Journal of Machine Learning Research 17 (2016) 1-26 Submitted 12/15; Revised 9/16; Published 12/16

An Error Bound for L_1-norm Support Vector Machine Coefficients in Ultra-high Dimension

Bo Peng PENG0199@UMN.EDU
Lan Wang WANGX346@UMN.EDU
School of Statistics
University of Minnesota
Minneapolis, MN 55455, USA

Yichao Wu WU@STAT.NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Editor: Jie Peng

Abstract

Comparing with the standard L_2-norm support vector machine (SVM), the L_1-norm SVM enjoys the nice property of simultaneously preforming classification and feature selection. In this paper, we investigate the statistical performance of L_1-norm SVM in ultra-high dimension, where the number of features p grows at an exponential rate of the sample size n. Different from existing theory for SVM which has been mainly focused on the generalization error rates and empirical risk, we study the asymptotic behavior of the coefficients of L_1-norm SVM. Our analysis reveals that the estimated L_1-norm SVM coefficients achieve near oracle rate, that is, with high probability, the L_2 error bound of the estimated L_1-norm SVM coefficients is of order $O_p(\sqrt{q \log p/n})$, where q is the number of features with nonzero coefficients. Furthermore, we show that if the L_1-norm SVM is used as an initial value for a recently proposed algorithm for solving non-convex penalized SVM (Zhang et al., 2016b), then in two iterative steps it is guaranteed to produce an estimator that possesses the oracle property in ultra-high dimension, which in particular implies that with probability approaching one the zero coefficients are estimated as exactly zero. Simulation studies demonstrate the fine performance of L_1-norm SVM as a sparse classifier and its effectiveness to be utilized to solve non-convex penalized SVM problems in high dimension.

Keywords: feature selection, L_1-norm SVM; non-convex penalty, oracle property, error bound, support vector machine, ulta-high dimension

1. Introduction

Support vector machine (SVM), originally introduced by Boser et al. (1992) and Vapnik (1995) and subsequently investigated by many others, is a popular and highly powerful technique for classification and has a solid mathematical foundation in statistical learning. In modern applications, we often face the challenge of classification at the presence of a very large number of redundant features. For example, in genomics it is of fundamental importance to build a classifier using a small number of genes from thousands of candidate genes for the purpose of disease diagnosis and drug discovery; in spam email classification, it is desirable to build an accurate classifier using a relatively small number of words from a dictionary that contains a huge number of different words. For such applications, the standard L_2-norm SVM suffers from some potential drawbacks. First, L_2-norm SVM does not automatically build in dimension reduction and hence usually does not yield an interpretable sparse decision rule. Second, the generalization performance of L_2-norm SVM can deteriorate by including many redundant features (e.g., Zhu et al., 2004).

The standard L_2-norm SVM has the well known *hinge loss*+L_2 *norm penalty* formulation. An effective way to preform simultaneous variable selection and classification using SVM is to replace the L_2-norm penalty with the L_1-norm penalty, which results in the L_1-norm SVM. See the earlier work of Bradley and Mangasarian (1998) and Song et al. (2002). Important advancement on the methodology and theory of L_1-norm SVM has been obtained in recent years, for example, Zhu et al. (2004) proposed a path-following algorithm and effectively demonstrated the advantages of L_1-norm SVM in high-dimensional sparse scenario; Tarigan and van de Geer (2004) investigated the adaptivity of SVMs with L_1 penalty and derived its adaptive rates; Tarigan, Van De Geer, et al. (2006) obtained an oracle inequality involving both model complexity and margin for L_1-norm SVM; Wang and Shen (2007) extended L_1-norm SVM to multi-class classification problems; Zou (2007) proposed to use adaptive L_1 penalty with the SVM; and Wegkamp and Yuan (2011) considered L_1-norm SVM with a built-in reject option.

The existing theory in the literature on SVM has been largely focused on the analysis of generalization error rate and empirical risk, see Greenshtein et al. (2006), Wang and Shen (2007), Van de Geer (2008), among others. These results neither contain nor directly imply the transparent error bound of the estimated coefficients of L_1-norm SVM studied in this paper. Our work makes a significant departure from most of the existing literature and is motivated by the recent growing interest of understanding the statistical properties of the estimated SVM coefficients (also referred to as the weight vector). For a linear binary SVM, the decision function is a hyperplane that separates two classes. The coefficients of SVM describe this hyperplane which directly predicts which class a new observation point belongs to. Moreover, the magnitudes of the SVM coefficients provide critical information on the importance of the features and can be used for feature ranking (Chang and Lin, 2008; Guyon et al., 2002). Koo et al. (2008) derived a novel Bahadur type representation of the coefficients of the L_2-norm SVM and established the asymptotic normality of the estimated coefficients when the number of features p is fixed. Park et al. (2012) studied the oracle properties of SCAD-penalized SVM coefficients, also for the fixed p case. The aforementioned worked has only considered small, fixed number of features. More recently, Zhang et al. (2016b) proposed a systemic framework for non-convex penalized SVM regarding variable selection consistency and oracle property in high dimension. Zhang et al. (2016a) investigated a consistent information criterion for tuning parameter selection for support vector machine in the diverging model space. Both of these two papers directly assume an appropriate initial value exists in the high-dimensional setting.

In this paper, we study the asymptotic behavior of the estimated L_1-norm SVM coefficients and derive that the error bound is of near-oracle rate $O(\sqrt{q \log p/n})$, where q is the number of features with nonzero coefficients, n is the sample size, and the number of candidate features p can be of exponential order of n (i.e., the ultra-high dimensional case). Furthermore, in Section 4 we show that this sharp error bound helps greatly extend the applicability of the recent algorithm and theory of high-dimensional non-convex-penalized SVM (Zhang et al., 2016b) by providing a statistically valid and computationally convenient initial value. The use of non-convex penalty function aims to further reduce the bias associated with the L_1 penalty and accurately identify the set of relevant features for classification. However, the presence of non-convex penalty results in computational complexity. Zhang et al. (2016b) proposed an algorithm and showed that given an appropriate initial value, in two iterative steps the algorithm is guaranteed to produce an estimator that possesses the oracle property in the ultra-high dimension and consequently with probability approaching one the zero coefficients are estimated as exactly zero. However, the availability of a qualified initial estimator is itself a challenging issue in high dimension. Zhang et al. (2016b) provided an initial estimator that would satisfy the requirement when $p = o(\sqrt{n})$. Our result shows that the L_1-norm SVM can be a valid initial estimator under general conditions when p grows at an exponential rate of n, which completes the algorithm and theory of Zhang et al. (2016b).

The rest of the paper is organized as follows. In Section 2, we introduce the basics and computation of the L_1-norm penalized support vector machine. Section 3 derives the near-oracle error bound for the estimated L_1-norm SVM coefficients in the ultra-high dimension. Section 4 investigates the application of the result in Section 3 for non-convex penalized SVM in the ultra-high dimension. Section 5 demonstrates through Monte Carlo experiments the effectiveness of L_1-norm SVM coefficients both as a sparse classifier and as an initial value for the non-convex penalized SVM algorithm. Technical proofs and additional notes are given in the appendices.

2. L_1-norm support vector machine

We consider the classical binary classification problem. Let $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ be a random sample from an unknown distribution $P(\mathbf{X}, Y)$. The response variable (class label) $Y_i \in \{1, -1\}$ has the marginal distribution: $P(Y_i = 1) = \pi_+$ and $P(Y_i = -1) = \pi_-$, where $\pi_+, \pi_- > 0$ and $\pi_+ + \pi_- = 1$. We write $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^T = (X_{i0}, (\mathbf{X}_{i-})^T)^T$, where $X_{i0} = 1$ corresponds to the intercept term. Let f and g be the conditional density functions of $\mathbf{X}_{i-}$ given $Y_i = 1$ and $Y_i = -1$, respectively. Moreover, in this paper we use the following notation for vector norms: for $\mathbf{x} = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$ and a positive integer m, we define $||\mathbf{x}||_m = \left(\sum_{i=1}^k |x_i|^m\right)^{1/m}$, $||\mathbf{x}||_\infty = \max(|x_1|, \ldots, |x_k|)$ and $||\mathbf{x}||_0 = \sum_{i=1}^k I(x_i \neq 0)$.

The standard linear SVM can be expressed as the following regularization problem

$$\min_{\boldsymbol{\beta}} n^{-1} \sum_{i=1}^n (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \lambda ||\boldsymbol{\beta}_-||_2^2, \tag{1}$$

where $(1-u)_+ = \max\{1-u, 0\}$ is often called the hinge loss function, λ is a tuning parameter and $\boldsymbol{\beta} = (\beta_0, (\boldsymbol{\beta}_-)^T)^T$ with $\boldsymbol{\beta}_- = (\beta_1, \beta_2, \ldots, \beta_p)^T$. Generally for a given vector $\mathbf{e}$, we use $\mathbf{e}_-$ to denote the subvector with the first entry of $\mathbf{e}$ omitted. Actually, optimization problem in (1) is known as the primal problem of the SVM, which can be efficiently solved by quadratic programming algorithms.

The L_1-norm SVM replaces the L_2 penalty in (1) by the L_1 penalty. That is, we consider the objective function

$$l_n(\boldsymbol{\beta}, \lambda) = n^{-1} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+ + \lambda ||\boldsymbol{\beta}_-||_1, \tag{2}$$

and define

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} l_n(\boldsymbol{\beta}, \lambda). \tag{3}$$

For a given data point X_i, it is classified into class $+$ (corresponding to $\hat{Y}_i = 1$) if $\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}(\lambda) > 0$ and into class $-$ (corresponding to $\hat{Y}_i = -1$) if $\mathbf{X}_i^T \widehat{\boldsymbol{\beta}}(\lambda) < 0$.

By introducing the slack variables, we can transform our optimization problem (3) as a linear programming problem (Zhu et al., 2004)

$$\begin{array}{ll} \min\limits_{\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\beta}} & \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} \zeta_j \right) \\ \text{subject to} & \xi_i \geq 0, \qquad i = 1, 2, \ldots, n, \\ & \xi_i \geq 1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta}, \qquad i = 1, 2, \ldots, n, \\ & \zeta_j \geq \beta_j, \zeta_j \geq -\beta_j, \qquad j = 1, 2, \ldots, p. \end{array} \tag{4}$$

Several R packages are available to solve such a standard linear programming problem, such as `lpSolve` and `linprog`.

