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

Bo Peng  
Lan Wang  
School of Statistics  
University of Minnesota  
Minneapolis, MN 55455, USA

Yichao Wu  
Department of Statistics  
North Carolina State University  
Raleigh, NC 27695, USA

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

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

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^n (1 - Y_i X_i^T \beta)_+ + \lambda ||\beta_-||_2^2,$$

where $(1-u)_+ = \max\{1-u, 0\}$ is often called the hinge loss function, $\lambda$ is a tuning parameter and $\beta = (\beta_0, (\beta_-)^T)^T$ with $\beta_- = (\beta_1, \beta_2, \ldots, \beta_p)^T$. Generally for a given vector $e$, we use $e_-$ to denote the subvector with the first entry of $e$ 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

$$l_n(\beta, \lambda) = n^{-1} \sum_{i=1}^{n} (1 - Y_i X^T_i \beta)_+ + \lambda ||\beta||_1,$$

(2)

and define

$$\hat{\beta}(\lambda) = \arg\min_{\beta} l_n(\beta, \lambda).$$

(3)

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

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

$$\min_{\xi, \zeta, \beta} \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} \zeta_j \right)$$

subject to

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, n,$$

$$\xi_i \geq 1 - Y_i X^T_i \beta, \quad i = 1, 2, \ldots, n,$$

$$\zeta_j \geq \beta_j, \zeta_j \geq -\beta_j, \quad j = 1, 2, \ldots, p.$$

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 $L_1$-norm SVM coefficients under the ultra-high dimensional setting. The choice of the tuning parameter $\lambda$ will be studied to achieve this error bound.

3.1 Preliminaries

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

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

(5)

where $L(\beta) = E(1 - Y \mathbf{X}^T \beta)_+$. 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 $\beta^*$ 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 $||\hat{\beta}(\lambda) - \beta^*||_2$ when $p \gg n$. In the ultra-high dimensional settings, it is often reasonable to assume that $\beta^* = (\beta^*_0, \beta^*_1, \ldots, \beta^*_p)^T$ 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 $T_+ = T \cup \{0\}$.
Next, we introduce the gradient vector and Hessian matrix of the population hinge loss function $L(\beta)$. We define
\[
S(\beta) = -E(I(1 - YX^T\beta \geq 0)YX)
\] (6)
as the $(p + 1)$-dimensional gradient vector and
\[
H(\beta) = E(\delta(1 - YX^T\beta)XX^T)
\] (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 $S(\beta)$ and $H(\beta)$ under certain conditions.

Throughout the paper, we assume the following regularity condition.

(A1) The densities $f$ and $g$ are continuous with common support $S \subset \mathbb{R}^p$ 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 $H(\beta)$ is well defined and continuous in $\beta$. The bound of $X_\cdot$ 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 $\lambda$ and a fact about $\hat{\beta}$

The estimated $L_1$-norm SVM parameter $\hat{\beta}(\lambda)$ defined in (3) depends on the tuning parameter $\lambda$. We will first show that a universal choice
\[
\lambda = c \sqrt{2A(\alpha) \log p/n},
\] (8)
where $c$ is some given constant, $\alpha$ is a small probability and $A(\alpha) > 0$ is a constant such that $4p^{-\frac{A(\alpha)}{3\alpha^2}} + 1 \leq \alpha$, can provide theoretical guarantee on the good performance of $\hat{\beta}(\lambda)$.

The above choice of $\lambda$ is motivated by a principle in the setting of penalized least squares regression (Bickel et al., 2009), which advocates to choose the penalty level $\lambda$ to dominate the subgradient of the loss function evaluated at the true value. Intuitively, the subgradient evaluated at $\beta^*$ 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 $\hat{\beta}$ 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
\[
\hat{S}(\beta) = -n^{-1} \sum_{i=1}^{n} I(1 - Y_iX_i^T\beta \geq 0)Y_iX_i.
\]

Then it follows from the argument as in Theorem 3.1 of Zhang et al. (2016b) that under some weak regularity conditions $||\hat{S}(\hat{\beta})||_\infty \leq \lambda$ with probability approaching one. It follows from Koo et al. (2008) that the oracle estimator $\hat{\beta}$ provides a consistent and asymptotically normal estimate of $\beta^*$. 5
Hence, in the ideal case where the population parameter $\beta^*$ is known, an intuitive choice of $\lambda$ is to set its value to be larger than the supremum norm of $\hat{S}(\beta^*)$ with large probability, that is

