and ∞ in a neighborhood of -1.

REMARK 3.1. *The Conditions (A1)-(A3) and (A6) are also assumed for fixed p in Koo et al. (2008). We need these assumptions to ensure that the oracle estimator is consistent in the scenario of diverging p . Condition (A3) states that the optimal classification decision function is not constant, which is required to ensure $\mathbf{S}(\beta)$ and $\mathbf{H}(\beta)$ are well-defined gradient vector and Hessian matrix of the hinge loss, see the Lemma 2 and Lemma 3 of Koo et al.*

(2008). The Conditions (A4) and (A7) are common in the literature of high-dimensional inference (Kim et al., 2008). More specifically, (A4) states that the divergence rate of the number of nonzero coefficients cannot be faster than root- n and (A7) simply states that the signals cannot decay too quickly. The Condition on the largest eigenvalues of the design matrix in (A5) is similar to the sparse Riesz condition and also assumed in Zhang and Huang (2008), Yuan (2010) and Zhang (2010). Note that the bound on the smallest eigenvalue is not specified. The Condition on the maximum norm in (A5) holds when \mathbf{Z}^* given Y follows multivariate normal distribution. (\mathbf{Z}_i, Y_i) are in general position if with probability one there are exactly $(q_n + 1)$ elements in $\mathbf{D} = \{i : 1 - Y_i \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}_1 = 0\}$ (Koenker, 2005). The condition for general position is true with probability one w.r.t. Lebesgue measure. The Condition (A8) requires that there is enough information around the nondifferentiable point of the hinge loss, similar to Condition (C5) in Wang et al. (2012) for quantile regression.

For an illustrative example that satisfies all the above conditions, assume $0 < \pi_+ = 1 - \pi_- < 1$ and the conditional distribution of \mathbf{X}^* given Y are continuous with support \mathcal{R}^p and have sub-Gaussian tail. It is clear that the Fisher's linear discriminant analysis is a special case where the conditional distribution of \mathbf{X}^* given Y are Gaussian. Let the number of signals be fixed. Condition (A1)-(A4) and (A7) are trivial. Condition (A5) holds by the properties of sub-Gaussian random variable. Koo et al. (2008) showed that Condition (A6) holds if the supports of the conditional densities of \mathbf{Z}^* given Y are convex, which are naturally satisfied for \mathcal{R}^q . Condition (A8) is trivially satisfied by the unbounded support of the conditional distribution of \mathbf{Z}^* given Y .

3.2. Local oracle property

In this subsection, we establish the theory of the local oracle property for the nonconvex penalized SVMs; namely, the oracle estimator is one of the local minimizers of the objective function $Q(\boldsymbol{\beta})$ defined in (3). We start with the following lemma on the consistency of the oracle estimator, which can be viewed as an extension of the consistency result in Koo et al. (2008) to the diverging p scenario.

LEMMA 3.1. *Assume that Conditions (A1)-(A7) are satisfied. The oracle estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ satisfies $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = O_p(\sqrt{q_n/n})$ when $n \rightarrow \infty$.*

Though the convexity of the nonconvex penalized hinge loss objective function $Q(\boldsymbol{\beta})$ is not guaranteed, it can be written as the difference of two convex functions:

$$Q(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) - h(\boldsymbol{\beta}), \quad (4)$$

where $g(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \lambda_n \sum_{j=1}^p |\beta_j|$ and $h(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^p |\beta_j| - \sum_{j=1}^p p_{\lambda_n}(|\beta_j|) = \sum_{j=1}^p H_{\lambda_n}(\beta_j)$. The form of $H_{\lambda}(\beta_j)$ depends on the penalty function. For the SCAD penalty, we have

$$H_{\lambda}(\beta_j) = [(\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)/\{2(a-1)\}]I(\lambda \leq |\beta_j| \leq a\lambda) + \{\lambda|\beta_j| - (a+1)\lambda^2/2\}I(|\beta_j| > a\lambda),$$

while for MCP, we have $H_{\lambda}(\beta_j) = \{\beta_j^2/(2a)\}I(0 \leq |\beta_j| < a\lambda) + (\lambda|\beta_j| - a\lambda^2/2)I(|\beta_j| \geq a\lambda)$. This decomposition is useful, as it naturally satisfies the form of the difference of convex functions (DC) algorithm (An and Tao, 2005).

To prove the oracle property of the nonconvex penalized SVMs, we will use a sufficient local optimality condition for the difference convex programming first presented in Tao and

An (1997). This sufficient condition is based on subgradient calculus. The subgradient can be viewed as an extension of the gradient of the smooth convex function to the nonsmooth convex function. Let $\text{dom}(g) = \{\mathbf{x} : g(\mathbf{x}) < \infty\}$ be the effective domain of a convex function g . The subgradient of $g(\mathbf{x})$ at a point \mathbf{x}_0 is defined as $\partial g(\mathbf{x}_0) = \{\mathbf{t} : g(\mathbf{x}) \geq g(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \mathbf{t}\}$. Note that at the nondifferentiable point, the subgradient contains a collection of vectors. One can easily check that the subgradient of the hinge loss function at the oracle estimator is the collection of vectors $s(\hat{\boldsymbol{\beta}}) = (s_0(\hat{\boldsymbol{\beta}}), \dots, s_p(\hat{\boldsymbol{\beta}}))^T$ with

$$s_j(\hat{\boldsymbol{\beta}}) = -n^{-1} \sum_{i=1}^n W_i Y_i \mathbf{X}_{ij} I(1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} > 0) - n^{-1} \sum_{i=1}^n W_i Y_i \mathbf{X}_{ij} v_j, \quad (5)$$

where $-1 \leq v_i \leq 0$ if $1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise, $j = 0, \dots, p$. Under some regularity conditions, we can study the asymptotic behaviors of the subgradient at the oracle estimator. The results are summarized in the following Theorem.

THEOREM 3.1. *Suppose that Conditions (A1)-(A8) hold, and the tuning parameter satisfies $\lambda = o(n^{-(1-c_2)/2})$ and $\log(p)q \log(n)n^{-\frac{1}{2}} = o(\lambda)$. For the oracle estimator $\hat{\boldsymbol{\beta}}$, there exists v_i^* which satisfies $v_i^* = 0$ if $1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} \neq 0$ and $v_i^* \in [-1, 0]$ if $1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} = 0$, such that for $s_j(\hat{\boldsymbol{\beta}})$ with $v_i = v_i^*$, with probability approaching one, we have*

$$s_j(\hat{\boldsymbol{\beta}}) = 0, \quad j = 0, 1, \dots, q,$$

$$|\hat{\beta}_j| \geq (a + \frac{1}{2})\lambda, \quad j = 1, \dots, q,$$

$$|s_j(\hat{\boldsymbol{\beta}})| \leq \lambda \text{ and } |\hat{\beta}_j| = 0, \quad j = q + 1, \dots, p,$$

Theorem 3.1 characterizes the subgradients of the hinge loss at the oracle estimator. It basically says that in a regular setting, with probability arbitrarily close to one, those components of the subgradients corresponding to the relevant covariates are exactly zero and those corresponding to irrelevant covariates are not far away zero.

We now present the sufficient optimality condition based on subgradient calculation. Corollary 1 of Tao and An (1997) states that if there exists a neighborhood U around the point \mathbf{x}^* such that $\partial h(\mathbf{x}) \cap \partial g(\mathbf{x}^*) \neq \emptyset, \forall \mathbf{x} \in U \cap \text{dom}(g)$, then \mathbf{x}^* is a local minimizer of $g(\mathbf{x}) - h(\mathbf{x})$. To verify this local sufficient condition, we study the asymptotic behaviors of subgradients of the two convex functions in the aforementioned decomposition (4) of $Q(\boldsymbol{\beta})$. Note that, based on (5), the subgradient function of $g(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ can be shown to be the following collection of vectors:

$$\begin{aligned} \partial g(\boldsymbol{\beta}) = \left\{ \boldsymbol{\xi} = (\xi_0, \dots, \xi_p)^T \in \mathcal{R}^{p+1} : \right. \\ \left. \xi_j = -n^{-1} \sum_{i=1}^n W_i Y_i \mathbf{X}_{ij} I(1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} > 0) - n^{-1} \sum_{i=1}^n W_i Y_i \mathbf{X}_{ij} v_j + \lambda l_j, j = 0, \dots, p \right\}, \end{aligned}$$

where $l_0 = 0, l_j = \text{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise for $1 \leq j \leq p$, and $-1 \leq v_i \leq 0$ if $1 - Y_i \mathbf{X}_i^T \hat{\boldsymbol{\beta}} = 0$ and $v_i = 0$ otherwise for $1 \leq i \leq n$. Furthermore, by Condition 2 of the class of nonconvex penalty functions, $\lim_{t \rightarrow 0^+} H'_\lambda(t) = \lim_{t \rightarrow 0^-} H'_\lambda(t) = \lambda \text{sgn}(t) - \lambda \text{sgn}(t) =$

0. Thus $h(\boldsymbol{\beta})$ is differentiable everywhere. Consequently the subgradient of $h(\boldsymbol{\beta})$ at point $\boldsymbol{\beta}$ is a singleton:

$$\partial h(\boldsymbol{\beta}) = \{\boldsymbol{\mu} = (\mu_0, \dots, \mu_p) \in \mathcal{R}^{p+1} : \mu_j = \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}, j = 0, \dots, p\}.$$

For the class of nonconvex penalty functions under consideration, $\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = 0$ for $j = 0$. For $1 \leq j \leq p$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = [\{\beta_j - \lambda \text{sgn}(\beta_j)\}/(a-1)]I(\lambda \leq |\beta_j| \leq a\lambda) + \lambda \text{sgn}(\beta_j)I(|\beta_j| > a\lambda)$$

for the SCAD penalty, and

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = (\beta_j/a)I(0 \leq |\beta_j| < a\lambda) + \lambda \text{sgn}(\beta_j)I(|\beta_j| \geq a\lambda)$$

for the MCP.

Combining this with Theorem 3.1, we will prove that with probability tending to one, for any $\boldsymbol{\beta}$ in a ball in \mathcal{R}^{p+1} with the center $\hat{\boldsymbol{\beta}}$ and radius $\frac{\lambda}{2}$, there exists a subgradient $\boldsymbol{\xi} = (\xi_0, \dots, \xi_p)^T \in \partial g(\hat{\boldsymbol{\beta}})$ such that $\frac{h(\boldsymbol{\beta})}{\partial \beta_j} = \xi_j$, $j = 0, 1, \dots, p$. Consequently the oracle estimator $\hat{\boldsymbol{\beta}}$ is itself a local minimizer of (3). This is summarized in the following theorem.