3. An error bound of L_1-norm SVM in ultra-high dimension

In this section, we will describe the near-oracle error bound for the estimated L1-norm SVM coefficients under the ultra-high dimensional setting. The choice of the tuning parameter λ will be studied to achieve this error bound.

3.1 Preliminaries

The key result of the paper is an error bound of $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$, where $\boldsymbol{\beta}^*$ is the minimizer of the population version of the hinge loss function, that is,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \tag{5}$$

where $L(\boldsymbol{\beta}) = E(1 - Y \mathbf{X}^T \boldsymbol{\beta})_+$. Lin (2002) suggested that there is a close connection between the minimizer of the population hinge loss function and the Bayes rule. The definition of $\boldsymbol{\beta}^*$ above is also used in Koo et al. (2008) and Park et al. (2012), both of which only considered the fixed p case. We are interested in the error bound of $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$ when $p \gg n$. In the ultra-high dimensional settings, it is often reasonable to assume that $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T$ is sparse in the sense that most of its components are exactly zero. We define the index set of active features as $T = \{1 \leq j \leq p : \beta_j^* \neq 0\}$. We denote the cardinality of T by $|T| = q$. To incorporate the intercept term, we also define $T_+ = T \bigcup \{0\}$.

Next, we introduce the gradient vector and Hessian matrix of the population hinge loss function $L(\boldsymbol{\beta})$. We define

$$S(\boldsymbol{\beta}) = -E(I(1 - Y\mathbf{X}^T\boldsymbol{\beta} \geq 0)Y\mathbf{X}) \tag{6}$$

as the $(p+1)$-dimensional gradient vector and

$$H(\boldsymbol{\beta}) = E(\delta(1 - Y\mathbf{X}^T\boldsymbol{\beta})\mathbf{X}\mathbf{X}^T) \tag{7}$$

as the $(p+1)\times(p+1)$-dimensional Hessian matrix where $I(\cdot)$ is the indicator function and $\delta(\cdot)$ is the Dirac delta function. Section 6.1 in Koo et al. (2008) has explained more details and theoretical properties of $S(\boldsymbol{\beta})$ and $H(\boldsymbol{\beta})$ under certain conditions.

Throughout the paper, we assume the following regularity condition.

(A1) The densities f and g are continuous with common support $\mathcal{S} \subset \mathbb{R}^p$ and have finite second moments. In addition, there exists a constant $M > 0$ such that $|X_j| \leq M$, $j \in \{1, \ldots, p\}$.

Remark 1. Condition (A1) ensures that $H(\boldsymbol{\beta})$ is well defined and continuous in $\boldsymbol{\beta}$. The bound of $\mathbf{X}_-$ can be relaxed with further technical complexity. More details can be found in Park et al. (2012) and Koo et al. (2008).

3.2 The choice of the tuning parameter λ and a fact about $\widehat{\boldsymbol{\beta}}$

The estimated L_1-norm SVM parameter $\widehat{\boldsymbol{\beta}}(\lambda)$ defined in (3) depends on the tuning parameter λ. We will first show that a universal choice

$$\lambda = c\sqrt{2A(\alpha)\log p/n}, \tag{8}$$

where c is some given constant, α is a small probability and $A(\alpha) > 0$ is a constant such that $4p^{-\frac{A(\alpha)}{M^2}+1} \leq \alpha$, can provide theoretical guarantee on the good performance of $\widehat{\boldsymbol{\beta}}(\lambda)$.

The above choice of λ is motivated by a principle in the setting of penalized least squares regression (Bickel et al., 2009), which advocates to choose the penalty level λ to dominate the subgradient of the loss function evaluated at the true value. Intuitively, the subgradient evaluated at $\boldsymbol{\beta}^*$ summarizes the estimation noise. See also the application of the same principle to choose the penalty level for quantile regression (Belloni and Chernozhukov, 2011; Wang, 2013). Another more technical motivation of this principle comes from the KKT condition in convex optimization theory. Let $\tilde{\boldsymbol{\beta}}$ be the oracle estimator (formally defined in Section 4) that minimizes the sample hinge loss function when the index set T is known in advance. Define the subgradient function

$$\widehat{S}(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n} I(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta} \geq 0)Y_i\mathbf{X}_i.$$

Then it follows from the argument as in Theorem 3.1 of Zhang et al. (2016b) that under some weak regularity conditions $||\widehat{S}(\tilde{\boldsymbol{\beta}})||_\infty \leq \lambda$ with probability approaching one. It follows from Koo et al. (2008) that the oracle estimator $\tilde{\boldsymbol{\beta}}$ provides a consistent and asymptotically normal estimate of $\boldsymbol{\beta}^*$.

Hence, in the ideal case where the population parameter $\boldsymbol{\beta}^*$ is known, an intuitive choice of λ is to set its value to be larger than the supremum norm of $\widehat{S}(\boldsymbol{\beta}^*)$ with large probability, that is

$$P(\lambda \geq c||\widehat{S}(\beta^*)||_\infty) \geq 1-\alpha, \tag{9}$$

where $c > 1$ is some given constant and α is a small probability. Lemma 1 below shows that the choice of λ given in (8) satisfies this requirement.

Lemma 1 *Assume that condition (A1) is satisfied. Suppose $\lambda = c\sqrt{2A(\alpha)\log p/n}$, we have*

$$P(\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty) \geq 1-\alpha$$

with α being a given small probability defined earlier in this section.

The proof of Lemma 1 is given in the Appendix A. The crux of the proof is to bound the tail probability of $\sum_{i=1}^n I(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* \geq 0)Y_i\mathbf{X}_i$ by applying Hoeffding's inequality and the union bound. Later in this section, we will show that this choice of λ warrants near-oracle rate performance of $\widehat{\boldsymbol{\beta}}(\lambda)$. Let $\mathbf{h} = \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}(\lambda)$. We state below an interesting fact about $\mathbf{h}$.

Lemma 2 *For $\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty$ and $\bar{C} = \frac{c-1}{c+1}$, we have*

$$\mathbf{h} \in \Delta_{\bar{C}},$$

where

$$\Delta_{\bar{C}} = \left\{\boldsymbol{\gamma} \in \mathbf{R}^{p+1} : ||\boldsymbol{\gamma}_{T_+}||_1 \geq \bar{C}||\boldsymbol{\gamma}_{T_+^c}||_1, \textit{where } T_+ = T\cup\{0\},\ T \subset \{1, 2, \ldots, p\} \textit{ and } |T| \leq q\right\},$$

with T_+^c denoting the complement of T_+, and $\boldsymbol{\gamma}_{T_+}$ denoting the $(p+1)$-dimensional vector that has the same coordinates as $\boldsymbol{\gamma}$ on T_+ and zero coordinates on T_+^c.

We call $\Delta_{\bar{C}}$ the *restricted set.* The proof of Lemma 2 is also given in Appendix A.

3.3 Regularity conditions

Let $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T$ denote the feature design matrix. We define *restricted eigenvalues* as follows

$$\lambda_{max} = \max_{\mathbf{d}\in\mathbb{R}^{p+1}:||\mathbf{d}||_0\leq 2(q+1)} \frac{\mathbf{d}^T\mathcal{X}^T\mathcal{X}\mathbf{d}}{n||\mathbf{d}||_2^2} \tag{10}$$

and

$$\lambda_{min}(H(\boldsymbol{\beta}^*); q) = \min_{\mathbf{d}\in\Delta_{\bar{C}}} \frac{\mathbf{d}^T H(\boldsymbol{\beta}^*)\mathbf{d}}{||\mathbf{d}||_2^2}. \tag{11}$$

These are similar to the sparse eigenvalue notion in Bickel, Ritov, and Tsybakov (2009) and Meinshausen and Yu (2009) for analyzing sparse least squares regression, see also Cai, Wang, and Xu (2010).

In addition to condition (A1) introduced in Section 2, we require the following regularity conditions for the main theory of this paper.

(A2) $q = O(n^{c_1})$ for some $0 \leq c_1 < 1/2$.

(A3) There exists a constant M_1 such that $\lambda_{max} \leq M_1$ almost surely.

(A4) $\lambda_{min}(H(\boldsymbol{\beta}^*); q) \geq M_2$, for some constant $M_2 > 0$.

(A5) $n^{(1-c_2)/2} \min_{j \in T} |\beta_j^*| \geq M_3$ for some constants $M_3 > 0$ and $2c_1 < c_2 \leq 1$.

(A6) Denote the conditional density of $\mathbf{X}^T\boldsymbol{\beta}^*$ 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 2. Conditions (A2) and (A5) are very common in high dimensional literature. Basically, condition (A2) states that the number of nonzero variables cannot diverge at a rate larger than $\sqrt{n}$. Condition (A5) controls the decay rate of true parameter $\boldsymbol{\beta}^*$. Condition (A3) is not restrictive, see the relevant discussions in Meinshausen and Yu (2009). Condition (A4) requires the smallest restricted eigenvalue has a lower bound. This would be satisfied if $H(\boldsymbol{\beta}^*)$ is positive definite. We provide a thorough discussion of this condition in Appendix B, including an example that demonstrates the validity of this condition. Condition (A6) warrants that there is sufficient information around the non-differentiable point of the hinge loss, similarly to Condition (C3) in Wang, Wu, and Li (2012) for quantile regression.

3.4 An error bound of $\widehat{\boldsymbol{\beta}}(\lambda)$ in ultra-high dimension

Before stating the main theorem, we first present an important lemma, which has to do with the empirical process behavior of the hinge loss function.