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

(9)

where $c > 1$ is some given constant and $\alpha$ is a small probability. Lemma 1 below shows that the choice of $\lambda$ 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

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

with $\alpha$ 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 $\sum_{n=1}^n I(1 - Y_i X_i^T \beta^* \geq 0) Y_i X_i$ by applying Hoeffding’s inequality and the union bound. Later in this section, we will show that this choice of $\lambda$ warrants near-oracle rate performance of $\hat{\beta}(\lambda)$. Let $h = \beta^* - \hat{\beta}(\lambda)$. We state below an interesting fact about $h$.

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

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

where

$$\Delta_C = \{ \gamma \in \mathbb{R}^{p+1} : ||\gamma_{T_+}||_1 \geq \bar{C}||\gamma_{T_+^c}||_1, \text{ where } T_+ = T \cup \{0\}, T \subset \{1, 2, \ldots, p\} \text{ and } |T| \leq q \},$$

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

We call $\Delta_C$ the restricted set. The proof of Lemma 2 is also given in Appendix A.

### 3.3 Regularity conditions

Let $\mathcal{X} = (X_1, X_2, \ldots, X_n)^T$ denote the feature design matrix. We define restricted eigenvalues as follows

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

(10)

and

$$\lambda_{\min}(H(\beta^*); q) = \min_{\mathbf{d} \in \Delta_C} \frac{\mathbf{d}^T H(\beta^*) \mathbf{d}}{||\mathbf{d}||_2^2}.$$ 

(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) $\lambda_{\min}(H(\beta^*); q) \geq M_2$, 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 $X^T \beta^*$ given $Y = +1$ and $Y = -1$ as $f^*$ and $g^*$, respectively. It is assumed that $f^*$ is uniformly bounded away from 0 and $\infty$ in a neighborhood of 1 and $g^*$ is uniformly bounded away from 0 and $\infty$ 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 $\beta^*$. 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 $H(\beta^*)$ 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 $\widehat{\beta}(\lambda)$ 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 $h \in \mathbb{R}^{p+1}$, let

$$B(h) = \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+$$

$$- E \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \right).$$

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

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

where $C_1 > 1$ is a constant.

Lemma 3 guarantees that $n^{-1} \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \right)$ 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.
Theorem 4 Suppose that conditions (A1)-(A6) hold, then the estimated $L_1$-norm SVM coefficients vector $\hat{\beta}(\lambda)$ satisfies

$$||\hat{\beta}(\lambda) - \beta^*||_2 \leq \sqrt{1 + \frac{1}{C} \left( \frac{2\lambda\sqrt{q} + 1}{M_2} + \frac{2C}{M_2} \sqrt{\frac{2q\log p}{n}} \left( \frac{5}{4} + \frac{1}{C} \right) \right)}$$

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,

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

when $\lambda = c\sqrt{2A(\alpha)\log p/n}$. Actually, in the inequality of Theorem 4, the first term satisfies

$$\frac{\lambda\sqrt{q}}{M_2} = \frac{2}{M_2} \sqrt{\frac{2A(\alpha)\log p}{n}} = O \left( \sqrt{\frac{q\log p}{n}} \right)$$

and it is also trivial to have the second term of the same order. Hence the near-oracle property of $\hat{\beta}(\lambda)$ will hold given $\lambda$ 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 $p = 0.1 \ast n^2$, $q = \lfloor n^{1/3} \rfloor$ and $\beta^* = ((1.1, \ldots, 1.1), 0, \ldots, 0)^T$, which allows $p$ and $q$ to vary with sample size $n$. Figure 1 depicts the average of $||\hat{\beta}(\lambda) - \beta^*||_2$ 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 $\lambda$. 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 $\beta^*$ is defined as

$$\tilde{\beta} = \arg \min_{\beta, \beta_{T^c}} \tilde{l}_n(\beta),$$

(12)

where $\tilde{l}_n(\beta) = n^{-1} \sum_{i=1}^{n}(1 - Y_i X_i^T \beta)_{+}$ is the sample hinge loss function and $\beta_{T^c}$ denotes the vector containing the components of $\beta$ 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 \( \beta^{(0)} \). At each step \( t \), we update the \( \beta \) to be \( \beta^{(t)} \) by solving

\[
\min_{\beta} \left\{ n^{-1} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta)_+ + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(t-1)}|)|\beta_j| \right\},
\]