THEOREM 3.2. *Assume that Conditions (A1)-(A8) hold. Let $B_n(\lambda)$ be the set of local minimizers of the objective function $Q(\boldsymbol{\beta})$ with regularization parameter λ . The oracle estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$ satisfies*

$$\Pr\{\hat{\boldsymbol{\beta}} \in B_n(\lambda)\} \rightarrow 1$$

as $n \rightarrow \infty$, if $\lambda = o(n^{-(1-c_2)/2})$, and $\log(p)q \log(n)n^{-\frac{1}{2}} = o(\lambda)$.

It can be shown that if we take $\lambda = n^{-1/2+\delta}$ for some $c_1 < \delta < c_2/2$, then the oracle property holds even for $p = o(\exp(n^{(\delta-c_1)/2}))$. Therefore, the local oracle property holds for the nonconvex penalized SVM even when the number of covariates grows exponentially with the sample size.

3.3. An algorithm with provable convergence to the oracle estimator

Note that Theorem 3.2 indicates that one of the local minimizers possesses the oracle property. However, there can potentially be multiple local minimizers and it remains challenging to identify the oracle estimator. In the high-dimensional setting, assuming that the local minimizer is unique would not be realistic.

In this article, instead of assuming the uniqueness of solutions, we work directly on the conditions under which the oracle estimator can be identified by some numerical algorithms that solve the nonconvex penalized SVM objective function. One possible algorithm is the local linear approximation (LLA) algorithm proposed by Zou and Li (2008). We focus on theoretical development first in this section and delay the detailed LLA algorithm for the nonconvex penalized SVMs to Section 4. Recently LLA has been shown to be capable of identifying the oracle estimator in the setup of folded concave penalized estimation with a

differentiable loss function (Wang et al., 2013; Fan et al., 2014). We generalize their results to nondifferentiable loss functions, so that it can fit in the framework of the nonconvex penalized SVMs. Similar to their work, the main condition required is the existence of an appropriate initial estimator inputed in the iterations of the LLA algorithm. Denote the initial estimator as $\tilde{\beta}^{(0)}$. Intuitively, if the initial estimator $\tilde{\beta}^{(0)}$ lies in a small neighborhood of the true value β_0 , the algorithm should converge to the good local minimizer around β_0 . This localizability will be formalized in terms of L_∞ distance later. With such an appropriate initial estimator, under the aforementioned regularity conditions, one can prove that the LLA algorithm converges to the oracle estimator with probability tending to one.

Let $\tilde{\beta}^{(0)} = (\tilde{\beta}_0^{(0)}, \dots, \tilde{\beta}_p^{(0)})^T$. Consider the following events:

- $F_{n1} = \{|\tilde{\beta}_j^{(0)} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p\}$,
- $F_{n2} = \{|\beta_{0j}| < (a+1)\lambda, \text{ for some } 1 \leq j \leq q\}$,
- $F_{n3} = \{\text{for all subgradients } s(\hat{\beta}), |s_j(\hat{\beta})| > (1-\frac{1}{a})\lambda \text{ for some } q+1 \leq j \leq p \text{ or } |s_j(\hat{\beta})| \neq 0 \text{ for some } 0 \leq j \leq q\}$,
- $F_{n4} = \{|\hat{\beta}_j| < a\lambda, \text{ for some } 1 \leq j \leq q\}$.

Denote the corresponding probability as $P_{ni} = \Pr(F_{ni}), i = 1, 2, 3, 4$. P_{n1} represents the localizability of the problem. When we have an appropriate initial estimator, we expect P_{n1} to converge to 0 as $n \rightarrow \infty$. P_{n2} is the probability that the true signal is too small to be detected by any method. P_{n3} describes the behavior of the subgradients at the oracle estimator. As stated in Theorem 3.1, there exists a subgradient such that its components corresponding to irrelevant variables are near 0 and those corresponding to relevant variables are exactly 0, so P_{n3} cannot be too large. P_{n4} has to do with the magnitude of the oracle estimator on relevant variables. Under regularity conditions, the oracle estimator will detect the true signals and hence P_{n4} will be very small.

Now we provide conditions for the LLA algorithm to find the oracle estimator $\hat{\beta}$ in the nonconvex penalized SVMs based on P_{n1}, P_{n2}, P_{n3} and P_{n4} .

THEOREM 3.3. *With probability at least $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$, the LLA algorithm initiated by $\tilde{\beta}^{(0)}$ finds the oracle estimator $\hat{\beta}$ after two iterations. Furthermore, if (A1)-(A8) hold, $\lambda = o(n^{-(1-c_2)/2})$ and $\log(p)q \log(n)n^{-\frac{1}{2}} = o(\lambda)$, then $P_{n2} \rightarrow 0, P_{n3} \rightarrow 0$ and $P_{n4} \rightarrow 0$ as $n \rightarrow \infty$.*

The first part of Theorem 3.3 provides a nonasymptotic lower bound on the probability that the LLA algorithm converges to the oracle estimator. As we will show in the appendix, if none of the events F_{ni} happen, the LLA algorithm initiated with $\tilde{\beta}^{(0)}$ will find the oracle estimator in the first iteration, and in the second iteration it will find the oracle estimator again and thus claim convergence. Therefore, the LLA algorithm can identify the oracle estimator after two iterations with one single correction and another step to stop the algorithm, and this result holds generally without the Conditions (A1)-(A8).

The second part of Theorem 3.3 indicates that under Conditions (A1)-(A8), the lower bound is determined only by the limiting behavior of the initial estimator. As long as an appropriate initial estimator is available, the problem of selecting the oracle estimator from potential multiple local minimizers is addressed. Let $\hat{\beta}^{L_1}$ be the solution to the

L_1 -penalized SVM. When the initial estimator $\tilde{\beta}^{(0)}$ is taken to be $\widehat{\beta}^{L_1}$ and the following Condition (A9) holds, by Theorem 3.3 the oracle estimator can be identified even in the ultra-high-dimensional setting. The result is summarized in the following Corollary.

$$(A9) \Pr(|\widehat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

COROLLARY 3.1. *Let $\widehat{\beta}(\lambda)$ be the solution found by the LLA algorithm initiated by $\widehat{\beta}^{L_1}$ after two iterations. Assume the same conditions in Theorem 3.3 and (A9) hold, then*

$$\Pr\{\widehat{\beta}(\lambda) = \widehat{\beta}\} \rightarrow 1 \text{ as } n \rightarrow \infty.$$

In the ultra-high-dimensional case, one may require more stringent conditions to guarantee (A9). For the nonconvex penalized least square regression, one can use the LASSO solution (Tibshirani, 1996) as the initial estimator and (A9) holds if one can further assume the restricted eigenvalue condition of the design matrix (Bickel et al., 2009). However, it is still largely unknown whether this conclusion also applies to a nonsmooth loss function. Without imposing any new regularity conditions, we next prove that in the case $p = o(\sqrt{n})$, the solution to the L_1 -penalized SVM satisfies (A9) under conditions quite similar to (A1)-(A8).

The following regularity Conditions are modified from (A1)-(A8). Conditions (A3), (A7)-(A8) are the same as aforementioned.

(A1*) The densities of \mathbf{X}^* given $Y = +1$ and $Y = -1$ are continuous and have a common support in \mathcal{R}^p .

$$(A2*) \mathbb{E}[X_j^2] < \infty \text{ for } 1 \leq j \leq p.$$

$$(A4*) p_n = O(n^{c_1}) \text{ for some } 0 \leq c_1 < 1/2.$$

(A5*) There exists a constant $M_1 > 0$ such that $\lambda_{max}(n^{-1}\mathbf{X}^T\mathbf{X}) \leq M_1$. It is further assumed that $\max_{1 \leq i \leq n} \|\mathbf{X}_i\| = O_p(\sqrt{p_n} \log n)$, (\mathbf{X}_i, Y_i) are in general position (Koenker, 2005, sect. 2.2), X_{ij} are sub-Gaussian random variables for $1 \leq i \leq n, q_n + 1 \leq j \leq p_n$.

$$(A6*) \lambda_{min}(H(\beta_0)) \geq M_3 \text{ for some constant } M_3 > 0.$$

Under the new regularity conditions, we can conclude that the solution to the L_1 -penalized SVM is an appropriate initial estimator. Combined with Theorem 3.3, the LLA algorithm initiated with a zero vector can identify the oracle estimator with one more iteration. The results are summarized in the following Theorem.

THEOREM 3.4. *Assume $\widehat{\beta}^{L_1}$ is the solution to the L_1 -penalized SVM with tuning parameter c_n . If the modified conditions hold, $\lambda = o(n^{-(1-c_2)/2})$, $p \log(n)n^{-1/2} = o(\lambda)$ and $c_n = o(n^{-1/2})$, then we have $\Pr(|\widehat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p) \rightarrow 0$ as $n \rightarrow \infty$.*

Further, the LLA algorithm initiated by $\widehat{\beta}^{L_1}$ finds the oracle estimator in two iterations with probability tending to one. That is, $\Pr\{\widehat{\beta}(\lambda) = \widehat{\beta}\} \rightarrow 1$ as $n \rightarrow \infty$.

Note that Theorem 3.4 can guarantee that the LLA algorithm initialized by the $\widehat{\beta}^{L_1}$ identifies the oracle estimator with high probability only when $p = o(\sqrt{n})$. However, our empirical studies suggest that even for cases with p much larger than n , the LLA algorithm initiated by $\widehat{\beta}^{L_1}$ usually converges within two iterations and the identified local minimizer has acceptable performance.

4. Implementation and tuning

To solve the nonconvex penalized SVMs, we use the LLA algorithm. More explicitly, we start with an initial value $\{\tilde{\beta}^{(0)} : \tilde{\beta}_j^{(0)} = 0, j = 1, 2, \dots, p\}$. At each step $t \geq 1$, we update by solving

$$\min_{\beta} \left\{ n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \beta)_+ + \sum_{j=1}^p p'_\lambda(|\tilde{\beta}_j^{(t-1)}|) |\beta_j| \right\}, \quad (6)$$

where $p'_\lambda(\cdot)$ denotes the derivative of $p_\lambda(\cdot)$. Following the literature, when $\tilde{\beta}_j^{(t-1)} = 0$, we take $p'_\lambda(0)$ as $p'_\lambda(0+) = \lambda$. The LLA algorithm is an instance of the majorize-minimize (MM) algorithm and converges to a local minimizer of the nonconvex objective function.