Lemma 3 *Assume that conditions (A1)-(A3) are satisfied. For* $\mathbf{h} \in \mathbb{R}^{p+1}$*, let*

$$\begin{aligned} B(\mathbf{h}) \;=\; & \frac{1}{n}\Big| \sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \\ & - E\Big(\sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+\Big)\Big|. \end{aligned}$$

Assume $p > n$*, then for all* n *sufficiently large*

$$P\left(\sup_{||\mathbf{h}||_0 \leq q+1, ||\mathbf{h}||_2 \neq 0} \frac{B(\mathbf{h})}{||\mathbf{h}||_2} \geq (1 + 2C_1\sqrt{M_1})\sqrt{\frac{2q\log p}{n}} \right) \leq 2p^{-2q(C_1^2-1)},$$

where $C_1 > 1$ *is a constant.*

Lemma 3 guarantees that $n^{-1}\left(\sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \sum_{i=1}^n (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+\right)$ is close to its expected value with high probability. This provides an important tool to handle the non-smoothness of the hinge loss function in proving the main theory, which is stated below.

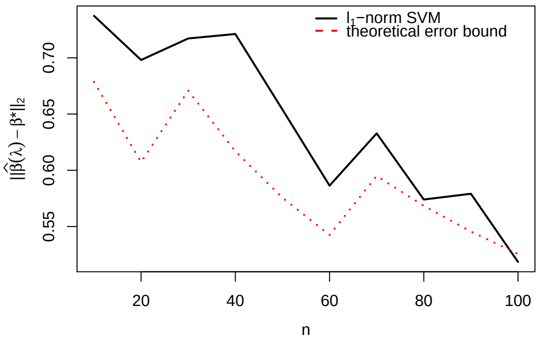

Figure 1: *L_2-norm estimation error comparison*

Theorem 4 *Suppose that conditions (A1)-(A6) hold, then the estimated L_1-norm SVM coefficients vector $\widehat{\boldsymbol{\beta}}(\lambda)$ satisfies*

$$||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2 \leq \sqrt{1 + \frac{1}{\bar{C}}}\Big(\frac{2\lambda\sqrt{q+1}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}\big(\frac{5}{4} + \frac{1}{\bar{C}}\big)\Big)$$

with probability at least $1 - 2p^{-2q(C_1^2-1)}$, where C is a constant, C_1 is given in Lemma 3 and $\bar{C}$ is defined in Lemma 2.

From this theorem, we can easily capture the near-oracle property for l_1 penalized SVM estimator, such that with high probability,

$$||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2 = O_p\left(\sqrt{\frac{q\log p}{n}}\right)$$

when $\lambda = c\sqrt{2A(\alpha)\log p/n}$. Actually, in the inequality of Theorem 4, the first term satisfies $\frac{\lambda\sqrt{q}}{M_2} = \frac{2}{M_2}\sqrt{\frac{2A(\alpha)q\log p}{n}} = O\left(\sqrt{\frac{q\log p}{n}}\right)$ and it is also trivial to have the second term of the same order. Hence the near-oracle property of $\widehat{\beta}(\lambda)$ will hold given λ above.

To numerically evaluate the above error bound of the L_1-norm SVM, we consider the simulation setting in Model 4 of Section 5.1. We choose $p = 0.1 * n^2$, $q = \lfloor n^{1/3} \rfloor$ and $\boldsymbol{\beta}^*_- = ((1.1, \ldots, 1.1)_q, 0, \ldots, 0)^T$, which allows p and q to vary with sample size n. Figure 1 depicts the average of $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$ across 200 simulation runs for different values of n for L_1-norm SVM and compares the curve with the theoretical error bound $(\sqrt{\frac{q\log p}{n}})$. We observe that these two curves display similar decreasing pattern and approach each other as n gets larger.

4. Application to non-convex penalized SVM in ultra-high dimension

In this section, we will step further to discuss the advantage of non-convex penalized SVM in ultra-high dimension. Similarly, the oracle property of non-convex penalized SVM coefficients will be investigated.

4.1 Why non-convex penalty?

Recently, several authors studied non-convex penalized SVM for simultaneous variable selection and classification, see Zhang et al. (2006), Becker et al. (2011), Park et al. (2012) and Zhang et al. (2016b). The idea is to replace the L_2 norm in standard SVM (1) by a non-convex penalty term in the form $\sum_{j=1}^{p} p_\lambda(|\beta_j|)$, where $p_\lambda(\cdot)$ is a symmetric penalty function with tuning parameter λ. Two commonly used non-convex penalty functions are the SCAD penalty and the MCP penalty. The SCAD penalty (Fan and Li, 2001) is defined by

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

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

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

for some $a > 1$.

The motivation of using non-convex penalty function is to further reduce the bias resulted from L_1 penalty and accurately identify the set of relevant features T. The use of non-convex penalty function was introduced in the setting of penalized least squares regression (Fan and Li, 2001; Zhang, 2010). These authors observed that L_1 penalized least squares regression requires stringent conditions, often not satisfied in real data analysis, to achieve variable selection consistency. The use of non-convex penalty function alleviates the bias caused by L_1 penalty which overpenalizes large coefficients, and leads to the so called *oracle property*. That is, under regularity conditions the resulted non-convex penalized estimator is able to estimate zero coefficients as exactly zero with probability approaching one, and estimate the nonzero coefficients as efficiently as if the set of relevant features is known in advance.

4.2 Oracle property in ultra-high dimension

The oracle property of non-convex penalized SVM coefficients is investigated by Park et al. (2012) for the case of fixed number of features and more recently by Zhang et al. (2016b) for the large p case. The oracle estimator of $\boldsymbol{\beta}^*$ is defined as

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{T_+^c}=\mathbf{0}} \tilde{l}_n(\boldsymbol{\beta}), \tag{12}$$

where $\tilde{l}_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} (1 - Y_i \mathbf{X}_i^T \boldsymbol{\beta})_+$ is the sample hinge loss function and $\boldsymbol{\beta}_{T_+^c}$ denotes the vector containing the components of $\boldsymbol{\beta}$ with indices in T_+^c and others to be zero.

To solve the non-convex penalized SVM, we choose to use the local linear approximation (LLA) algorithm. The LLA algorithm starts with an initial value $\boldsymbol{\beta}^{(0)}$. At each step t, we update the $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}^{(t)}$ by solving

$$\min_{\beta} \{n^{-1}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta})_+ + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(t-1)}|)|\beta_j|\}, \tag{13}$$

where $p'_\lambda(\cdot)$ denotes the derivative of the penalty function $p_\lambda(\cdot)$. Specifically, we have $p'_\lambda(0) = p'_\lambda(0+) = \lambda$.

Zhang et al. (2016b) showed that if an appropriate initial estimator exists, then under quite general regularity conditions, the LLA algorithm can identify the oracle estimator with probability approaching one in just two iterative steps (see their Theorem 3.4). This result provides a systematic framework for non-convex penalized SVM in high dimension. However it relies on the availability of a qualified initial value $\widehat{\boldsymbol{\beta}}^{(0)} = (\widehat{\beta}_0^{(0)}, \widehat{\beta}_1^{(0)}, \ldots, \widehat{\beta}_p^{(0)})^T$that satisfies

$$P(|\widehat{\beta}_j^{(0)} - \beta_j^*| > \lambda, \text{ for some } 1 \le j \le p) \to 0 \text{ as } n \to \infty. \tag{14}$$

Yet the availability of such an appropriate initial value is itself a challenging problem in ultra-high dimension. Zhang et al. (2016b) showed that such an initial estimator is guaranteed when $p = o(\sqrt{n})$. The error bound we derived on L_1-norm SVM ensures that a qualified initial value is indeed available under general conditions in ultra-high dimension and hence greatly extends the applicability of the result of Zhang et al. (2016b). In the following we restate Theorem 3.4 of Zhang et al. (2016b) for the ultra-high dimensional case.

Theorem 5 *Assume $\widehat{\beta}(\lambda)$ is the solution to the L_1-norm SVM with tuning parameter $\lambda = c\sqrt{2A(\alpha)\log p/n}$ defined above. Suppose that conditions (A1)-(A6) hold, then we have $P(|\widehat{\beta}_j(\lambda) - \beta_j^*| > \lambda,$ for some $1 \le j \le p) \to 0$ as $n \to \infty$. Furthermore, the LLA algorithm initiated by $\widehat{\beta}(\lambda)$ finds the oracle estimator in two iterations with probability tending to 1, i.e., $P(\widehat{\beta}^{nc}(\lambda) = \tilde{\beta})$, where $\widehat{\beta}^{nc}(\lambda)$ is the solution for non-convex penalized SVM with given λ.*

5. Simulation experiments

In this section, we will investigate the finite sample performance of the L_1-norm SVM. We will also study its application to non-convex penalized SVM in high dimension.

5.1 Monte Carlo results for L_1-norm SVM

We generate random data from each of the following four models.

- Model 1: $Pr(Y=1) = Pr(Y=-1) = 0.5$, $\mathbf{X}_-|(Y=1) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}_-|(Y=-1) \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q=5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T \in \mathbb{R}^p, \boldsymbol{\Sigma} = (\sigma_{ij})$ with diagonal entries equal to 1, nonzero entries $\sigma_{ij} = -0.2$ for $1 \le i \ne j \le q$ and other entries equal to 0. The Bayes rule is $\text{sign}(1.39X_1 + 1.47X_2 + 1.56X_3 + 1.65X_4 + 1.74X_5)$ with Bayes error 6.3%.

- Model 2: $Pr(Y=1) = Pr(Y=-1) = 0.5$, $\mathbf{X}_{-}|(Y=1) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}_{-}|(Y=-1) \sim MN(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $q=5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T \in \mathbb{R}^p, \boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = -0.4^{|i-j|}$ for $1 \leq i, j \leq q$ and other entries equal to 0. The Bayes rule is $\text{sign}(3.09X_1 + 4.45X_2 + 5.06X_3 + 4.77X_4 + 3.58X_5)$ with Bayes error 0.6%.

- Model 3: model stays the same as Model 2, but $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ij} = -0.4^{|i-j|}$ for $1 \leq i, j \leq q$ and $\sigma_{ij} = 0.4^{|i-j|}$ for $q < i, j \leq p$. The Bayes rule is still $sign(3.09X_1 + 4.45X_2 + 5.06X_3 + 4.77X_4 + 3.58X_5)$ with Bayes error 0.6%.