(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 \( \tilde{\beta}^{(0)} = (\tilde{\beta}_0^{(0)}, \tilde{\beta}_1^{(0)}, \ldots, \tilde{\beta}_p^{(0)})^T \) that satisfies

\[
P(|\tilde{\beta}_j^{(0)} - \beta_j^*| > \lambda, \text{ for some } 1 \leq j \leq p) \to 0 \text{ as } n \to \infty.
\]

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 \( \tilde{\beta}(\lambda) \) 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 \( P(|\tilde{\beta}_j(\lambda) - \beta_j^*| > \lambda, \text{ for some } 1 \leq j \leq p) \to 0 \text{ as } n \to \infty \). Furthermore, the LLA algorithm initiated by \( \tilde{\beta}(\lambda) \) finds the oracle estimator in two iterations with probability tending to 1, i.e., \( P(\tilde{\beta}^{nc}(\lambda) = \beta) \), where \( \tilde{\beta}^{nc}(\lambda) \) is the solution for non-convex penalized SVM with given \( \lambda \).

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: \( Pr(Y = 1) = Pr(Y = -1) = 0.5, X_\sim|(Y = 1) \sim MN(\mu, \Sigma), X_\sim|(Y = -1) \sim MN(-\mu, \Sigma), q = 5, \mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T \in \mathbb{R}^p, \Sigma = (\sigma_{ij}) \) with diagonal entries equal to 1, nonzero entries \( \sigma_{ij} = -0.2 \) for \( 1 \leq i \neq j \leq q \) 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: \( Pr(Y = 1) = Pr(Y = -1) = 0.5, X_-|(Y = 1) \sim MN(\mu, \Sigma), X_-|(Y = -1) \sim MN(-\mu, \Sigma), q = 5, \mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T \in \mathbb{R}^p, \Sigma = (\sigma_{ij}) \) 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 \( \Sigma = (\sigma_{ij}) \) 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 \( \text{sign}(3.09X_1 + 4.45X_2 + 5.06X_3 + 4.77X_4 + 3.58X_5) \) with Bayes error 0.6%.

- Model 4: \( X_- \sim MN(0_p, \Sigma), \Sigma = (\sigma_{ij}) \) with nonzero elements \( \sigma_{ij} = 0.4^{i-j} \) for \( 1 \leq i, j \leq p \), \( Pr(Y = 1|X_-) = \Phi(X^T\beta^*) \), where \( \Phi(\cdot) \) is the cumulative density function of the standard normal distribution, \( \beta^* = (1.1, 1.1, 1.1, 1.1, 0, \ldots, 0)^T \) 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 discriminant 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 \( \lambda \) 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 \( ||\hat{\beta}(\lambda) - \beta^*||_2 \) where Appendix B provides details on the calculation of \( \beta^* \) and the absolute value of the sample correlation between \( X^T\beta(\lambda) \) and \( X^T\beta^* \). 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**: \( ||\hat{\beta}(\lambda) - \beta^*||_2 \).
- **AAC**: Absolute absolute correlation \( \text{corr}(X^T\hat{\beta}(\lambda), X^T\beta^*) \).
- **Signal**: the average of number of nonzero regression coefficients \( \hat{\beta}_i \neq 0 \) 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 \( \hat{\beta}_i(\lambda) \neq 0 \) 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

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

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 $\hat{\beta}(\lambda)$ 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

\[
SVMIC_\gamma(T) = \sum_{i=1}^{n} 2\xi_i + \log(n)|T| + 2\gamma\left(\frac{p}{|T|}\right), \quad 0 \leq \gamma \leq 1,
\]

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

Table 2: Simulation results for SCAD penalized SVM

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

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

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

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 Founda-
tion (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

\[
P\left(c\sqrt{2A(\alpha)}\log p/n \leq c\|\widehat{S}(\beta^*)\|_\infty\right) \\
\leq \sum_{j=0}^{p} P\left(\sqrt{2A(\alpha)}\log p/n \leq n^{-1}\left|\sum_{i=1}^{n} I(1 - Y_iX_i^T\beta^* \geq 0)Y_iX_{ij}\right|\right).
\]