With slack variables, the convex optimization problem in (6) can be easily recast as a linear programming (LP) problem

$$\min_{\xi, \eta, \beta} \left\{ n^{-1} \sum_{i=1}^n W_i \xi_i + \sum_{j=1}^p p'_\lambda(|\tilde{\beta}_j^{(t-1)}|) \eta_j \right\}$$

subject to

$$\begin{aligned} \xi_i &\geq 0; i = 1, 2, \dots, n, \\ \xi_i &\geq 1 - Y_i \mathbf{X}_i^T \beta; i = 1, 2, \dots, n, \\ \eta_j &\geq \beta_j, \eta_j \geq -\beta_j; j = 1, 2, \dots, p. \end{aligned}$$

We propose using the stopping rule that $p'_\lambda(|\tilde{\beta}_j^{(t-1)}|)$ stabilizes for $j = 1, 2, \dots, p$, namely, when $\sum_{j=1}^p (p'_\lambda(|\tilde{\beta}_j^{(t-1)}|) - p'_\lambda(|\tilde{\beta}_j^{(t)}|))^2$ is sufficiently small.

For the choice of tuning parameter λ , Claeskens et al. (2008) suggested the SVM information criterion (SVMIC). For a subset S of $\{1, 2, \dots, p\}$, the SVMIC is defined as

$$\text{SVMIC}(S) = \sum_{i=1}^n \xi_i + \log(n)|S|,$$

where $|S|$ is the cardinality of S and $\xi_i, i = 1, 2, \dots, n$ denote the corresponding optimal slack variables. This criterion directly follows the spirit of the Bayesian information criterion (BIC) by Schwarz (1978). Chen and Chen (2008) showed that BIC can be too liberal when the model space is large and proposed the extended BIC (EBIC):

$$\text{EBIC}_\gamma(S) = -2 \log \text{Likelihood} + \log(n)|S| + 2\gamma \binom{p}{|S|}, \quad 0 \leq \gamma \leq 1.$$

By combining these ideas, we suggest the SVM-extend BIC (SVMIC_γ)

$$\text{SVMIC}_\gamma(S) = \sum_{i=1}^n 2W_i \xi_i + \log(n)|S| + 2\gamma \binom{p}{|S|}, \quad 0 \leq \gamma \leq 1.$$

Note that SVMIC_γ reduces to SVMIC when $\gamma = 0$ and $w = 0.5$. We use $\gamma = 0.5$ as suggested by Chen and Chen (2008) and choose the λ that minimizes SVMIC_γ .

5. Simulation and real data examples

We carry out Monte Carlo studies to evaluate the finite-sample performance of the non-convex penalized SVMs. We compare the performance of SCAD-penalized SVM, MCP-penalized SVM, standard L_2 SVM, L_1 -penalized SVM, adaptively weighted L_1 -penalized SVM (Zou, 2007) and hybrid Huberized SVM (Wang et al., 2007) (denoted by SCAD-svm, MCP-svm, L_2 -svm, L_1 -svm, Adap L_1 -svm, and Hybrid-svm, respectively) with weight parameter $w = 0.5$. The main interest here is the ability to identify the relevant covariates and the control of test error when $p > n$.

5.1. Simulation study

We consider two data generation processes. The first, adapted from Park et al. (2012), is essentially a standard linear discriminant analysis (LDA) setting. The second is related to probit regression.

- Model 1: $\Pr(Y = 1) = \Pr(Y = -1) = 0.5$, $\mathbf{X}^*|Y = 1 \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}^*|Y = -1 \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q = 5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \dots, 0)^T \in \mathbf{R}^p$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \dots, p$ and $\sigma_{ij} = \rho = -0.2$ for $1 \leq i \neq j \leq q$. The Bayes rule is $\text{sign}(2.67X_1 + 2.83X_2 + 3X_3 + 3.17X_4 + 3.33X_5)$ with Bayes error: 6.3%.
- Model 2: $\mathbf{X}^* \sim MN(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ii} = 1$ for $i = 1, 2, \dots, p$ and $\sigma_{ij} = 0.4^{|i-j|}$ for $1 \leq i \neq j \leq p$, $\Pr(Y = 1|\mathbf{X}^*) = \Phi((\mathbf{X}^*)^T \boldsymbol{\beta}^*)$ where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\boldsymbol{\beta}^* = (1.1, 1.1, 1.1, 1.1, 0, \dots, 0)^T$, $q = 4$. The Bayes rule is $\text{sign}(X_1 + X_2 + X_3 + X_4)$ with Bayes error 10.4%.

We consider different (n, p) settings for each data generation process with p much larger than n . Similarly to Mazumder et al. (2011), an independent tuning dataset of size $10n$ is generated to tune any regularization parameter for all methods by minimizing the estimated prediction error calculated over the tuning dataset. We also report the performance of the SCAD- and MCP-penalized SVMs using SVMIC_γ to select the tuning parameter λ . Notice that tuning by a large independent tuning dataset of $10n$ approximates the ideal “population tuning”, which is usually not available in practice. By giving all the other methods the best possible tuning, we are controlling the effect of tuning parameter selection and conservative about the performance of the nonconvex penalized SVMs tuned by SVMIC_γ . As we will see later, the results of SCAD- and MCP-penalized SVMs using the independent tuning dataset are slightly better than the corresponding results using SVMIC_γ tuning; and all other methods have no ability to select the correct model exactly, even with an unrealistically good tuning parameter. The range of λ is $\{2^{-6}, \dots, 2^3\}$. We use $a=3.7$ for the SCAD penalty and $a = 3$ for the MCP as suggested in the literature. We generate an independent test dataset of size n to report the estimated test error. The columns “Signal” and “Noise” summarize the average number of selected relevant and irrelevant covariates, respectively. The numbers in the “Correct” column summarize the percentages of selecting the exactly true model over replications.

Table 1 shows the results of Model 1 for different (n, p) settings. The numbers in parentheses are the corresponding standard errors based on 100 replications. When tuned by using an independent tuning set of size $10n$, both SCAD- and MCP-penalized SVMs identify more relevant variables than any other methods and they also reduce the number

Table 1. Simulation results for Model 1

Method	n	p	Signal	Noise	Correct	Test Error
SCAD-svm	100	400	4.94(0.03)	0.89(0.19)	64%	8.71%(0.4%)
	100	800	4.93(0.03)	0.93(0.14)	51%	9.39%(0.4%)
	200	800	5.00(0.00)	0.09(0.05)	96%	7.20%(0.2%)
	200	1600	5.00(0.00)	0.07(0.04)	96%	7.24%(0.2%)
MCP-svm	100	400	4.90(0.04)	0.88(0.17)	53%	8.96%(0.4%)
	100	800	4.92(0.03)	1.37(0.20)	40%	10.59%(0.5%)
	200	800	5.00(0.00)	0.06(0.04)	97%	7.30%(0.2%)
	200	1600	5.00(0.00)	0.09(0.03)	92%	6.79%(0.2%)
SCAD-svm ^(SVMIC$_{\gamma}$)	100	400	4.64(0.08)	0.48(0.11)	64%	10.32%(0.6%)
	100	800	4.63(0.09)	0.57(0.09)	52%	11.68%(0.7%)
	200	800	5.00(0.00)	0.03(0.02)	97%	7.24%(0.2%)
	200	1600	4.99(0.01)	0.05(0.03)	95%	7.23%(0.2%)
MCP-svm ^(SVMIC$_{\gamma}$)	100	400	4.46(0.10)	0.44(0.08)	45%	11.81%(0.6%)
	100	800	4.34(0.11)	0.68(0.11)	38%	13.13%(0.7%)
	200	800	5.00(0.00)	0.09(0.03)	92%	7.34%(0.2%)
	200	1600	5.00(0.00)	0.06(0.03)	95%	7.19%(0.2%)
L_1 -svm	100	400	4.87(0.05)	32.97(1.47)	0%	16.08%(0.5%)
	100	800	4.63(0.07)	44.34(2.18)	0%	19.71%(0.6%)
	200	800	5.00(0.00)	21.33(1.70)	0%	9.59%(0.3%)
	200	1600	4.99(0.01)	33.37(0.96)	0%	10.88%(0.3%)
Hybrid-svm	100	400	4.78(0.05)	24.74(1.37)	0%	16.34%(0.5%)
	100	800	4.62(0.06)	27.16(1.30)	0%	19.93%(0.6%)
	200	800	5.00(0.00)	12.86(0.99)	0%	9.93%(0.2%)
	200	1600	4.99(0.01)	10.85(0.98)	0%	10.53%(0.3%)
Adap L_1 -svm	100	400	4.39(0.08)	13.14(0.90)	0%	16.76%(0.5%)
	100	800	3.99(0.08)	12.50(0.69)	0%	20.19%(0.6%)
	200	800	4.86(0.04)	3.93(0.25)	1%	10.04%(0.3%)
	200	1600	4.49(0.06)	1.01(0.09)	4%	13.43%(0.4%)
L_2 -svm	100	400	5.00(0.00)	395.00(0.00)	0%	39.23%(0.5%)
	100	800	5.00(0.00)	795.00(0.00)	0%	42.99%(0.5%)
	200	800	5.00(0.00)	795.00(0.00)	0%	39.22%(0.3%)
	200	1600	5.00(0.00)	1595.00(0.00)	0%	42.50%(0.4%)

Table 2. Simulation results for Model 2 with $n = 250$ and $p = 800$

Method	Signal	Noise	Correct	Test Error
SCAD-svm	3.99(0.01)	0.26(0.08)	92.5%	11.4%(0.1%)
MCP-svm	3.99(0.01)	0.17(0.07)	93.5%	11.3%(0.1%)
SCAD-svm ^(SVMIC$_{\gamma}$)	3.96(0.02)	0.05(0.02)	94%	11.5%(0.1%)
MCP-svm ^(SVMIC$_{\gamma}$)	3.98(0.01)	0.07(0.02)	92.5%	11.4%(0.1%)
L_1 -svm	4.00(0.00)	6.84(0.42)	7.5%	12.4%(0.1%)
Hybrid-svm	4.00(0.00)	4.03(0.41)	10.5%	11.9%(0.1%)
Adap L_1 -svm	4.00(0.00)	2.90(0.28)	38%	11.8%(0.1%)
L_2 -svm	4.00(0.00)	796.00(0.00)	0%	32.5%(0.2%)