- Model 4: $\mathbf{X}_{-} \sim MN(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = (\sigma_{ij})$ with nonzero elements $\sigma_{ij} = 0.4^{|i-j|}$ for $1 \leq i, j \leq p$, $Pr(Y=1|\mathbf{X}_{-}) = \Phi(\mathbf{X}_{-}^T \boldsymbol{\beta}_{-}^*)$, where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, $\boldsymbol{\beta}_{-}^* = (1.1, 1.1, 1.1, 1.1, 0, \ldots, 0)^T$ and $q=4$. The Bayes rule is $\text{sign}(1.1X_1 + 1.1X_2 + 1.1X_3 + 1.1X_4)$ with Bayes error 10.4%.

Model 1 and Model 4 are identical to the ones in Zhang et al. (2016b). In particular, Model 1 focuses on a standard linear discriminate analysis setting. On the other hand, Model 4 is a typical probit regression case. Models 2 and 3 are designed with autoregressive covariance as correlation decaying off-diagonal-wise. We consider sample size $n = 100$ with $p = 1000$ and 1500, and $n = 200$ with $p = 1500$ and 2000. Similarly as in Cai, Liu, and Luo (2011), we use an independent tuning data set of size $2n$ to tune our λ by minimizing the prediction error using five-fold cross validation. The tuning range spans from 2^{-6} to 2 as equally-spaced sequence with 100 elements. For each simulation scenario, we conduct 200 runs. Then we generate an independent test data set of size n to report the estimated test error.

We evaluate the performance of L_1-norm SVM by its testing misclassification error rate, estimator error and variable selection ability. In particular, we measure the estimation accuracy by two criteria: the L_2 estimation error $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$ where Appendix B provides details on the calculation of $\boldsymbol{\beta}^*$ and the absolute value of the sample correlation between $\mathbf{X}^T\widehat{\boldsymbol{\beta}}(\lambda)$ and $\mathbf{X}^T\boldsymbol{\beta}^*$. The absolute value of the sample correlation (AAC) is also used as accuracy measure in Cook et al. (2007). To summarize, we will report

- **Test error**: the misclassification error rate.

- L_2 **error**: $||\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^*||_2$.

- **AAC**: Absolute absolute correlation $\text{corr}(\mathbf{X}^T\widehat{\boldsymbol{\beta}}(\lambda), \mathbf{X}^T\boldsymbol{\beta}^*)$.

- **Signal**: the average of number of nonzero regression coefficients $\widehat{\boldsymbol{\beta}}_i \neq 0$ with $i = 1, 2, 3, 4, 5$ for Model 1-3 and with $i = 1, 2, 3, 4$ for Model 4. This measures the ability of L_1-norm SVM selecting relevant features.

- **Noise**: the average of number of nonzero regression coefficients $\widehat{\boldsymbol{\beta}}_i(\lambda) \neq 0$ with $i \notin \{1, 2, 3, 4, 5\}$ for Model 1-3 and with $i \notin \{1, 2, 3, 4\}$ for Model 4. This measures the ability of L_1-norm SVM not selecting noise features.

Table 1 summarizes the simulation results for all four models. The numbers in the parentheses are the corresponding standard errors based on 200 replications. Overall, the L_1-norm SVM performs satisfactorily for classification with relatively low error rates in all

Table 1: Simulation results for L_1-norm SVMs

Model	n	p	Test error	L_2 error	AAC	Signal	Noise
Model 1	100	1000	0.17(0.06)	0.53(0.14)	0.89(0.03)	4.84(0.41)	38.20(5.50)
	100	1500	0.19(0.05)	0.59(0.14)	0.89(0.03)	4.75(0.47)	40.27(5.41)
	200	1500	0.10(0.03)	0.27(0.07)	0.96(0.02)	5.00(0.07)	19.80(4.12)
	200	2000	0.10(0.02)	0.27(0.06)	0.96(0.02)	5.00(0.00)	23.61(4.80)
Model 2	100	1000	0.06(0.04)	0.34(0.12)	0.95(0.02)	4.88(0.35)	21.25(4.22)
	100	1500	0.07(0.04)	0.39(0.12)	0.95(0.02)	4.79(0.41)	28.80(4.61)
	200	1500	0.02(0.01)	0.21(0.07)	0.97(0.01)	4.99(0.10)	5.41(2.25)
	200	2000	0.02(0.02)	0.22(0.07)	0.97(0.01)	4.99(0.10)	6.88(2.50)
Model 3	100	1000	0.06(0.05)	0.36(0.14)	0.95(0.02)	4.8.(0.40)	19.93(3.87)
	100	1500	0.06(0.04)	0.37(0.13)	0.95(0.02)	4.83(0.40)	27.55(4.85)
	200	1500	0.02(0.02)	0.22(0.07)	0.97(0.02)	5.00(0.07)	5.18(2.19)
	200	2000	0.02(0.02)	0.20(0.08)	0.97(0.02)	5.00(0.07)	6.72(2.67)
Model 4	100	1000	0.16(0.04)	0.52(0.13)	0.94(0.03)	3.88(0.33)	12.87(3.65)
	100	1500	0.17(0.05)	0.55(0.14)	0.93(0.03)	3.81(0.42)	12.09(3.56)
	200	1500	0.13(0.03)	0.33(0.09)	0.97(0.01)	4.00(0.00)	11.12(3.53)
	200	2000	0.15(0.03)	0.43(0.07)	0.94(0.02)	4.00(0.00)	48.34(7.71)

the models. Actually, the error rates are all quite close to the Bayes errors. It is also successful in eliminating most of the irrelevant features. The performance improves with increased sample size. In terms of estimation accuracy, the L_2 error decreases as p decreases and n increases, which echoes the result in main theorem. We observe that AAC is greater than 0.9 in most cases, implying that the direction of $\widehat{\boldsymbol{\beta}}(\lambda)$ matches that of the Bayes rule.

It is worth noting that the earlier literature have already performed thorough numerical analysis to compare the performance of L_1-norm SVM with L_2-norm SVM and logistic regression. For example, Zhu et al. (2004) observes that the performance of L_1-norm SVM and L_2-norm SVM is similar when there is no redundant features; however, the performance of L_2-norm SVM can be adversely affected by the presence of redundant features. Rocha et al. (2009) numerically compared L_1-norm SVM with logistic regression classifier and discovered that they are comparable but their relative finite-sample advantage depends on the sample size and design. See similar observation in Zou (2007), Zhang et al. (2016b), among others. Although L_1-norm SVM can outperform regular L_2-norm SVM when there are many redundant features, it shares the drawback of L_1 penalized least squares regression that it overpenalizes large coefficients and tends to have larger false positives (including more noise features) comparing with the non-convex penalized SVM, which will be investigated in Section 5.2.

5.2 Monte Carlo results for non-convex penalized SVM

In this subsection, we consider the same four models as in Section 5.1. Instead of the L_1-norm SVM, we use it as the initial value for the non-convex penalized SVM algorithm proposed in Zhang et al. (2016b). We consider two popular choices of non-convex penalty functions: SCAD penalty (with $a = 3.7$) and MCP penalty (with $a = 3$). As suggested in Zhang et al. (2016b), we used the recently developed high-dimensional BIC criterion to choose the tuning parameter for non-convex penalized SVMs. More specifically, the SVM-extended BIC is defined as

$$SVMIC_\gamma(T) = \sum_{i=1}^{n} 2\xi_i + \log(n)|T| + 2\gamma \binom{p}{|T|}, \qquad 0 \leq \gamma \leq 1,$$

where in practice we can set $\gamma = 0.5$ as suggested by Chen and Chen (2008) and choose the λ that minimizes the above $SVMIC_\gamma$ for non-convex penalized SVM.

Table 2: Simulation results for SCAD penalized SVM

Model	n	p	Test error	L_2 error	AAC	Signal	Noise
Model 1	100	1000	0.10(0.05)	0.25(0.17)	0.95(0.04)	4.88(0.38)	4.92(5.82)
	100	1500	0.12(0.06)	0.35(0.20)	0.93(0.05)	4.84(0.53)	9.31(8.89)
	200	1500	0.08(0.03)	0.15(0.10)	0.98(0.03)	4.99(0.12)	0.48(0.51)
	200	2000	0.07(0.02)	0.10(0.05)	0.99(0.01)	5.00(0.00)	0.66(0.80)
Model 2	100	1000	0.04(0.05)	0.25(0.17)	0.95(0.05)	4.73(0.51)	1.47(1.38)
	100	1500	0.05(0.05)	0.28(0.18)	0.94(0.05)	4.64(0.55)	1.42(1.38)
	200	1500	0.03(0.03)	0.19(0.10)	0.96(0.03)	4.91(0.29)	2.77(3.53)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.07)	1.40(1.81)
Model 3	100	1000	0.05(0.04)	0.30(0.16)	0.94(0.04)	4.53(0.58)	0.58(0.84)
	100	1500	0.04(0.04)	0.24(0.15)	0.95(0.04)	4.75(0.46)	1.08(1.15)
	200	1500	0.02(0.01)	0.14(0.06)	0.98(0.01)	4.99(0.10)	1.30(1.53)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.00)	1.32(1.83)
Model 4	100	1000	0.15(0.05)	0.51(0.20)	0.94(0.04)	3.50(0.59)	7.54(5.20)
	100	1500	0.17(0.05)	0.61(0.18)	0.93(0.04)	3.57(0.71)	8.86(6.37)
	200	1500	0.12(0.03)	0.19(0.10)	0.99(0.01)	3.98(0.14)	3.19(2.45)
	200	2000	0.14(0.03)	0.39(0.19)	0.97(0.03)	3.69(0.51)	0.95(1.07)

Tables 2 and 3 summarize the simulation results for SCAD and MCP penalty functions, respectively. We observe that the SCAD-penalized SVM and MCP-penalized MCP have similar performance, both demonstrating a clear advantage of selecting the relevant features and excluding irrelevant ones over L_1-norm SVM. The Noise size decreases dramatically to less than 3 as the sample size gets larger. The Signal size is almost 5 when $n = 200$ for Model 1-3 and 4 for Model 4, implying the success of selecting the exact true model. We also observe that non-convex penalized SVM has uniformly smaller L_2 error and larger AAC than L_1-norm SVM. This resonates with the observation in the literature that eliminating irrelevant features enhances classification performance. The Monte Carlo study confirms the effectiveness of the algorithm of Zhang et al. (2016b) for feature selection for SVM in high dimension when using L_1-norm SVM as an initial value.