Notice that we have \(S(\beta^*) = 0\) because of minimizer \(\beta^*\) and the definition of gradient
vector. Then, for each \(i\) and \(j\), \(E(Y_iX_{ij}I(1 - Y_iX_i^T\beta^* \geq 0)) = 0\), by Hoeffding’s inequality,

\[
P\left(\sqrt{2A(\alpha)}\log p/n \leq n^{-1}\left|\sum_{i=1}^{n} I(1 - Y_iX_i^T\beta^* \geq 0)Y_iX_{ij}\right|\right)
\leq 2\exp\left(-\frac{4A(\alpha)n\log p}{4nM^2}\right) = 2p^{-\frac{A(\alpha)}{M^2}}.
\]

Therefore \(P\left(c\sqrt{2A(\alpha)}\log p/n \leq c\|\widehat{S}(\beta^*)\|_\infty\right) \leq (p + 1) \cdot 2p^{-\frac{A(\alpha)}{M^2}} \leq \alpha.\)

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

\[
\frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\widehat{\beta})_+ + \lambda\|\widehat{\beta}_-\|_1 \leq \frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\beta^*)_+ + \lambda\|\beta^*_+\|_1,
\]

\[
\frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\beta^* + Y_iX_i^Th)_+ - \frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\beta^*)_+ \leq \lambda\|\beta^*_+\|_1 - \lambda\|\widehat{\beta}_-\|_1.
\]

Recalling \(T = \{1 \leq j \leq p : \beta^*_j \neq 0\}\) and \(T_+ = T \cup \{0\}\), we have

\[
\|\beta^*_+\|_1 - \|\widehat{\beta}_-\|_1 \leq \|\beta^*_+\|_1 - \|\widehat{\beta}_-\|_1
\]

\[
\leq \|h_{T_+}\|_1 - \|h_{T^-}\|_1.
\]

This implies

\[
\frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\beta^* + Y_iX_i^Th)_+ - \frac{1}{n}\sum_{i=1}^{n}(1 - Y_iX_i^T\beta^*)_+ \leq \lambda(\|h_{T_+}\|_1 - \|h_{T^-}\|_1).
\]
Since the subdifferential of \( l_n(\beta) \) at the point of \( \beta^* \) is \( \hat{S}(\beta^*) \) and recall the assumption \( \lambda \geq c\|\hat{S}(\beta^*)\|_\infty \), we have

\[
\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \\
\geq \hat{S}^T(\beta^*)h \\
\geq -\|h\|_1 \cdot \|\hat{S}(\beta^*)\|_\infty \\
\geq -\frac{\lambda}{c}(\|h_{T^+}\|_1 + \|h_{T^-}\|_1).
\]

Hence, we have

\[
\lambda(\|h_{T^+}\|_1 - \|h_{T^-}\|_1) \geq -\frac{\lambda}{c}(\|h_{T^+}\|_1 + \|h_{T^-}\|_1),
\]

\[
\|h_{T^+}\|_1 \geq \tilde{C}\|h_{T^-}\|_1,
\]

where \( \tilde{C} = \frac{c-1}{c+1} \). We have thus proved that \( h \in \Delta_C \).

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

\[
\left| (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - (1 - Y_i X_i^T \beta^*)_+ \right| \leq \frac{1}{\|h\|_2} \|X_i^T h\|.
\]

By Hoeffding’s inequality, we have \( \forall t > 0 \),

\[
P\left( \frac{B(h)}{\|h\|_2} \geq \frac{t}{\sqrt{n}} \right) \leq 2 \exp\left( -\frac{2nt^2}{4\|Xh\|_2^2/\|h\|_2^2} \right).
\]

Hence by assumption (A3),

\[
P\left( \frac{B(h)}{\|h\|} \geq \frac{t}{\sqrt{n}} \right) \leq 2 \exp\left( -\frac{t^2}{2\lambda_{\max}} \right) \leq 2 \exp\left( -\frac{t^2}{2M_1} \right).
\]

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

\[
P\left( \frac{B(h)}{\|h\|} \geq C\sqrt{\frac{2q\log p}{n}} \right) \leq 2 \exp\left( -\frac{C^2q\log p}{M_1} \right) \leq 2p^{-C^2q/M_1} \leq 2p^{-C^2(q+1)/(2M_1)}.
\]