Table 3. Classification error of MAQC-II dataset

Method	Test error	Genes
SCAD-svm	9.8%(0.2%)	2.06(0.43)
MCP-svm	9.6%(0.2%)	1.04(0.02)
L_1 -svm	10.9%(0.2%)	28.74(1.36)
Adap L_1 -svm	13.1%(0.2%)	34.30(1.03)
Hybrid-svm	10.0%(0.1%)	1391.60(94.86)
L_2 -svm	10.8%(0.2%)	3000.00(0.00)

of falsely selected variables dramatically. When tuned by $SVMIC_\gamma$, SCAD- and MCP-penalized SVMs select slightly fewer signals when $n = 100$, but this is based on the fact that other methods select a much larger model without proper control of noise. A large proportion of the missed relevant covariates are from X_1 as it has the weakest signal. Notice that $SVMIC_\gamma$ performs almost the same as “population tuning” when n is relatively large. In general, the nonconvex penalized SVMs have an overwhelmingly high probability to select the exact true mode as n and p increase, while other methods show very weak, if any at all, ability to recover the exact true model. This is consistent with our theory of asymptotic oracle property of nonconvex penalized SVMs. The test errors of SCAD- and MCP-penalized SVMs are uniformly smaller than those of any other method in all settings, even in the settings with a small sample size $n = 100$ and tuned by $SVMIC_\gamma$, where they select slightly fewer signals. This is due to the fact that in high-dimensional classification problem, a large number of falsely selected variables will greatly blur the prediction power of the relevant variables.

Table 2 shows the results of Model 2 for $n = 250$ and $p = 800$. The numbers in the parentheses are the corresponding standard errors based on 200 replications. We observe similar performance patterns in terms of both variable selection and prediction error. Due to the higher correlation between signal and noise, in Model 2 it is generally more difficult to select the relevant covariates. Both SCAD- and MCP-penalized SVM still have reasonable performance in identifying the underlying true model and result in more accurate prediction. Note that under this data generation process the adaptively weighted L_1 -penalized SVM behaves similar to nonconvex penalized SVMs, though its oracle property is largely unknown.

5.2. Real data application

We next use a real dataset to illustrate the performance of the nonconvex penalized SVM. This dataset is part of the MicroArray Quality Control (MAQC)-II project, available at the GEO database with accession number GSE20194. It contains 278 patient samples from two classes: 164 with have positive estrogen receptor (ER) status and 114 with have negative estrogen receptor (ER) status. Each sample is described by 22283 genes.

The original data have been standardized for each predictor. To reduce the computational burden, only the 3000 genes with largest absolute values of the two sample t -statistics are used. Such simplification has been considered in Cai and Liu (2011). Though only 3000 genes are used, the classification result is satisfactory. We randomly split the data into an equally balanced training set with 50 samples with positive ER status and 50 samples with negative ER status, and the rest were designated as the test set. As in the simulation study, we use $a=3.7$ for the SCAD penalty and $a=3$ for the MCP penalty. To get a fair comparison, a 5-fold cross validation is implemented on the training set to select a tuning parameter by

a grid search over $\{2^{-15}, \dots, 2^3\}$ for all methods and the test error is calculated on the test data. The above procedure is repeated 100 times.

Table 3 summarizes the average classification error and number of selected genes. The numbers in the parentheses are the corresponding standard errors based on 100 replications. Nonconvex penalized SVMs achieve significantly lower test error than all the other methods except for the doubly penalized hybrid SVM. Although the doubly penalized hybrid SVM performs similar to SCAD- and MCP-penalized SVMs in terms of test error, it selects a much more complex model in general. In addition, the number of genes selected by nonconvex penalized SVMs is stable, while the model size selected by hybrid SVM ranges from 102 genes to 2576 genes across the 100 replications. Such stability is desirable, so that the procedure is robust to the random partition of the data. The numerical results confirm that SCAD- and MCP-penalized SVMs can achieve both promising prediction power and excellent gene selection ability.

6. Discussion

In this article we study the nonconvex penalized SVMs with a diverging number of covariates in terms of variable selection. When the true model is sparse, under some regularity conditions, we prove that it enjoys the oracle property. That is, one of the local minimizers of the nonconvex penalized SVM behaves like the oracle estimator as if the true sparsity is known in advance and only the relevant variables are used to form the decision boundary. We also show that as long as we have an appropriate initial estimator, we can identify the oracle estimator with probability tending to one.

6.1. Connection to Bayes rule

In this paper, the true model and the oracle property are built on β_0 , which is the minimizer of the population version of the hinge loss. This definition has a strong connection to the Bayes rule, which is theoretically optimal if the underlying distribution is known. In the equal-weight case ($w=1/2$), the Bayes rule is given by $\text{sign}(\mathbf{X}^T \beta_{\text{Bayes}})$ with $\beta_{\text{Bayes}} = \arg \min_{\beta} \mathbb{E}\{I(\text{sign}(\mathbf{X}^T \beta) \neq Y)\}$. To appreciate the connection, we first note that β_{Bayes} and β_0 are equivalent to each other in the important special case of Fisher linear discriminant analysis. Indeed, consider an informative example setting with $\pi_+ = \pi_- = 1/2$, $\mathbf{X}^*|Y = +1 \sim N(\boldsymbol{\mu}_+, \boldsymbol{\Sigma})$ and $\mathbf{X}^*|Y = -1 \sim N(\boldsymbol{\mu}_-, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ denote different mean vectors for two classes and $\boldsymbol{\Sigma}$ a same variance covariance matrix. It is known that in this case the Bayes rule boundary is given by

$$(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T \boldsymbol{\Sigma}^{-1} \{\mathbf{x}^* - 1/2(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\} = 0.$$

Note that β_0 as the minimizer of the population hinge loss satisfies the gradient condition

$$\mathbf{S}(\beta_0) = -\mathbb{E}\{I(1 - Y\mathbf{X}^T \beta_0 \geq 0)Y\mathbf{X}\} = 0,$$

which is equivalent to following equations:

$$\begin{aligned} \Pr(1 - \mathbf{X}^T \beta_0 \geq 0|Y = +1) &= \Pr(1 + \mathbf{X}^T \beta_0 \geq 0|Y = -1), \\ \mathbb{E}\{I(1 - \mathbf{X}^T \beta_0 \geq 0)\mathbf{X}^*|Y = +1\} &= \mathbb{E}\{I(1 + \mathbf{X}^T \beta_0 \geq 0)\mathbf{X}^*|Y = -1\}. \end{aligned} \quad (7)$$

For any $\beta_{0,\perp}^*$ that satisfies $(\beta_0^*)^T \Sigma \beta_{0,\perp}^* = 0$, $(\mathbf{X}^*)^T \beta_0^*$ and $(\mathbf{X}^*)^T \beta_{0,\perp}^*$ are conditionally independent given Y and thus we can decompose the conditional expectation in (7) into two parts. It can be seen from (7) that

$$\begin{aligned} \beta_{00} &= -1/2(\beta_0^*)^T(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-), \\ (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T \beta_{0,\perp}^* &= 0, \forall \beta_{0,\perp}^* \text{ satisfying } (\beta_0^*)^T \Sigma \beta_{0,\perp}^* = 0. \end{aligned}$$

That is, $(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ lies in the space spanned by $\Sigma \beta_0^*$. The decision boundary defined by the true value is then

$$\mathbf{x}^T \beta_0 \equiv C(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T \Sigma^{-1} \{\mathbf{x}^* - 1/2(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)\} = 0$$

for some constant C . Therefore, the Bayes rule is equivalent to β_0 .

In more general settings, β_{Bayes} and β_0 may not be the same. However, Lin (2000) showed that the nonlinear SVM approaches the Bayes rule in a direct fashion, and its expected misclassification rate quickly converges to that of the Bayes rule even though its extension to linear SVM is largely unknown. Furthermore, denote $R(f)$ and $R_0(f)$ to be the risk in terms of the 0-1 loss and hinge loss, respectively, for any measurable f ; that is, $R(f) = \mathbb{E}\{I(\text{sign}(f(\mathbf{X})) \neq Y)\}$ and $R_0(f) = \mathbb{E}\{(1 - Yf(\mathbf{X}))_+\}$. It is known that minimizing $R(f)$ directly is very difficult because minimizing the empirical 0-1 loss is infeasible in practice (Bartlett et al., 2006). Instead, we can always shift the target from the 0-1 loss to a convex surrogate such as the hinge loss. Assume that the minimizers of $R(f)$ and $R_0(f)$ are both linear functions, and by definitions they are $\mathbf{X}^T \beta_{Bayes}$ and $\mathbf{X}^T \beta_0$, respectively. By Theorem 1 of Bartlett et al. (2006), we have the optimal excess risk upper bound

$$R(\mathbf{X}^T \beta) - R(\mathbf{X}^T \beta_{Bayes}) \leq R_0(\mathbf{X}^T \beta) - R_0(\mathbf{X}^T \beta_0)$$

for any β . Hence pursuing oracle property on β_0 has the potential to efficiently control the excess risk. As can be seen in this paper, the main advantages of working with the hinge loss instead of the 0-1 loss are the theoretical tractability and convenience in practical implementation.

6.2. Other issues

As one referee pointed out, the objective function (2) in the definition of our oracle estimator is piecewise linear and may have multiple minimizers. The same issue applies to the L_1 -penalized SVM and the nonconvex penalized SVM. Based on our theoretical development, non-uniqueness of the minimizer of (2) is not essential. When the minimizer is not unique, our theoretical results still hold for any particular minimizer. In this sense, we can first use the nonconvex penalized SVM to identify important predictors. In the next step, to obtain a unique classifier, a refitting can be applied by using the standard L_2 penalized SVM on those identified important predictors. For Model 1 in Section 5.1, we considered this refitting. This additional refitting step does not lead to much improvement: it reduces the average test errors in some settings but not in others. Thus the refitting result is not reported here.