Table 3: Simulation results for MCP penalized SVM

Model	n	p	Test error	L_2 error	AAC	Signal	Noise
Model 1	100	1000	0.11(0.05)	0.28(0.17)	0.95(0.04)	4.87(0.42)	5.46(5.45)
	100	1500	0.13(0.07)	0.36(0.20)	0.93(0.05)	4.84(0.47)	9.00(8.49)
	200	1500	0.07(0.02)	0.11(0.07)	0.99(0.02)	4.99(0.10)	0.48(0.51)
	200	2000	0.07(0.02)	0.10(0.04)	0.99(0.01)	5.00(0.00)	0.83(0.83)
Model 2	100	1000	0.03(0.03)	0.20(0.12)	0.96(0.03)	4.84(0.38)	0.88(0.97)
	100	1500	0.11(0.10)	0.47(0.27)	0.89(0.08)	4.08(0.85)	3.56(2.65)
	200	1500	0.02(0.01)	0.14(0.05)	0.98(0.01)	5.00(0.00)	1.50(2.22)
	200	2000	0.02(0.01)	0.14(0.06)	0.98(0.02)	5.00(0.07)	1.38(1.80)
Model 3	100	1000	0.04(0.04)	0.26(0.15)	0.95(0.04)	4.67(0.54)	0.60(0.82)
	100	1500	0.04(0.04)	0.24(0.15)	0.95(0.04)	4.75(0.46)	1.01(1.07)
	200	1500	0.02(0.01)	0.14(0.06)	0.98(0.01)	5.00(0.07)	1.27(1.72)
	200	2000	0.02(0.01)	0.15(0.06)	0.98(0.02)	5.00(0.00)	1.47(2.04)
Model 4	100	1000	0.15(0.05)	0.50(0.20)	0.94(0.04)	3.66(0.52)	7.20(4.49)
	100	1500	0.17(0.05)	0.62(0.16)	0.92(0.04)	3.35(0.68)	4.96(3.58)
	200	1500	0.12(0.03)	0.20(0.12)	0.99(0.01)	3.98(0.12)	1.99(1.72)
	200	2000	0.13(0.03)	0.34(0.17)	0.97(0.02)	3.83(0.43)	0.86(0.80)

6. Conclusion and discussion

We investigate the statistical properties of L_1-norm SVM coefficients in ultra-high dimension. We proved that L_1-norm SVM coefficients achieve a near-oracle rate of estimation error. To deal with the non-smoothness of the hinge loss function, we employ empirical processes techniques to derive the theory. Furthermore, we showed that under some general regularity conditions, the L_1-norm SVM provides an appropriate initial value for the recent algorithm developed by Zhang et al. (2016b) for non-convex penalized SVM in high dimension. Combined with the theory in that paper, we extended the applicability and validity of their result to the ultra-high dimension.

Our work is motivated by the importance of identifying individual features for SVM in analyzing high-dimensional data, which frequently arise in genomics and many other fields. We not only closed a theoretical gap on the estimation error bound on L_1-SVM when $p \gg n$, but also verified that (Section 4) this leads to consistently identifying important features when combined with a two-step iterative algorithm in the ultra-high dimensional setting. Hence, we have guarantee for both algorithm convergence and theoretical performance. We believe such results are of direct interest to JMLR readers given the popularity of SVM in practice. Our work has substantial difference from the existing work in the literature. The existing theory on SVM has been largely focused on the analysis of generalization error rate and empirical risk. These results neither contain nor directly imply the transparent error bound of the estimated coefficients of L_1-norm SVM studied in this paper. Furthermore, the techniques used in the paper for deriving the L_2 error bound when $p \gg n$ are completely different from those used in $p < n$ setting. Although our approach for deriving the L_2-error bound is inspired by the recent work in the literature for Lasso. There is substantial new technical challenge to deal with the nonsmooth Hinge loss function and requires more delicate application of empirical process techniques. Also, unlike Lasso, we do not require Gaussian or sub-Gaussian conditions in the technical derivation.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant DMS-1308960) for Peng and Wang's research and the National Science Foundation (NSF grant DMS-1055210) and National Institutes of Health (NIH grants R01 CA149569 and P01 CA142538) for Wu's research respectively.

Appendix A: Technical Proofs

Proof of Lemma 1. By the union bound, we have

$$\begin{aligned} & P\big(c\sqrt{2A(\alpha)\log p/n} \le c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty\big) \\ \le\ & \sum_{j=0}^{p} P\Big(\sqrt{2A(\alpha)\log p/n} \le n^{-1}\big|\sum_{i=1}^{n} I(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* \ge 0)Y_iX_{ij}\big|\Big). \end{aligned}$$

Notice that we have $S(\boldsymbol{\beta}^*) = 0$ because of minimizer $\boldsymbol{\beta}^*$ and the definition of gradient vector. Then, for each i and j, $E(Y_iX_{ij}I(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* \ge 0)) = 0$, by Hoeffding's inequality,

$$\begin{aligned} & P\Big(\sqrt{2A(\alpha)\log p/n} \le n^{-1}\big|\sum_{i=1}^{n} I(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* \ge 0)Y_iX_{ij}\big|\Big) \\ \le\ & 2\exp\Big(-\frac{4A(\alpha)n\log p}{4nM^2}\Big) = 2p^{-\frac{A(\alpha)}{M^2}}. \end{aligned}$$

Terefore $P\big(c\sqrt{2A(\alpha)\log p/n} \le c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty\big) \le (p+1)\cdot 2p^{-\frac{A(\alpha)}{M^2}} \le \alpha$.

Proof of Lemma 2. Since $\widehat{\beta}$ minimizes $l_n(\boldsymbol{\beta})$, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\widehat{\boldsymbol{\beta}})_+ + \lambda||\widehat{\boldsymbol{\beta}}_-||_1 \le \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ + \lambda||\boldsymbol{\beta}^*_-||_1, \\ & \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \le \lambda||\boldsymbol{\beta}^*_-||_1 - \lambda||\widehat{\boldsymbol{\beta}}_-||_1. \end{aligned}$$

Recalling $T = \{1 \le j \le p : \beta^*_j \ne 0\}$ and $T_+ = T\bigcup\{0\}$, we have

$$\begin{aligned} ||\boldsymbol{\beta}^*_-||_1 - ||\widehat{\boldsymbol{\beta}}_-||_1 &\le ||\boldsymbol{\beta}^*_{T_+}||_1 - ||\widehat{\boldsymbol{\beta}}_-||_1 \\ &\le ||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1. \end{aligned}$$

This implies

$$\frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1-Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \le \lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1). \qquad (15)$$

Since the subdifferential of $l_n(\boldsymbol{\beta})$ at the point of $\boldsymbol{\beta}^*$ is $\widehat{S}(\boldsymbol{\beta}^*)$ and recall the assumption $\lambda \geq c||\widehat{S}(\boldsymbol{\beta}^*)||_\infty$, we have

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \\
\geq\ & \widehat{S}^T(\boldsymbol{\beta}^*)\mathbf{h} \\
\geq\ & -||\mathbf{h}||_1 \cdot ||\widehat{S}(\boldsymbol{\beta}^*)||_\infty \\
\geq\ & -\frac{\lambda}{c}(||\mathbf{h}_{T_+}||_1 + ||\mathbf{h}_{T_+^c}||_1).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^c}||_1) &\geq -\frac{\lambda}{c}(||\mathbf{h}_{T_+}||_1 + ||\mathbf{h}_{T_+^c}||_1), \\
||\mathbf{h}_{T_+}||_1 &\geq \bar{C}||\mathbf{h}_{T_+^c}||_1,
\end{aligned}$$

where $\bar{C} = \frac{c-1}{c+1}$. We have thus proved that $\mathbf{h} \in \Delta_{\bar{C}}$.

Proof of Lemma 3. We first consider a fixed $\mathbf{h} \in \mathbb{R}^{p+1}$ such that $||\mathbf{h}||_0 \leq q+1$ and $||\mathbf{h}||_2 \neq 0$. Note that the Hinge loss function is Lipschitz continuous and we have

$$\frac{\big|(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - (1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+\big|}{||\mathbf{h}||_2} \leq \frac{|\mathbf{X}_i^T\mathbf{h}|}{||\mathbf{h}||_2}.$$

By Hoeffding's inequality, we have $\forall\ t > 0$,

$$P\left(\frac{B(\mathbf{h})}{||\mathbf{h}||_2} \geq \frac{t}{\sqrt{n}}|\mathcal{X}\right) \leq 2\exp\Big(-\frac{2nt^2}{4||\mathcal{X}\mathbf{h}||_2^2/||\mathbf{h}||_2^2}\Big).$$

Hence by assumption (A3),

$$P\left(\frac{B(\mathbf{h})}{||\mathbf{h}||} \geq \frac{t}{\sqrt{n}}|\mathcal{X}\right) \leq 2\exp\Big(-\frac{t^2}{2\lambda_{max}}\Big) \leq 2\exp\Big(-\frac{t^2}{2M_1}\Big).$$

Let $t = C\sqrt{2q\log p}$, where C is an arbitrary given positive constant. Then

$$P\Big(\frac{B(\mathbf{h})}{||\mathbf{h}||} \geq C\sqrt{\frac{2q\log p}{n}}\Big) \quad\leq\quad 2\exp\Big(-\frac{C^2 q\log p}{M_1}\Big) \leq 2p^{-C^2q/M_1} \leq 2p^{-C^2(q+1)/(2M_1)}.$$