Next we will derive an upper bound for \( \sup_{\|h\|_0 \leq q+1} \frac{B(h)}{\|h\|} \). We consider covering \( \{h \in \mathbb{R}^{p+1}, \|h\|_0 \leq q + 1 \} \) with \( \epsilon \)-balls such that for any \( h_1 \) and \( h_2 \) in the same ball we have \( \|h_1\|_2 - \|h_2\|_2 \leq \epsilon \), where \( \epsilon \) is a small positive number. The number of \( \epsilon \)-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 \( h \) is a \((p+1)\)-dimensional vector with at most \( q+1 \) nonzero coordinates and \( h / \|h\|_2 \) has unit length in \( L_2 \) norm, the covering number we require is at most \((3p/\epsilon)^{q+1}\). Let \( N \) denote such an \( \epsilon \)-net. By the union bound,

\[
P\left( \sup_{h \in N} \frac{B(h)}{\|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 \left( \frac{3}{\epsilon} \right)^{p1-C^2/(2M_1)} \).
for any given positive constant $C$. Furthermore, for any $h_1, h_2 \in \mathbb{R}^{p+1}$ such that $||h_1||_0 \leq q + 1$, $||h_2||_0 \leq q + 1$, $||h_1||_2 \neq 0$ and $||h_2||_2 \neq 0$, we have

$$\left| \frac{B(h_1)}{||h_1||_2} - \frac{B(h_2)}{||h_2||_2} \right| \leq \frac{2}{n} \|\mathcal{X}(h_1/||h_1||_2 - h_2/||h_2||_2)\|_1$$

$$\leq \frac{2}{\sqrt{n}} \|\mathcal{X}(h_1/||h_1||_2 - h_2/||h_2||_2)\|_2$$

$$\leq 2\sqrt{M_1}\epsilon.$$

Therefore,

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

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

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

$$\leq P \left( \sup_{h \in N} \frac{B(h)}{||h||_2} \geq (C - 1) \sqrt{\frac{2q \log p}{n}} \right)$$

$$\leq 2 \left( \frac{2M_1 n}{q \log p} \right)^{\frac{q+1}{2}} (3p^{1-(C-1)^2/(2M_1)})^{q+1}$$

$$\leq 2 \left( 2\sqrt{M_1 n}3p^{1-(C-1)^2/(2M_1)} \right)^{q+1}.$$

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

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

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

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

**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 $i \in \{2, 3, \ldots, n - 1\}$

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

17
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

\[
g(k) \leq g \left( n \left( \frac{(x_1 + x_n)^2 - x_n^2}{x_1^2 - x_n^2} \right) \right) = \sqrt{n}(x_1 - x_n) \left( \frac{1}{2} - \frac{x_1 + 3x_n}{4(x_1 + x_n)} \right).
\]

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 \( h = \beta^* - \hat{\beta} \), then it follows from Lemma 2 that \( h \in \Delta_C \). 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

\[
\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h)_+ - \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j} h_{S_k})_+ - \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+
\]

\[
= \sum_{j \geq 1} \frac{1}{n} \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j} h_{S_k})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j-1} h_{S_k})_+ \right)
\]

\[
+ \frac{1}{n} \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h_{S_0})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \right),
\]

(16)
where the first equation follows from the definition of $h_{S_k}$, $k \geq 0$; 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 $j$th term in the sum of the above decomposition has the increment indexed by $h_{S_j}$, which has at most $q+1$ nonzero coordinates. Lemma 3 implies that uniformly for $j = 1, 2, \ldots$, with probability at least $1 - 2p^{-2q(q^2-1)}$,

$$\frac{1}{n} \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j} h_{S_k})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j-1} h_{S_k})_+ \right) \geq \frac{1}{n} E \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h_{S_k})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T \sum_{k=0}^{j-1} h_{S_k})_+ \right) - C \sqrt{\frac{2q \log p}{n}} \|h_{S_j}\|_2,$$

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

$$\frac{1}{n} \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h_{S_j})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \right) \geq M(h) - C \sqrt{\frac{2q \log p}{n}} \sum_{j \geq 0} \|h_{S_j}\|_2$$

(17)

where $M(h) = \frac{1}{n} E \left( \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^* + Y_i X_i^T h_{S_j})_+ - \sum_{i=1}^{n} (1 - Y_i X_i^T \beta^*)_+ \right)$.