An alternative approach to deal with this non-uniqueness is to consider a joint penalty by using both a nonconvex penalty and a standard L_2 penalty. The objective function then becomes

$$n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \beta)_+ + \sum_{j=1}^{p_n} p_{\lambda_{1n}}(|\beta_j|) + \sum_{j=1}^{p_n} \lambda_{2n} \beta_j^2$$

Table 4. Comparison between SCAD and joint penalized SVMs using Model 1

Method	p	Signal	Noise	Correct	Test Error
SCAD-svm	600	5.00(0.00)	0.17(0.07)	93%	7.04%(0.2%)
	800	5.00(0.00)	0.13(0.06)	93%	7.25%(0.2%)
Joint SCAD+ L_2 -svm	600	5.00(0.00)	1.22(0.28)	65%	7.12%(0.2%)
	800	5.00(0.00)	2.64(0.62)	50%	7.10%(0.2%)

for two different tuning parameters λ_{1n} and λ_{2n} . The corresponding oracle estimator is then defined as the minimizer of the objective function for the standard L_2 SVM using only the relevant covariates. One advantage of this joint penalty formulation over the method proposed in this paper is that the uniqueness of the oracle estimator is guaranteed in the finite sample case. However, it involves simultaneously selecting two tuning parameters, and this may not be convenient in practice. We conduct a simple numerical experiment using Model 1 in Section 5.1 with $n = 200$ and $p = 600$ or 800 . The simulation results are summarized in Table 4. As shown in Table 4, our numerical example suggests that the performance of this joint penalty method is similar to the approach proposed in this paper.

Several issues remain unsolved. In this article we only study the SVMs in nonseparable cases in the limit. Although the nonseparable cases are important in practical applications, it would be interesting to show similar results for separable cases. The asymptotic analysis of separable cases requires the positiveness of the limit of the regularization term, which is different from the analysis in this article. Another issue is the availability of an appropriate initial estimator in the ultra-high dimension. Our empirical studies suggest that the solution to the L_1 -penalized SVM is an appropriate initial estimator and the LLA algorithm converges very quickly even for cases with p much larger than n . However it still lacks theoretical justification in this case since our Theorem 3.4 only provides theoretical support for the case with $p = o(\sqrt{n})$. Future research topic could extend Theorem 3.4 to the case with a faster diverging dimensionality.

7. Appendix

We first prove Lemma 3.1.

PROOF (PROOF OF LEMMA 3.1). Let $l(\beta_1) = n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{Z}_i^T \beta_1)_+$. Note that $\hat{\beta}_1 = \arg \min_{\beta_1} l(\beta_1)$. We will show that when $\forall \eta > 0$, there exists a constant Δ such that for all n sufficiently large, $\Pr\{\inf_{\|\mathbf{u}\|=\Delta} l(\beta_{01} + \sqrt{q/n}\mathbf{u}) > l(\beta_{01})\} \geq 1 - \eta$. Because $l(\beta_1)$ is convex, with probability at least $1 - \eta$, $\hat{\beta}_1$ is in the ball $\{\beta_1 : \|\beta_1 - \beta_{01}\| \leq \Delta \sqrt{q/n}\}$. Denote $\Lambda_n(\mathbf{u}) = nq^{-1}\{l(\beta_{01} + \sqrt{q/n}\mathbf{u}) - l(\beta_{01})\}$. Observe that $\mathbb{E}\{\Lambda_n(\mathbf{u})\} = nq^{-1}\{L(\beta_{01} + \sqrt{q/n}\mathbf{u}) - L(\beta_{01})\}$. Recall also that $\beta_0 = \arg \min_{\beta} \mathbb{E}\{W(1 - Y\mathbf{X}^T\beta)\}$. If we restrict the last $p - q$ elements to be 0, it can be easily seen that $\beta_{01} = \arg \min_{\beta_1} \mathbb{E}\{W(1 - Y\mathbf{Z}^T\beta_1)\} = \arg \min_{\beta_1} L(\beta_1)$, thus $S(\beta_{01}) = 0$. By Taylor series expansion of $L(\beta_1)$ around β_{01} , we have $\mathbb{E}\{\Lambda_n(\mathbf{u})\} = \frac{1}{2}\mathbf{u}^T H(\tilde{\beta})\mathbf{u} + o_p(1)$, where $\tilde{\beta} = \beta_{01} + \sqrt{q/nt}\mathbf{u}$ for some $0 < t < 1$. As shown in Koo et al. (2008), for $0 \leq j, k \leq q$, the (j, k) -th element of the Hessian Matrix $H(\beta_{01})$ is continuous given (A1) and (A2); thus $H(\beta)$ is continuous. By continuity of $H(\beta)$ at β_{01} , then $\frac{1}{2}\mathbf{u}^T H(\tilde{\beta})\mathbf{u} = \frac{1}{2}\mathbf{u}^T H(\beta_{01})\mathbf{u} + o(1)$ as $n \rightarrow \infty$. Define $\mathbf{W}_n = -\sum_{i=1}^n \zeta_i W_i Y_i \mathbf{Z}_i$ where $\zeta_i = I(1 - Y_i \mathbf{Z}_i^T \beta_{01} \geq 0)$. Recall that $S(\beta_{01}) = -\mathbb{E}[\zeta_i W_i Y_i \mathbf{Z}_i] = 0$. If we define

$$R_{i,n}(\mathbf{u}) = W_i(1 - Y_i \mathbf{Z}_i^T (\beta_{01} + \frac{\sqrt{q}}{\sqrt{n}}\mathbf{u}))_+ - W_i(1 - Y_i \mathbf{Z}_i^T \beta_{01})_+ + \zeta_i W_i Y_i \mathbf{Z}_i^T \sqrt{q/n}\mathbf{u}$$

then we have

$$\Lambda_n(\mathbf{u}) = \mathbb{E}\{\Lambda_n(\mathbf{u})\} + \mathbf{W}_n^T \mathbf{u} / \sqrt{qn} + q^{-1} \sum_{i=1}^n [R_{i,n}(\mathbf{u}) - \mathbb{E}\{R_{i,n}(\mathbf{u})\}]. \quad (8)$$

Then similar to Equation (28) in Koo et al. (2008) we have

$$q^{-2} \sum_{i=1}^n E[|R_{i,n}(\mathbf{u}) - E\{R_{i,n}(\mathbf{u})\}|^2] \leq C\Delta^2 E\{q^{-1}(1 + \|\mathbf{Z}\|^2)U(\sqrt{1 + \|\mathbf{Z}\|^2}\Delta\sqrt{q/n})\},$$

where $U(t) = I(|1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01}| < t)$. (A2) implies that $E\{q^{-1}(1 + \|\mathbf{Z}\|^2)\} < \infty$. Hence, for any $\epsilon > 0$, we can choose a positive constant C such that $E[q^{-1}(1 + \|\mathbf{Z}\|^2)I\{q^{-1}(1 + \|\mathbf{Z}\|^2) > C\}] < \epsilon/2$, then

$$\begin{aligned} & E\{q^{-1}(1 + \|\mathbf{Z}\|^2)U(\sqrt{1 + \|\mathbf{Z}\|^2}\Delta\sqrt{q/n})\} \\ & \leq E[q^{-1}(1 + \|\mathbf{Z}\|^2)I\{q^{-1}(1 + \|\mathbf{Z}\|^2) > C\}] + C\Pr(|1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01}| < C\Delta\sqrt{q/n}). \end{aligned}$$

We can take a large N such that $\Pr(|1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01}| < C\Delta\sqrt{q/n}) < \frac{\epsilon}{2C}$ for all $n > N$ by (A4). This proves that $q^{-2} \sum_{i=1}^n E\{|R_{i,n}(\mathbf{u}) - E[R_{i,n}(\mathbf{u})]|^2\} \rightarrow 0$ as $n \rightarrow \infty$. Observe that $E(\mathbf{W}_n^T \mathbf{u} / \sqrt{qn}) = 0$, and

$$\text{Var}(\mathbf{W}_n^T \mathbf{u} / \sqrt{qn}) \leq Cn^{-1}q^{-1} \sum_{i=1}^n (\mathbf{Z}_i^T \mathbf{u})^2 \leq Cq^{-1}\lambda_{\max}(n^{-1}\mathbf{X}_A^T \mathbf{X}_A)\|\mathbf{u}\|^2 \rightarrow 0$$

as $n \rightarrow \infty$. Therefore, the first term of (8) will dominate other terms as $n \rightarrow \infty$. By (A6) we have $\frac{1}{2}\mathbf{u}^T H(\boldsymbol{\beta}_{01})\mathbf{u} > 0$. Thus we can choose a sufficiently large Δ such that $\Lambda_n(\mathbf{u}) > 0$ with probability $1 - \eta$ for $\|\mathbf{u}\| = \Delta$ and all sufficiently large n .

The proof of Theorem 3.1 relies on the following Lemmas.

LEMMA 7.1.

$$\Pr\left\{\max_{q+1 \leq j \leq p} n^{-1} \left| \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0) \right| > \lambda/2\right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

PROOF. Recall that $\mathbb{E}\{W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\} = 0$. By (A5) and Lemma 14.9 of Bühlmann and Van De Geer (2011), we have $\Pr\{n^{-1} |\sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)| > \lambda/2\} \leq \exp(-Cn\lambda^2)$. Note that

$$\begin{aligned} & \Pr\left\{\max_{q+1 \leq j \leq p} n^{-1} \left| \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0) \right| > \lambda/2\right\} \\ & = \Pr\left\{\cup_{q+1 \leq j \leq p} \left\{n^{-1} \left| \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0) \right| > \lambda/2\right\}\right\} \leq p \exp(-Cn\lambda^2) \rightarrow 0 \end{aligned}$$

as $n \rightarrow \infty$ by the fact that $\log(p) = o(n\lambda^2)$.

LEMMA 7.2. For any $\Delta > 0$,

$$\Pr\left\{\max_{q+1 \leq j \leq p} \sup_{\|\beta_1 - \beta_{01}\| \leq \Delta \sqrt{q/n}} \left| \sum_{i=1}^n W_i Y_i X_{ij} [I(1 - Y_i \mathbf{Z}_i^T \beta_1 \geq 0) - I(1 - Y_i \mathbf{Z}_i^T \beta_{01} \geq 0)] - \Pr(1 - Y_i \mathbf{Z}_i^T \beta_1 \geq 0) + \Pr(1 - Y_i \mathbf{Z}_i^T \beta_{01} \geq 0) \right| > n\lambda \right\} \rightarrow 0$$

as $n \rightarrow \infty$.