Next we will derive an upper bound for $\sup\limits_{||\mathbf{h}||_0 \leq q+1, ||\mathbf{h}||_2 \neq 0} \frac{B(\mathbf{h})}{||\mathbf{h}||}$. We consider covering $\{\mathbf{h} \in \mathbb{R}^{p+1}, ||\mathbf{h}||_0 \leq q+1\}$ with ϵ-balls such that for any $\mathbf{h}_1$ and $\mathbf{h}_2$ in the same ball we have $\big|\frac{\mathbf{h}_1}{||\mathbf{h}_1||_2} - \frac{\mathbf{h}_2}{||\mathbf{h}_2||_2}\big| \leq \epsilon$, where ϵ is a small positive number. The number of ϵ-balls that is required to cover a k-dimensional unit ball is bounded by $(3/\epsilon)^k$, see for example Rogers (1963) and Bourgain and Milman (1987). Since $\mathbf{h}$ is a $(p+1)$-dimensional vector with at most $q+1$ nonzero coordinates and $\mathbf{h}/||\mathbf{h}||_2$ has unit length in L_2 norm, the covering number we require is at most $(3p/\epsilon)^{q+1}$. Let N denote such an ϵ-net. By the union bound,

$$P\left(\sup_{\mathbf{h}\in N}\frac{B(\mathbf{h})}{||\mathbf{h}||_2} \geq C\sqrt{\frac{2q\log p}{n}}\right) \leq 2\left(\frac{3p}{\epsilon}\right)^{q+1} p^{-C^2(q+1)/(2M_1)} = 2\Big(\frac{3}{\epsilon}p^{1-C^2/(2M_1)}\Big)^{q+1},$$

for any given positive constant C. Furthermore, for any $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^{p+1}$ such that $||\mathbf{h}_1||_0 \leq q+1$, $||\mathbf{h}_2||_0 \leq q+1$, $||\mathbf{h}_1||_2 \neq 0$ and $||\mathbf{h}_2||_2 \neq 0$, we have

$$\begin{aligned} \Big|\frac{B(\mathbf{h}_1)}{||\mathbf{h}_1||_2} - \frac{B(\mathbf{h}_2)}{||\mathbf{h}_2||_2}\Big| &\leq \frac{2}{n}||\mathcal{X}\big(\mathbf{h}_1/||\mathbf{h}_1||_2 - \mathbf{h}_2/||\mathbf{h}_2||_2\big)||_1 \\ &\leq \frac{2}{\sqrt{n}}||\mathcal{X}\big(\mathbf{h}_1/||\mathbf{h}_1||_2 - \mathbf{h}_2/||\mathbf{h}_2||_2\big)||_2 \\ &\leq 2\sqrt{M_1}\epsilon. \end{aligned}$$

Therefore,

$$\sup_{||\mathbf{h}||_0 \leq q+1, ||\mathbf{h}||_2 \neq 0} \frac{B(\mathbf{h})}{||\mathbf{h}||} \leq \sup_{\mathbf{h} \in N} \frac{B(\mathbf{h})}{||\mathbf{h}||} + 2\sqrt{M_1}\epsilon.$$

Let $\epsilon = \sqrt{\frac{q \log p}{2M_1 n}}$, we have

$$\begin{aligned} & P\left(\sup_{||\mathbf{h}||_0 \leq q+1, ||\mathbf{h}||_2 \neq 0} \frac{B(\mathbf{h})}{||\mathbf{h}||_2} \geq C\sqrt{\frac{2q \log p}{n}}\right) \\ \leq\ & P\left(\sup_{\mathbf{h} \in N} \frac{B(\mathbf{h})}{||\mathbf{h}||_2} \geq (C-1)\sqrt{\frac{2q \log p}{n}}\right) \\ \leq\ & 2\big(\frac{2M_1 n}{q \log p}\big)^{\frac{q+1}{2}}\big(3p^{1-(C-1)^2/(2M_1)}\big)^{q+1} \\ \leq\ & 2\big(\sqrt{2M_1 n} 3p^{1-(C-1)^2/(2M_1)}\big)^{q+1}. \end{aligned}$$

Since $p > n$, take $C = 1 + 2C_1\sqrt{M_1}$ for some $C_1 > 1$, then for all n sufficiently large,

$$P\Big(\sup_{||\mathbf{h}||_0 \leq q+1, ||\mathbf{h}||_2 \neq 0} \frac{B(\mathbf{h})}{||\mathbf{h}||} \geq (1 + 2C_1\sqrt{M_1})\sqrt{\frac{2q \log p}{n}}\Big) \leq 2p^{-2q(C_1^2-1)}.$$

Lemma 6 *For any $x \in \mathbb{R}^n$,*

$$||x||_2 - \frac{||x||_1}{\sqrt{n}} \leq \frac{\sqrt{n}}{4}\big(\max_{1 \leq i \leq n} |x_i| - \min_{1 \leq i \leq n} |x_i|\big).$$

■

Proof. This proof was given in Cai, Wang, and Xu (2010). We include it here for completeness and easy reference. It is obvious that the result holds when $|x_1| = |x_2| = \ldots = |x_n|$. Without loss of generality, we now assume that $x_1 \geq x_2 \geq \ldots \geq x_n \geq 0$ and not all x_i are equal. Let

$$f(x) = ||x||_2 - \frac{||x||_1}{\sqrt{n}}.$$

Note that for any $i \in \{2, 3, \ldots, n-1\}$

$$\frac{\partial f}{\partial x_i} = \frac{x_i}{||x||_2} - \frac{1}{\sqrt{n}}.$$

This implies that when $x_i \leq \frac{||x||_2}{\sqrt{n}}$, $f(x)$ is decreasing w.r.t x_i; otherwise $f(x)$ is increasing w.r.t x_i. Hence, if we fix x_1 and x_n, when $f(x)$ achieves its maximum, x must be of the form that $x_1 = x_2 = \ldots = x_k$ and $x_{k+1} = \ldots = x_n$ for some $1 \leq k \leq n$. Now,

$$f(x) = \sqrt{k(x_1^2 - x_n^2) + nx_n^2} - \frac{k}{\sqrt{n}}(x_1 - x_n) - \sqrt{n}x_n.$$

Treat this as a function of k for $k \in (0, n)$.

$$g(x) = \sqrt{k(x_1^2 - x_n^2) + nx_n^2} - \frac{k}{\sqrt{n}}(x_1 - x_n) - \sqrt{n}x_n.$$

By taking the derivatives, it is easy to see that

$$\begin{aligned} g(k) &\leq g\left(n\frac{(\frac{x_1+x_n}{2})^2 - x_n^2}{x_1^2 - x_n^2}\right) \\ &= \sqrt{n}(x_1 - x_n)\left(\frac{1}{2} - \frac{x_1 + 3x_n}{4(x_1 + x_n)}\right). \end{aligned}$$

Since $\frac{1}{2} - \frac{x_1+3x_n}{4(x_1+x_n)} \geq \frac{1}{4}$, we have

$$||x||_2 \leq \frac{||x||_1}{\sqrt{n}} + \frac{\sqrt{n}}{4}(x_1 - x_n).$$

We can also see that the above inequality becomes an equality if and only if $x_{k+1} = \ldots = x_n = 0$ and $k = \frac{n}{4}$.

Proof of Theorem 4. Let $\mathbf{h} = \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}$, then it follows from Lemma 2 that $\mathbf{h} \in \Delta_{\bar{C}}$. Assume without loss of generality that $|h_0| \geq |h_1| \geq \ldots \geq |h_p|$. Create a partition of $\{0, 1, 2, \ldots, p\}$ as

$$S_0 = \{0, 1, 2, \ldots, q\}, S_1 = \{q + 1, q + 2 \ldots, 2q + 1\}, S_2 = \{2q + 2, 2q + 3 \ldots, 3q + 2\}, \ldots$$

where S_i, $i = 1, 2, \ldots$, has cardinality q+1, except the last set which may have cardinality smaller than q+1. This partition leads to the following decomposition

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \frac{1}{n}\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \\ = \;& \frac{1}{n}\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k\geq 0}\mathbf{h}_{S_k})_+ - \frac{1}{n}\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+ \\ = \;& \sum_{j\geq 1}\frac{1}{n}\left(\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j}\mathbf{h}_{S_k})_+ - \sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j-1}\mathbf{h}_{S_k})_+\right) \\ &+\frac{1}{n}\left(\sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h}_{S_0})_+ - \sum_{i=1}^n(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+\right), \end{aligned} \tag{16}$$

where the first equation follows from the definition of $\mathbf{h}_{S_k}$, $k \geq 0$; the second equation holds by observing that the intermediate terms cancel out each other. The purpose of the above decomposition is to obtain more accurate probability bounds by appealing to Lemma 3. This is made possible by noting that the jth term in the sum of the above decomposition has the increment indexed by $\mathbf{h}_{S_j}$, which has at most q+1 nonzero coordinates. Lemma 3 implies that uniformly for $j = 1, 2, \ldots$, with probability at least $1 - 2p^{-2q(C_1^2-1)}$,

$$\begin{aligned}
& \frac{1}{n}\left(\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j}\mathbf{h}_{S_k})_+ - \sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j-1}\mathbf{h}_{S_k})_+\right) \\
\geq\ & \frac{1}{n}E\left(\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j}\mathbf{h}_{S_k})_+ - \sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\sum_{k=0}^{j-1}\mathbf{h}_{S_k})_+\right) \\
& -C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_j}||_2,
\end{aligned}$$

where $C = 1 + 2C_1\sqrt{M_1}$. Hence by (16), with probability at least $1 - 2p^{-2q(C_1^2-1)}$,

$$\frac{1}{n}\Big(\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+\Big) \geq M(\mathbf{h}) - C\sqrt{\frac{2q\log p}{n}}\sum_{j\geq 0}||\mathbf{h}_{S_j}||_2 \tag{17}$$

where $M(\mathbf{h}) = \frac{1}{n}E(\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^* + Y_i\mathbf{X}_i^T\mathbf{h})_+ - \sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta}^*)_+)$.

It is straightforward to show that $||\mathbf{h}_{S_0}||_1 \geq ||\mathbf{h}_{T_+}||_1 \geq \bar{C}||\mathbf{h}_{T_+^C}||_1 \geq \bar{C}||\mathbf{h}_{S_0^C}||_1$. By Lemma 6, we have

$$\begin{aligned}
\sum_{j\geq 1}||\mathbf{h}_{S_j}||_2 &\leq \sum_{j\geq 1}\frac{||\mathbf{h}_{S_j}||_1}{\sqrt{q+1}} + \frac{\sqrt{q+1}}{4}|h_q| \\
&\leq \frac{||\mathbf{h}_{S_0^C}||_1}{\sqrt{q+1}} + \frac{||\mathbf{h}_{S_0}||_1}{4\sqrt{q+1}} \\
&\leq (\frac{1}{\sqrt{q+1}\bar{C}} + \frac{1}{4\sqrt{q+1}})||\mathbf{h}_{S_0}||_1 \\
&\leq (\frac{1}{4} + \frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2. \qquad (18)
\end{aligned}$$

By the definition of $\mathbf{h}$, (15), (17) and (18), we have

$$\begin{aligned}
M(\mathbf{h}) &\leq \lambda(||\mathbf{h}_{T_+}||_1 - ||\mathbf{h}_{T_+^C}||_1) + (\frac{1}{4} + \frac{1}{\bar{C}})C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_0}||_2 + C\sqrt{\frac{2q\log p}{n}}||\mathbf{h}_{S_0}||_2 \\
&\leq \lambda\sqrt{q+1}||\mathbf{h}_{S_0}||_2 + C\sqrt{\frac{2q\log p}{n}}(\frac{5}{4} + \frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2. \qquad (19)
\end{aligned}$$

Condition (A4) imply that

$$M(\mathbf{h}) = \frac{1}{2}\mathbf{h}^T H(\boldsymbol{\beta}^*)\mathbf{h} + o(||\mathbf{h}||_2^2) \geq \frac{1}{2}M_2||\mathbf{h}||_2^2 + o(||\mathbf{h}||_2^2). \tag{20}$$

Combining (19) and (20), we have

$$\frac{1}{2}M_2||\mathbf{h}||_2^2 + o(||\mathbf{h}||_2^2) \quad \leq \quad \lambda\sqrt{q+1}||\mathbf{h}_{S_0}||_2 + C\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2.$$

Note that $||\mathbf{h}||_2^2 = ||\mathbf{h}_{S_0}||_2^2 + \sum_{j\geq 1}||\mathbf{h}_{S_j}||_2^2 \geq ||\mathbf{h}_{S_0}||_2^2$, and

$$\sum_{j\geq 1}||\mathbf{h}_{S_j}||_2^2 \quad \leq \quad |h_q|\sum_{j\geq 1}||\mathbf{h}_{S_j}||_1 \leq \frac{1}{\bar{C}}|h_q|||\mathbf{h}_{S_0}||_1 \leq \frac{1}{\bar{C}}||\mathbf{h}_{S_0}||_2^2.$$

So $||\mathbf{h}||_2^2 \leq (1+\frac{1}{\bar{C}})||\mathbf{h}_{S_0}||_2^2$. This implies $o(||\mathbf{h}||_2^2) = o(||\mathbf{h}_{S_0}||_2^2)$. To wrap up, we have

$$||\mathbf{h}_{S_0}||_2 + o(||\mathbf{h}_{S_0}||_2) \leq \frac{2\lambda\sqrt{q+1}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\bar{C}}).$$

Hence,

$$||\mathbf{h}||_2 + o(||\mathbf{h}||_2) \leq \sqrt{1+\frac{1}{\bar{C}}}\left(\frac{2\lambda\sqrt{q+1}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\bar{C}})\right).$$

We therefore have

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leq \sqrt{1+\frac{1}{\bar{C}}}\Big(\frac{2\lambda\sqrt{q+1}}{M_2} + \frac{2C}{M_2}\sqrt{\frac{2q\log p}{n}}(\frac{5}{4}+\frac{1}{\bar{C}})\Big)$$

with probability at least $1-2p^{-2q(C_1^2-1)}$.

Proof of Theorem 5. It follows by combining the result of Theorem 3.3 with that of Theorem 4 in of Zhang et al. (2016b).

Appendix B: Discussions of Condition (A4)

We note that Condition (A4) is satisfied if the smallest eigenvalues of $H(\boldsymbol{\beta}^*)$ has a positive lower bound. In the following, we provide a set of sufficient conditions to guarantee the positive definiteness of $H(\boldsymbol{\beta}^*)$.

(A1*) For some $1 \leq k \leq p$,

$$\int_{\mathcal{S}} I(X_k \geq V_k^-)X_i g(\mathbf{X})d\mathbf{X} < \int_{\mathcal{S}} I(X_k \leq U_k^+)X_i f(\mathbf{X})d\mathbf{X}$$

or

$$\int_{\mathcal{S}} I(X_k \leq V_k^+)X_i g(\mathbf{X})d\mathbf{X} > \int_{\mathcal{S}} I(X_k \geq U_k^-)X_i f(\mathbf{X})d\mathbf{X}$$

Here U_k^+, $V_k^+ \in [-\infty, +\infty]$ are upper bounds such that $\int_{\mathcal{S}} I(X_k \leq U_k^+)f(\mathbf{X})d\mathbf{X} = \min(1, \frac{\pi_-}{\pi_+})$ and $\int_{\mathcal{S}} I(X_k \leq V_k^+)f(\mathbf{X})d\mathbf{X} = \min(1, \frac{\pi_+}{\pi_-})$. Similarly, lower bounds U_k^-, $V_k^- \in [-\infty, +\infty]$ and are defined as $\int_{\mathcal{S}} I(X_k \geq U_k^-)f(\mathbf{X})d\mathbf{X} = \min(1, \frac{\pi_-}{\pi_+})$ and $\int_{\mathcal{S}} I(X_k \geq V_k^-)g(\mathbf{X})d\mathbf{X} = \min(1, \frac{\pi_+}{\pi_-})$.

(A2*) For an orthogonal transformation A_j that maps $\dfrac{\boldsymbol{\beta}^*_-}{||\boldsymbol{\beta}^*_-||_2}$ to the j-th unit vector $\mathbf{e}_j$ for some $j \in \{1, 2, 3, \ldots, p\}$, there exists rectangles

$$D^+ = \{x \in M^+ : l_i \le (A_j x)_i \le v_i \text{ with } l_i < v_i \text{ for } i \ne j\}$$

and

$$D^- = \{x \in M^- : l_i \le (A_j x)_i \le v_i \text{ with } l_i < v_i \text{ for } i \ne j\}$$

such that $f(x) \ge B_1 > 0$ on D^+, and $g(x) \ge B_2 > 0$ on D^-, where $M^+ = \{x \in \mathbf{R}^p | x^T \boldsymbol{\beta}^*_- + \boldsymbol{\beta}^*_0 = 1\}$ and $M^- = \{x \in \mathbf{R}^p | x^T \boldsymbol{\beta}^*_- + \boldsymbol{\beta}^*_0 = -1\}$.

Also with some technical modification, Condition (A1) in our paper can be further relaxed to

(A3*) The densities f and g are continuous with common support $\mathcal{S} \subset \mathbb{R}^p$ and have finite second moments.

As an interesting side result, Lemma 5 in Koo et al. (2008) showed that Condition (A4) holds under (A1*)-(A3*). Although their paper's results on the Bahadur representation of L_1-norm SVM coefficients are restricted to the classical fixed p case, a careful examination of the derivation showed that this particular lemma holds irrespective of the dimension of p.

In the following, we demonstrate that Conditions (A1*)-(A3*) hold in a nontrivial example where we have two multivariate normal distributions in $\mathcal{R}^p$. The marginal distribution of Y is given by $\pi_+ = \pi_- = 1/2$. Let f and g be the density functions of $\mathbf{X}_-$ given $Y = 1$ and -1, respectively. Here, we assume f and g are multivariate normal densities with different mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and a common covariance matrix $\boldsymbol{\Sigma}$. This setup was also considered in Koo et al. (2008) but we will provide more details to show condition (A4) is satisfied in our high-dimensional setting. In particular, we will provide some details for deriving the analytic forms of $\boldsymbol{\beta}^*$ and $H(\boldsymbol{\beta}^*)$, which complements the results in Koo et al. (2008).

For normal density functions f and g, it is straightforward to check Condition (A3*) is satisfied. While $U^+_k = V^+_k = +\infty$ and $U^-_k = V^-_k = -\infty$, Condition (A1*) also holds. Since D^+ and D^- are bounded rectangles in $\mathbb{R}^p$, the normal densities f and g are always bounded away from zero on D^+ and D^-. Thus (A2*) is satisfied. Denote the density and cumulative distribution function of standard normal distribution $N(0, 1)$ as ϕ and Φ, respectively. Then we have $S(\boldsymbol{\beta}^*) = 0$, where $S(\cdot)$ is defined in (6), that is

$$E_f(I(1 - \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)) = E_g(I(1 + \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)) \tag{21}$$

and

$$E_f(I(1 - \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)\mathbf{X}_-) = E_g(I(1 + \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)\mathbf{X}_-) \tag{22}$$

For left hand of equation (21), we have $\mathbf{X}^T_- \boldsymbol{\beta}^*_- \sim N(\boldsymbol{\mu}^T \boldsymbol{\beta}^*_-, \boldsymbol{\beta}^{*T}_- \boldsymbol{\Sigma} \boldsymbol{\beta}^*_-)$, thus

$$E_f(I(1 - \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)) = P_f(1 - \beta^*_0 - \mathbf{X}^T_- \boldsymbol{\beta}^*_- \ge 0) = \Phi(c_f), \tag{23}$$

where $c_f = \frac{1 - \beta^*_0 - \boldsymbol{\mu}^T \boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*_-||_2}$. Similarly, $E_g(I(1 + \mathbf{X}^T \boldsymbol{\beta}^* \ge 0)) = \Phi(c_g)$. where $c_g = \frac{1 + \beta^*_0 + \boldsymbol{\nu}^T \boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}^*_-||_2}$.