PROOF. We generalize an approach by Welsh (1989). We cover the ball $\{\beta_1 : \|\beta_1 - \beta_{01}\| \leq \Delta \sqrt{q/n}\}$ with a net of balls with radius $\Delta \sqrt{q/n^5}$. It can be shown that this net can be constructed with cardinality $N \leq dn^{4q}$ for some $d > 0$. Denote the N balls by $B(\mathbf{t}_1), \dots, B(\mathbf{t}_N)$, where $\mathbf{t}_k, k = 1, \dots, N$ are the centers. Denote $\kappa_i(\beta_1) = 1 - Y_i \mathbf{Z}_i^T \beta_1$, and

$$\begin{aligned} J_{nj1} &= \sum_{k=1}^N \Pr\left(\left| \sum_{i=1}^n W_i Y_i X_{ij} [I\{\kappa_i(\mathbf{t}_k) \geq 0\} - I\{\kappa_i(\beta_{01}) \geq 0\}] - \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\} + \Pr\{\kappa_i(\beta_{01}) \geq 0\} \right| > n\lambda/2\right), \\ J_{nj2} &= \sum_{k=1}^N \Pr\left(\sup_{\tilde{\beta}_1 \in B(\mathbf{t}_k)} \left| \sum_{i=1}^n W_i Y_i X_{ij} [I\{\kappa_i(\tilde{\beta}_1) \geq 0\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\}] - \Pr\{\kappa_i(\tilde{\beta}_1) \geq 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\} \right| > n\lambda/2\right). \end{aligned}$$

Then by (A5),

$$\Pr\left(\sup_{\|\beta_1 - \beta_{01}\| \leq \Delta \sqrt{q/n}} \left| \sum_{i=1}^n W_i Y_i X_{ij} [I\{\kappa_i(\beta_1) \geq 0\} - I\{\kappa_i(\beta_{01}) \geq 0\}] - \Pr\{\kappa_i(\beta_1) \geq 0\} + \Pr\{\kappa_i(\beta_{01}) \geq 0\} \right| > n\lambda \right) \leq J_{nj1} + J_{nj2}.$$

To evaluate J_{nj1} , let $U_i = W_i Y_i X_{ij} [I\{\kappa_i(\mathbf{t}_k) \geq 0\} - I\{\kappa_i(\beta_{01}) \geq 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\} + \Pr\{\kappa_i(\beta_{01}) \geq 0\}]$. The U_i are independent mean-zero random variable, and $\text{Var}(U_i) = \mathbb{E}(U_i^2) = \mathbb{E}(U_i^2 | Y_i = 1) \Pr(Y_i = 1) + \mathbb{E}(U_i^2 | Y_i = -1) \Pr(Y_i = -1)$. Denote F and G the CDF of the conditional distribution of $\mathbf{Z}^T \beta_{01}$ given $Y = +1$ and $Y = -1$. Observe that

$$\begin{aligned} \mathbb{E}(U_i^2 | Y_i = 1) &\leq C \{F_i(1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))(1 - F_i(1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))) + F_i(1)(1 - F_i(1)) \\ &\quad - 2F_i(\min(1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k), 1)) + 2F_i(1)F_i(1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))\} \\ &\leq C |\mathbf{Z}_i^T(\mathbf{t}_k - \beta_{01})|, \end{aligned}$$

and it follows by (A8) that

$$\begin{aligned} &\mathbb{E}(U_i^2 | Y_i = -1) \\ &\leq C \{G_i(-1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))(1 - G_i(-1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))) + G_i(-1)(1 - G_i(-1)) \\ &\quad - 2(1 - G_i(\max(-1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k), -1))) + 2(1 - G_i(-1))(1 - G_i(-1 + \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k)))\} \\ &\leq C |\mathbf{Z}_i^T(\mathbf{t}_k - \beta_{01})|. \end{aligned}$$

Thus we have

$$\sum_{i=1}^n \text{Var}(U_i) \leq nC \max_i \|\mathbf{Z}_i\| \|\mathbf{t}_k - \beta_{01}\| = nO(\sqrt{q} \log(n))O(\sqrt{q/n}) = O(\sqrt{nq} \log(n)).$$

Applying Lemma 14.9 of Bühlmann and Van De Geer (2011), for some positive constant C_1 and C_2 under the assumptions on the rate of λ ,

$$J_{nj1} \leq 2N \exp\left(-\frac{n^2\lambda^2/4}{C_1\sqrt{nq}\log(n) + C_2n\lambda}\right) \leq C \exp\{4q\log(n) - Cn\lambda\}. \quad (9)$$

To evaluate J_{nj2} , note that $I(x \geq s)$ is decreasing in s . Denote

$$V_i = [I\{\kappa_i(\tilde{\beta}_1) \geq 0\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\} - \Pr\{\kappa_i(\tilde{\beta}_1) \geq 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\}].$$

We have $-B_i \leq V_i \leq A_i$ for any $\tilde{\beta}_1 \in B(\mathbf{t}_k)$, where

$$A_i = [I\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq \Delta\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\}],$$

$$B_i = [I\{\kappa_i(\mathbf{t}_k) \geq 0\} - I\{\kappa_i(\mathbf{t}_k) \geq \Delta\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\}].$$

Therefore, we have

$$\begin{aligned} & \Pr\left(\sup_{\tilde{\beta}_1 \in B(\mathbf{t}_k)} \left| \sum_{i=1}^n W_i Y_i X_{ij} [I\{\kappa_i(\tilde{\beta}_1) \geq 0\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\} - \Pr\{\kappa_i(\tilde{\beta}_1) \geq 0\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\}] \right| > n\lambda/2\right) \\ & \leq \Pr\left(C \max_i |X_{ij}| \sup_{\tilde{\beta}_1 \in B(\mathbf{t}_k)} \left| \sum_{i=1}^n V_i \right| > n\lambda/2\right) \leq \Pr\left(C \max_i |X_{ij}| \max\left(\sum_{i=1}^n A_i, \sum_{i=1}^n B_i\right) > n\lambda/2\right) \end{aligned}$$

by the fact that $A_i > 0, B_i > 0$. Note that

$$\begin{aligned} \sum_{i=1}^n A_i &= \sum_{i=1}^n [I\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\}] \\ & \quad + \sum_{i=1}^n [\Pr\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq \Delta\sqrt{q/n^5}\}] \end{aligned}$$

and

$$\begin{aligned} & \sum_{i=1}^n [\Pr\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq \Delta\sqrt{q/n^5}\}] \\ &= [F_i(1 + \Delta\sqrt{q/n^5} - \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k)) - F_i(1 - \Delta\sqrt{q/n^5} - \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))] \Pr(Y_i = 1) \\ & \quad + [G_i(-1 + \Delta\sqrt{q/n^5} - \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k)) - G_i(-1 - \Delta\sqrt{q/n^5} - \mathbf{Z}_i^T(\beta_{01} - \mathbf{t}_k))] \Pr(Y_i = -1) \\ & \leq Cn \log(q) \sqrt{q/n^5} \sqrt{q} = C \log(q) q n^{-3/2} \end{aligned}$$

by (A8). Denote

$$O_i = [I\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\} - \Pr\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} + \Pr\{\kappa_i(\mathbf{t}_k) \geq 0\}].$$

Thus for sufficiently large n by $\lambda = o(n^{-(1-c_2)/2})$ and A(7), we have

$$\sum_{k=1}^N \Pr\left(C \sum_{i=1}^n A_i > n\lambda/2\right) \leq \sum_{k=1}^N \Pr\left(C \sum_{i=1}^n O_i > n\lambda/2 - C \log(q) q n^{-3/2}\right) \leq \sum_{k=1}^N \Pr\left(C \sum_{i=1}^n O_i > n\lambda/4\right).$$

Notice that O_i are independent mean-zero random variables, and

$$\mathbb{E}(O_i^2) = \mathbb{E}[I\{\kappa_i(\mathbf{t}_k) \geq -\Delta\sqrt{q/n^5}\} - I\{\kappa_i(\mathbf{t}_k) \geq 0\}]^2 \leq \sqrt{q/n^5} \max_i \|\mathbf{Z}_i\| = Cq \log(n)n^{-5/2},$$

using a similar idea to deriving the upper bound of $\mathbb{E}(U_i^2)$. Applying Bernstein's inequality and the fact that $\max_i |X_{ij}| = O_p(\sqrt{\log(n)})$ for sub-Gaussian random variable, for some positive constant C_1 and C_2 ,

$$\sum_{k=1}^N \Pr(C \max_i |X_{ij}| \sum_{i=1}^n A_i > n\lambda/2) \leq N \exp\left(-\frac{n^2 \lambda^2 / 4}{C_1 q n^{-3/2} \log(n)^{3/2} + C_2 n \lambda}\right) \leq C \exp\{4q \log(n) - Cn\lambda\}.$$

Similarly, we can prove that $\sum_{k=1}^N \Pr(C \max_i |X_{ij}| \sum_{i=1}^n B_i > n\lambda/2) \leq C \exp\{4q \log(n) - Cn\lambda\}$. Therefore, we have

$$J_{nj2} \leq C \exp\{4q \log(n) - Cn\lambda\}. \quad (10)$$

Using (9) and (10), then the probability of Lemma 7.2 is bounded by

$$\sum_{j=q+1}^p (J_{nj1} + J_{nj2}) \leq C \exp\{\log(p) + 4q \log(n) - Cn\lambda\} \rightarrow 0 \quad (11)$$

which completes the proof.

Now we prove Theorem 3.1.

PROOF (PROOF OF THEOREM 3.1). The unpenalized hinge loss objective function is convex. By convex optimization theorem, there exists v_i^* such that $s_j(\hat{\boldsymbol{\beta}}) = 0$, $j = 0, 1, \dots, q$, with $v_i = v_i^*$.

Note that $\min_{1 \leq j \leq q} |\hat{\beta}_j| \geq \min_{1 \leq j \leq q} |\beta_{0j}| - \max_{1 \leq j \leq q} |\hat{\beta}_j - \beta_{0j}|$. By (A7) we have $n^{(1-c_2)/2} \min_{1 \leq j \leq q} |\beta_{0j}| \geq M_1$, and $\max_{1 \leq j \leq q} |\hat{\beta}_j - \beta_{0j}| = O_p(\sqrt{q/n})$ by Lemma 3.1. Thus we have $\min_{1 \leq j \leq q} |\hat{\beta}_j| = O_p(n^{-(1-c_2)/2})$. By $\lambda = o(n^{-(1-c_2)/2})$, we have $\Pr(|\hat{\beta}_j| \geq (a + \frac{1}{2})\lambda) \rightarrow 1$, for $j = 0, 1, \dots, q$.