To obtain an analytic expression of $E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0))$, we consider an orthogonal matrix $\mathbf{P}$ that satisfies $\frac{\mathbf{P}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2} = (1, 0, 0, \ldots, 0)^T$. Such a matrix $\mathbf{P}$ can always be constructed. Actually, let $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_p)^T$ and $\mathbf{P}_1 = \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2}$. By using Gram-Schmidt process, we can generate other orthogonal vectors $\mathbf{P}_i$ based on $\mathbf{P}_1$ with $i = 2, 3, \ldots, p$. Since $\mathbf{P}\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}_- - \boldsymbol{\mu}) = \mathbf{Z}$, a standard multivariate normal random vector, we have $I - \mathbf{X}^T\boldsymbol{\beta}^* = c_f||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2 - Z^T\mathbf{P}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-$. Thus

$$
\begin{aligned}
E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) &= E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)(\boldsymbol{\Sigma}^{1/2}\mathbf{P}^T\mathbf{Z} + \boldsymbol{\mu})) \\
&= E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)\boldsymbol{\mu}) + E_{\boldsymbol{\phi}}(I(c_f - Z_1 \geq 0)\boldsymbol{\Sigma}^{1/2}\mathbf{P}^T\mathbf{Z}).
\end{aligned}
$$

where $\boldsymbol{\phi}$ is the joint probability density function of a p-dimensional standard multivariate normal distribution. We will compute the above expectation componentwise. Let $\Sigma^{1/2} = \Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_p)^T$. For $k = 1, \ldots, p$, we have

$$
E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)X_k) = \mu_k\Phi(c_f) + E_{\boldsymbol{\phi}}\big(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^{p} P_iZ_i\big)
$$

where. Since $Z_2, \ldots, Z_p$ have mean zero and are independent of Z_1,

$$
\begin{aligned}
E_{\boldsymbol{\phi}}\big(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^{p}(P_iZ_i)\big) &= E_\phi\big(I(c_f - Z_1 \geq 0)\Lambda_k^T P_1 Z_1\big) \\
&= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2}E_\phi\big(I(c_f - Z_1 \geq 0)Z_1\big) \\
&= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2}\int_{-\infty}^{+\infty} I\big(c_f - x \geq 0\big)x\phi(x)dx \\
&= \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2}\int_{-\infty}^{c_f} x\phi(x)dx
\end{aligned}
$$

Since $x\phi(x)$ is an odd function and $\phi(x)$ is symmetric, we have

$$
\int_{-\infty}^{c_f} x\phi(x)dx = \int_{-\infty}^{|c_f|} x\phi(x)dx = -\frac{1}{\sqrt{2\pi}}\int_{c_f^2}^{+\infty}\frac{1}{2}\exp(-\frac{z}{2})dz = -\phi(c_f).
$$

Therefore, for $k = 1, \ldots, p$, $E_f\big(I(1-\mathbf{X}^T\boldsymbol{\beta}^* \geq 0)X_k\big) = \mu_k\Phi(c_f) - \Lambda_k^T\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}^*_-||_2}\phi(c_f)$. Hence

$$
E_f(I(1 - \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) = \boldsymbol{\mu}\Phi(c_f) - \phi(c_f)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1.
$$

Similarly,

$$
E_g(I(1 + \mathbf{X}^T\boldsymbol{\beta}^* \geq 0)\mathbf{X}_-) = \boldsymbol{\nu}\Phi(c_g) + \phi(c_g)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1
$$

Then, we have

$$\Phi(c_f) = \Phi_(c_g) \tag{24}$$

and

$$\boldsymbol{\mu}\Phi(c_f) - \phi(c_f)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1 = \boldsymbol{\nu}\Phi(c_g) + \phi(c_g)\boldsymbol{\Sigma}^{1/2}\mathbf{P}_1 \tag{25}$$

From (24), we have $\tilde{c} = c_f = c_g$, which implies

$$\boldsymbol{\beta}_-^{*T}(\boldsymbol{\mu} + \boldsymbol{\nu}) = -2\beta_0^* \tag{26}$$

From (25),

$$\frac{\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2} = \frac{\Phi(\tilde{c})}{2\phi(\tilde{c})}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu}) \tag{27}$$

Let $d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) = ((\boldsymbol{\mu}-\boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu}))^{1/2}$ be the Mahalanobis distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ and $R(x) = \frac{\phi(x)}{\Phi(x)}$. As $\boldsymbol{\Sigma}^{1/2}\dfrac{\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$ has l_2 norm equal to 1, we have $||\frac{\Phi(\tilde{c})}{2\phi(\tilde{c})}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}-\boldsymbol{\nu})||_2 = 1$, *i.e.*, $R(\tilde{c}) = \frac{d_\Sigma(\boldsymbol{\mu},\boldsymbol{\nu})}{2}$. $R(x)$ is a monotonically decreasing function, thus we have $\tilde{c} = R^{-1}\big(\frac{d_\Sigma(\boldsymbol{\mu},\boldsymbol{\nu})}{2}\big)$. Meanwhile, $\tilde{c} = c_f = \frac{1-\beta_0^*-\boldsymbol{\mu}^T\boldsymbol{\beta}_-^*}{||\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_-^*||_2}$, we can solve the problem based on (26) and (27),

$$\beta_0^* = -\frac{(\boldsymbol{\mu} - \boldsymbol{\nu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} + \boldsymbol{\nu})}{2\tilde{c}d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) + d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})} \tag{28}$$

From (25),

$$\boldsymbol{\beta}_-^* = \frac{2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})}{2\tilde{c}d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}) + d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})} \tag{29}$$

By plugging (28) and (29) into (7), we can calculate $H(\boldsymbol{\beta}^*)$ as

$$H(\boldsymbol{\beta}^*) = \frac{\phi(\tilde{c})}{4}(2\tilde{c} + d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu}))\begin{pmatrix} 2 & (\boldsymbol{\mu} + \boldsymbol{\nu})^T \\ \boldsymbol{\mu} + \boldsymbol{\nu} & H_{22}(\boldsymbol{\beta}^*) \end{pmatrix} \tag{30}$$

where

$$H_{22}(\boldsymbol{\beta}^*) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\nu}\boldsymbol{\nu}^T + 2\boldsymbol{\Sigma} + 2\left(\left(\frac{\tilde{c}}{d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu})}\right)^2 + \frac{\tilde{c}}{d_\Sigma(\boldsymbol{\mu}, \boldsymbol{\nu})} - \frac{1}{d_\Sigma^2(\boldsymbol{\mu}, \boldsymbol{\nu})}\right)(\boldsymbol{\mu} - \boldsymbol{\nu})(\boldsymbol{\mu} - \boldsymbol{\nu})^T$$

As we have obtained the analytic form of $H(\boldsymbol{\beta}^*)$, we consider Model 1 in Section 5.1 as an example. In Model 1, $q = 5$, $\boldsymbol{\mu} = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T$ and $\boldsymbol{\nu} = (-0.1, -0.2, -0.3, -0.4, -0.5, 0, \ldots, 0)^T \in \mathbb{R}^p$ and $\pi^+ = \pi^- = 1/2$. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ consists of nonzero entries $\sigma_{ij} = -0.2$ for $1 \leq i \neq j \leq q$ and other entries equal to 0. From (28) and (29), we have $\boldsymbol{\beta}^* = (0, 1.39, 1.47, 1.56, 1.65, 1.74, 0, \ldots, 0)^T$. Based on (30), we derived $H(\boldsymbol{\beta}^*)$ and numerically validated its positive-definiteness.

References

Natalia Becker, Grischa Toedt, Peter Lichter, and Axel Benner. Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC bioinformatics*, 12(1):138, 2011.

Alexandre Belloni and Victor Chernozhukov. l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

Jean Bourgain and Vitaly D Milman. New volume ratio properties for convex symmetric bodies in $\mathcal{R}^n$. *Inventiones Mathematicae*, 88(2):319–340, 1987.

Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

Tony Cai, Lie Wang, and Guangwu Xu. New bounds for restricted isometry constants. *IEEE Transactions on Information Theory*, 56(9):4388–4394, 2010.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear svm. *Causation and Prediction Challenge Challenges in Machine Learning*, 2:47, 2008.

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584, 2007.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Eitan Greenshtein et al. Best subset selection, persistence in high-dimensional statistical learning and optimization under l1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

Ja-Yong Koo, Yoonkyung Lee, Yuwon Kim, and Changyi Park. A bahadur representation of the linear support vector machine. *The Journal of Machine Learning Research*, 9: 1343–1368, 2008.

Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

Changyi Park, Kwang-Rae Kim, Rangmi Myung, and Ja-Yong Koo. Oracle properties of scad-penalized support vector machine. *Journal of Statistical Planning and Inference*, 142(8):2257–2270, 2012.

Guilherme V Rocha, Xing Wang, and Bin Yu. Asymptotic distribution and sparsistency for l1-penalized parametric m-estimators with applications to linear svm and logistic regression. *arXiv preprint arXiv:0908.1940*, 2009.

CA Rogers. Covering a sphere with spheres. *Mathematika*, 10(02):157–164, 1963.

Minghu Song, Curt M Breneman, Jinbo Bi, Nagamani Sukumar, Kristin P Bennett, Steven Cramer, and Nihal Tugcu. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *Journal of Chemical Information and Computer Sciences*, 42(6):1347–1357, 2002.

Bernadetta Tarigan and Sara Anna van de Geer. *Adaptivity of Support Vector Machines with L_1 Penalty.* University of Leiden. Mathematical Institute, 2004.

Bernadetta Tarigan, Sara A Van De Geer, et al. Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 1995.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Lie Wang. The l_1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

Lifeng Wang and Xiaotong Shen. On l_1-norm multiclass support vector machines: methodology and theory. *Journal of the American Statistical Association*, 102(478):583–594, 2007.

Marten Wegkamp and Ming Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Hao Helen Zhang, Jeongyoun Ahn, Xiaodong Lin, and Cheolwoo Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1):88–95, 2006.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. A consistent information criterion for support vector machines in diverging model spaces. *Journal of Machine Learning Research*, 17(16):1–26, 2016a.

Xiang Zhang, Yichao Wu, Lan Wang, and Runze Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):53–76, 2016b.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems*, 16(1):49–56, 2004.

Hui Zou. An improved 1-norm svm for simultaneous classification and variable selection. *Journal of Machine Learning Research*, 2:675–681, 2007.