By the definition of the oracle estimator, we have $|\hat{\beta}_j| = 0$, $j = q+1, \dots, p$. It suffices to show that $\Pr\{|s_j(\hat{\boldsymbol{\beta}})| > \lambda, \text{ for some } j = q+1, \dots, p\} \rightarrow 0$. Let $\mathbf{D} = \{i : 1 - Y_i \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}_1 = 0\}$; then for $j = q+1, \dots, p$, we have

$$s_j(\hat{\boldsymbol{\beta}}) = -n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}_1 \geq 0) - n^{-1} \sum_{i \in \mathbf{D}} W_i Y_i X_{ij} (v_j - 1),$$

where $-1 \leq v_i \leq 0$ if $i \in \mathbf{D}$ and $v_i = 0$ otherwise. By (A5) (\mathbf{Z}_i, Y_i) are in general positions, with probability one there are exactly $(q+1)$ elements in \mathbf{D} . Then by (A4), with probability one $|n^{-1} \sum_{i \in \mathbf{D}} W_i Y_i X_{ij} (v_j - 1)| = O(qn^{-1} \log(q)) = o(\lambda)$. Thus we only need to show that

$\Pr\{\max_{q+1 \leq j \leq p} |n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \widehat{\boldsymbol{\beta}}_1 \geq 0)| > \lambda\} \rightarrow 0$. Observe that

$$\begin{aligned} & \Pr\left\{\max_{q+1 \leq j \leq p} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \widehat{\boldsymbol{\beta}}_1 \geq 0)\right| > \lambda\right\} \\ & \leq \Pr\left\{\max_{q+1 \leq j \leq p} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} [I(1 - Y_i \mathbf{Z}_i^T \widehat{\boldsymbol{\beta}}_1 \geq 0) - I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)]\right| > \frac{\lambda}{2}\right\} \\ & \quad + \Pr\left\{\max_{q+1 \leq j \leq p} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\right| > \frac{\lambda}{2}\right\}. \end{aligned} \quad (12)$$

By Lemma 7.1 the second term of (12) is $o_p(1)$. Notice that from Lemma 3.1, the first term of (12) is bounded by

$$\begin{aligned} & \Pr\left[\max_{q+1 \leq j \leq p} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} \{I(1 - Y_i \mathbf{Z}_i^T \widehat{\boldsymbol{\beta}}_1 \geq 0) - I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\}\right| > \frac{\lambda}{2}\right] \\ & \leq \Pr\left[\max_{q+1 \leq j \leq p} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \Delta \sqrt{q/n}} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} \{I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0) - I(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\}\right| > \frac{\lambda}{4}\right] \\ & \quad - \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0) + \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0) \\ & \quad + \Pr\left[\max_{q+1 \leq j \leq p} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \Delta \sqrt{q/n}} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} \{\Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0)\right.\right. \\ & \quad \left.\left. - \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\right\}\right] > \frac{\lambda}{4}]. \end{aligned} \quad (13)$$

By Lemma 7.2, the first term of (13) is $o_p(1)$. Thus we only need to bound the second term of (13). Notice that

$$\begin{aligned} & |\Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0) - \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)| \\ & \leq |F_i(1 + \mathbf{Z}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})) - F_i(1)| \Pr(Y_i = 1) + |G_i(-1 + \mathbf{Z}_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01})) - G_i(-1)| \Pr(Y_i = -1). \end{aligned}$$

Then we have

$$\begin{aligned} & \max_{q+1 \leq j \leq p} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \Delta \sqrt{q/n}} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} \{\Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0) - \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\}\right| \\ & \leq C \max_{i,j} |X_{ij}| \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \Delta \sqrt{q/n}} n^{-1} \sum_{i=1}^n \|\mathbf{Z}_i\| \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_p(\sqrt{\log pn}) O(\sqrt{q/n}) O_p(\sqrt{q} \log(n)) \\ & = o_p(\lambda). \end{aligned}$$

Thus

$$\begin{aligned} & \Pr\left[\max_{q+1 \leq j \leq p} \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq \Delta \sqrt{q/n}} \left|n^{-1} \sum_{i=1}^n W_i Y_i X_{ij} \{\Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_1 \geq 0)\right.\right. \\ & \quad \left.\left. - \Pr(1 - Y_i \mathbf{Z}_i^T \boldsymbol{\beta}_{01} \geq 0)\right\}\right] > \frac{\lambda}{4} = o_p(1), \end{aligned}$$

which completes the proof.

Now we prove Theorem 3.2.

PROOF (PROOF OF THEOREM 3.2). We will show $\widehat{\beta}$ is a local minimizer of $Q(\beta)$ by writing $Q(\beta)$ as $g(\beta) - h(\beta)$.

By Theorem 3.1, we have $\Pr\{\mathcal{G} \subseteq \partial g(\widehat{\beta})\} \rightarrow 1$, where

$$\bar{\mathcal{G}} = \{\xi = (\xi_0, \dots, \xi_p) : \xi_0 = 0; \xi_j = \lambda \text{sgn}(\widehat{\beta}_j), j = 1, \dots, q; \xi_j = s_j(\beta) + \lambda l_j, j = q+1, \dots, p.\},$$

where $l_j \in [-1, +1]$, $j = q+1, \dots, p$.

Consider any β in the \mathbf{R}^{p+1} with the center $\widehat{\beta}$ and radius $\frac{\lambda}{2}$. It suffices to show that there exist $\xi^* \in \bar{\mathcal{G}}$ such that $\Pr\{\xi_j^* = \frac{\partial h(\beta)}{\partial \beta_j}\} \rightarrow 1$ as $n \rightarrow \infty$.

Since $\frac{\partial h(\beta)}{\partial \beta_0} = 0$, we have $\xi_0^* = \frac{\partial h(\beta)}{\partial \beta_0}$.

For $j = 1, \dots, q$, we have $\min_{1 \leq j \leq q} |\beta_j| \geq \min_{1 \leq j \leq q} |\widehat{\beta}_j| - \max_{1 \leq j \leq q} |\widehat{\beta}_j - \beta_j| \geq (a + \frac{1}{2})\lambda - \frac{\lambda}{2} = a\lambda$ with probability one by Theorem 3.1. Therefore by Condition 2 of the class of penalties $\Pr\{\frac{\partial h(\beta)}{\partial \beta_j} = \lambda \text{sgn}(\beta_j)\} \rightarrow 1$ for $j = 1, \dots, q$. For sufficiently large n , $\text{sgn}(\beta_j) = \text{sgn}(\widehat{\beta}_j)$. Thus we have $\Pr\{\xi_j^* = \frac{\partial h(\beta)}{\partial \beta_j}\} \rightarrow 1$ as $n \rightarrow \infty$ for $j = 1, \dots, q$.

For $j = q+1, \dots, p$, we have $\Pr\{|\beta_j| \leq |\widehat{\beta}_j| + |\beta_j - \widehat{\beta}_j| \leq \lambda\} \rightarrow 1$ by Theorem 3.1. Therefore we have $\Pr\{\frac{\partial h(\beta)}{\partial \beta_j} = 0\} \rightarrow 1$ for SCAD and $\Pr\{\frac{\partial h(\beta)}{\partial \beta_j} = -\frac{\beta_j}{a}\} \rightarrow 1$ for MCP. Observe that by Condition 2 we have $\Pr\{|\frac{\partial h(\beta)}{\partial \beta_j}| \leq \lambda\} \rightarrow 1$ for the class of penalties. By Lemma 1 we have $\Pr\{|s_j(\widehat{\beta}_j)| \leq \lambda\} \rightarrow 1$ for $j = q+1, \dots, p$. We can always find $l_j \in [-1, +1]$ such that $\Pr\{\xi_j^* = s_j(\widehat{\beta}) + \lambda l_j = \frac{\partial h(\beta)}{\partial \beta_j}\} \rightarrow 1$ for $j = 1, \dots, q$, for both penalties. This completes the proof.

The proof of Theorem 3.3 consists of two parts. First we will show that LLA algorithm initiated by $\tilde{\beta}^{(0)}$ gives the oracle estimator after one iteration. Then we will show that once LLA algorithm finds the oracle estimator $\widehat{\beta}$, the LLA algorithm will find it again in the next iteration, that is, the LLA algorithm will converge.

PROOF (PROOF OF THEOREM 3.3). Assume that none of the events F_{ni} is true, for $i = 1, \dots, 4$. The probability that none of these event is true is at least $1 - P_{n1} - P_{n2} - P_{n3} - P_{n4}$. Then we have

$$\begin{aligned} |\tilde{\beta}_j^{(0)}| &= |\tilde{\beta}_j^{(0)} - \beta_{0j}| \leq \lambda, q+1 \leq j \leq p, \\ |\tilde{\beta}_j^{(0)}| &\geq |\beta_{0j}| - |\tilde{\beta}_j^{(0)} - \beta_{0j}| \geq a\lambda, 1 \leq j \leq q. \end{aligned}$$

By Condition 2 of the class of nonconvex penalties, we have $p'_\lambda(|\tilde{\beta}_j^{(0)}|) = 0$ for $1 \leq j \leq q$. Therefore the solution of the next iteration of $\tilde{\beta}^{(1)}$ is the solution to the convex optimization

$$\tilde{\beta}^{(1)} = \arg \min_{\beta} n^{-1} \sum_{i=1}^n W_i (1 - Y_i \mathbf{X}_i^T \beta)_+ + \sum_{q+1 \leq j \leq p} p'_\lambda(|\tilde{\beta}_j^{(0)}|) \cdot |\beta_j|. \quad (14)$$

By the fact the F_{n3} is not true, there exist some subgradients of oracle estimator $s(\widehat{\boldsymbol{\beta}})$ such that $s_j(\widehat{\boldsymbol{\beta}}) = 0$ for $0 \leq j \leq q$ and $|s_j(\widehat{\boldsymbol{\beta}})| < (1 - \frac{1}{a})\lambda$ for $q+1 \leq j \leq p$. Note that by the definition of subgradient, we have

$$\begin{aligned} n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ &\geq n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \widehat{\boldsymbol{\beta}})_+ + \sum_{0 \leq j \leq p} s_j(\widehat{\boldsymbol{\beta}})(\beta_j - \widehat{\beta}_j) \\ &= n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \widehat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leq j \leq p} s_j(\widehat{\boldsymbol{\beta}})(\beta_j - \widehat{\beta}_j). \end{aligned}$$

Then we have for any $\boldsymbol{\beta}$

$$\begin{aligned} &\{n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \sum_{q+1 \leq j \leq p} p'_\lambda(|\tilde{\beta}_j^{(0)}|)|\beta_j|\} - \{n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \widehat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leq j \leq p} p'_\lambda(|\tilde{\beta}_j^{(0)}|)|\widehat{\beta}_j|\} \\ &\geq \sum_{q+1 \leq j \leq p} \{p'_\lambda(|\tilde{\beta}_j^{(0)}|) - s_j(\widehat{\boldsymbol{\beta}}) \cdot \text{sgn}(\beta_j)\} \cdot |\beta_j| \geq \sum_{q+1 \leq j \leq p} \{(1 - \frac{1}{a})\lambda - s_j(\widehat{\boldsymbol{\beta}}) \cdot \text{sgn}(\beta_j)\} \cdot |\beta_j| \geq 0. \end{aligned}$$

The strict inequality holds unless $\beta_j = 0$ for all $q+1 \leq j \leq p$. Since we consider the non-separable case that the oracle estimator is unique, we know the oracle estimator is the unique minimizer of (14) and hence $\tilde{\boldsymbol{\beta}}^{(1)} = \widehat{\boldsymbol{\beta}}$. This proves that the LLA algorithm finds the oracle estimator after one iteration.

In the case that F_{n2} is not true, we have $|\widehat{\beta}_j| > a\lambda$ for all $1 \leq j \leq q$. Hence by Condition 2 of the class of penalties $p'_\lambda(|\widehat{\beta}_j|) = 0$ for all $1 \leq j \leq q$ and $p'_\lambda(|\widehat{\beta}_j|) = p'_\lambda(0) = \lambda$ for all $q+1 \leq j \leq p$. Once the LLA algorithm finds $\widehat{\boldsymbol{\beta}}$, the solution to the next LLA iteration $\tilde{\boldsymbol{\beta}}^{(2)}$ is the minimizer of the convex optimization problem

$$\tilde{\boldsymbol{\beta}}^{(2)} = \arg \min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \sum_{q+1 \leq j \leq p} \lambda |\beta_j|. \quad (15)$$

Then we have for any $\boldsymbol{\beta}$

$$\begin{aligned} &\{n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \sum_{q+1 \leq j \leq p} \lambda |\beta_j|\} - \{n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \widehat{\boldsymbol{\beta}})_+ + \sum_{q+1 \leq j \leq p} \lambda |\widehat{\beta}_j|\} \\ &\geq \sum_{q+1 \leq j \leq p} \{\lambda - s_j(\widehat{\boldsymbol{\beta}}) \cdot \text{sgn}(\beta_j)\} \cdot |\beta_j| \geq 0. \end{aligned}$$

and hence $\tilde{\boldsymbol{\beta}}^{(2)} = \widehat{\boldsymbol{\beta}}$ is the unique minimizer of (15). That is, the LLA algorithm finds the oracle estimator again and stops.

As $n \rightarrow \infty$, by Theorem 3.1 we have $P_{n2} \rightarrow 0$ and $P_{n4} \rightarrow 0$. The proof for $P_{n3} \rightarrow 0$ is similar to the proof for Theorem 3.1 by changing the constant to be $(1 - \frac{1}{a})$.

Now we prove Theorem 3.4.

PROOF (PROOF OF THEOREM 3.4). Let $\|\cdot\|_1$ be the L_1 norm of a vector. Denote $l_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n W_i(1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + c_n \|\boldsymbol{\beta}\|_1$. Note that

$$\begin{aligned} &E[np^{-1}\{l_n(\boldsymbol{\beta}_0 + \sqrt{p/n}\mathbf{u}) - l_n(\boldsymbol{\beta}_0)\}] \\ &= E[np^{-1}\{W(1 - Y\mathbf{X}^T(\boldsymbol{\beta}_0 + \sqrt{p/n}\mathbf{u}))_+ - W(1 - Y\mathbf{X}^T\boldsymbol{\beta}_0)_+\}] \\ &\quad + np^{-1}c_n(\|\boldsymbol{\beta}_0 + \sqrt{p/n}\mathbf{u}\|_1 - \|\boldsymbol{\beta}_0\|_1) \end{aligned}$$

for some constant Δ that $\|\mathbf{u}\| = \Delta$. Observe that $\|\beta_0 + \sqrt{p/n}\mathbf{u}\|_1 - \|\beta_0\|_1 \leq \|\sqrt{p/n}\mathbf{u}\|_1 = \sqrt{p/n}\|\mathbf{u}\|_1$. By the fact that $c_n = o(n^{-1/2})$, we have $np^{-1}c_n(\|\beta_0 + \sqrt{p/n}\mathbf{u}\|_1 - \|\beta_0\|_1) \rightarrow 0$ as $n \rightarrow \infty$. Then similar to the proof of Lemma 3.1, we can show that the expectation is dominated by $\frac{1}{2}u^T H(\beta_0)\mathbf{u} > 0$ and $\Pr\{\inf_{\|\mathbf{u}\|=\Delta} l_n(\beta_0 + \sqrt{p/n}\mathbf{u}) > l_n(\beta_0)\} \geq 1 - \eta$. Hence $\|\hat{\beta}^{L_1} - \beta_0\| = O_p(\sqrt{p/n})$. Because $pn^{-\frac{1}{2}} = o(\lambda)$, $\Pr(|\hat{\beta}_j^{L_1} - \beta_{0j}| > \lambda, \text{ for some } 1 \leq j \leq p) \rightarrow 0$ as $n \rightarrow \infty$. Then using Theorem 3.1 and Corollary 3.1 we have $\Pr\{\hat{\beta}(\lambda) = \hat{\beta}\} \rightarrow 1$, which completes the proof.

Acknowledgments

We thank the co-Editors Professor Gareth Roberts and Professor Piotr Fryzlewicz, the Associate Editor and three referees for very constructive comments and suggestions which have improved the presentation of the paper. We also thank Amanda Applegate for her help on the presentation of this paper. The research is partially supported by National Science Foundation grants DMS-1055210 (Wu) and DMS-1308960 (Wang) and National Institutes of Health grants R01-CA149569 (Zhang and Wu), P01-CA142538 (Wu), P50-DA10075 (Li) and P50-DA036107 (Li).

References

- An, L. T. H. and P. D. Tao (2005). The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research* 133(1-4), 23–46.
- Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473), 138–156.
- Becker, N., G. Toedt, P. Lichter, and A. Benner (2011). Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC Bioinformatics* 12(1), 138.
- Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 37(4), 1705–1732.
- Bradley, P. and O. Mangasarian (1998). Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference (ICML98)*, pp. 82–90.
- Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.
- Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* 106(496), 1566–1577.
- Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* 95(3), 759–771.
- Claeskens, G., C. Croux, and J. Van Kerckhoven (2008). An information criterion for variable selection in support vector machines. *The Journal of Machine Learning Research* 9, 541–558.

- Donoho, D. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–32.
- Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of Statistics* 36(6), 2605–2637.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.
- Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B (Methodological)* 70, 849–911.
- Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, to appear.
- Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer Series in Statistics.
- Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine learning* 46(1), 389–422.
- Kim, Y., H. Choi, and H. Oh (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* 103(484), 1665–1673.
- Kim, Y. and S. Kwon (2012). Global optimality of nonconvex penalized estimators. *Biometrika* 99(2), 315–325.
- Koenker, R. (2005). *Quantile regression*, Volume 38. Cambridge University Press.
- Koo, J., Y. Lee, Y. Kim, and C. Park (2008). A Bahadur representation of the linear support vector machine. *The Journal of Machine Learning Research* 9, 1343–1368.
- Lin, Y. (2000). Some asymptotic properties of the support vector machine. Technical report. Technical report 1029, Department of Statistics, University of Wisconsin-Madison.
- Lin, Y. (2002). Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery* 6, 259–275.
- Lin, Y., Y. Lee, and G. Wahba (2002). Support vector machines for classification in non-standard situations. *Machine Learning* 46(1-3), 191–202.
- Mazumder, R., J. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* 106(495), 1125–1138.
- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3), 1436–1462.
- Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 246–270.
- Park, C., K.-R. Kim, R. Myung, and J.-Y. Koo (2012). Oracle properties of scad-penalized support vector machine. *Journal of Statistical Planning and Inference* 142(8), 2257–2270.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- Tao, P. and L. An (1997). Convex analysis approach to dc programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica* 22(1), 289–355.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), 267–288.
- Vapnik, V. (1996). *The nature of statistical learning theory*. Springer, New York.
- Wang, L., Y. Kim, and R. Li (2013). Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics* 41, 2505–2536.
- Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* 107(497), 214–222.
- Wang, L., J. Zhu, and H. Zou (2006). The doubly regularized support vector machine. *Statistica Sinica* 16(2), 589–615.
- Wang, L., J. Zhu, and H. Zou (2007). Hybrid huberized support vector machines for microarray classification. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 983–990.
- Wegkamp, M. and M. Yuan (2011). Support vector machines with a reject option. *Bernoulli* 17, 1368–1385.
- Welsh, A. (1989). On m -processes and m -estimation. *The Annals of Statistics* 17(1), 337–361.
- Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research* 99, 2261–2286.
- Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2), 894–942.
- Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36(4), 1567–1594.
- Zhang, H., J. Ahn, X. Lin, and C. Park (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* 22(1), 88–95.
- Zhao, P. and B. Yu (2007). On model selection consistency of lasso. *Journal of Machine Learning Research* 7(2), 2541–2563.
- Zhu, J., S. Rosset, T. Hastie, and R. Tibshirani (2004). 1-norm support vector machines. *Advances in Neural Information Processing Systems* 16(1), 49–56.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418–1429.
- Zou, H. (2007). An improved 1-norm svm for simultaneous classification and variable selection. In *Eleventh International Conference on Artificial Intelligence and Statistics*.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 36(4), 1509–1533.

Zou, H. and M. Yuan (2008). The f -infinity norm support vector machine. *Statistica Sinica* 18, 379–398.