It is straightforward to show that $\|h_{S_0}\|_1 \geq \|h_{T^C}||_1 \geq C ||h_{T^C}||_1 \geq \|h_{S_0}\|_1$. By Lemma 6, we have

$$\sum_{j \geq 1} \|h_{S_j}\|_2 \leq \sum_{j \geq 1} \frac{\|h_{S_j}\|_1}{\sqrt{q+1}} + \sqrt{\frac{q+1}{4}} |h_q| \leq \frac{\|h_{S_0}\|_1}{\sqrt{q+1}} + \frac{\|h_{S_0}\|_1}{4 \sqrt{q+1}} \leq \left( \frac{1}{\sqrt{q+1C}} + \frac{1}{4 \sqrt{q+1}} \right) \|h_{S_0}\|_1 \leq \left( \frac{1}{4} + \frac{1}{C} \right) \|h_{S_0}\|_2.$$

(18)

By the definition of $h$, (15), (17) and (18), we have

$$M(h) \leq \lambda (\|h_{T^+}\|_1 - \|h_{T^C}\|_1) + \left( \frac{1}{4} + \frac{1}{C} \right) C \sqrt{\frac{2q \log p}{n}} \|h_{S_0}\|_2 + C \sqrt{\frac{2q \log p}{n}} \|h_{S_0}\|_2 \leq \lambda \sqrt{q+1} \|h_{S_0}\|_2 + C \sqrt{\frac{2q \log p}{n}} \left( \frac{5}{4} + \frac{1}{C} \right) \|h_{S_0}\|_2.$$

(19)

Condition (A4) imply that

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

(20)
Combining (19) and (20), we have

\[
\frac{1}{2} M_2 \| h \|_2^2 + o(\| h \|_2^2) \leq \lambda \sqrt{q + 1} \| h_{S_0} \|_2 + C \sqrt{\frac{2q \log p}{n} \left( \frac{5}{4} + \frac{1}{C} \right)} \| h_{S_0} \|_2.
\]

Note that \( \| h \|_2 = \| h_{S_0} \|_2 + \sum_{j \geq 1} \| h_{S_j} \|_2 \geq \| h_{S_0} \|_2 \), and

\[
\sum_{j \geq 1} \| h_{S_j} \|_2^2 \leq |h_q| \sum_{j \geq 1} \| h_{S_j} \|_1 \leq \frac{1}{C} |h_q| \| h_{S_0} \|_1 \leq \frac{1}{C} \| h_{S_0} \|_2.
\]

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

\[
\| h_{S_0} \|_2 + o(\| h \|_2) \leq \frac{2 \lambda \sqrt{q + 1}}{M_2} + 2 C \frac{2q \log p}{n} \left( \frac{5}{4} + \frac{1}{C} \right).
\]

Hence,

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

We therefore have

\[
\| \hat{\beta} - \beta^* \|_2 \leq \sqrt{1 + \frac{1}{C} \left( \frac{2 \lambda \sqrt{q + 1}}{M_2} + 2 C \frac{2q \log p}{n} \left( \frac{5}{4} + \frac{1}{C} \right) \right)}
\]

with probability at least \( 1 - 2 p^{-2q(C^2 - 1)} \).

**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 \( H(\beta^*) \) has a positive lower bound. In the following, we provide a set of sufficient conditions to guarantee the positive definiteness of \( H(\beta^*) \).

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

\[
\int_S I(X_k \geq V_k^-) X_i g(X) dX < \int_S I(X_k \leq U_k^+) X_i f(X) dX
\]

or

\[
\int_S I(X_k \leq V_k^+) X_i g(X) dX > \int_S I(X_k \geq U_k^-) X_i f(X) dX
\]

Here \( U_k^+ \), \( V_k^+ \) \( \in \left[ -\infty, +\infty \right) \) are upper bounds such that \( \int_S I(X_k \leq U_k^+) f(X) dX = \min(1, \frac{\pi}{\pi_+}) \) and \( \int_S I(X_k \leq V_k^+) f(X) dX = \min(1, \frac{\pi}{\pi_+}) \). Similarly, lower bounds \( U_k^- \), \( V_k^- \) \( \in \left[ -\infty, +\infty \right] \) and are defined as \( \int_S I(X_k \geq U_k^-) f(X) dX = \min(1, \frac{\pi}{\pi_-}) \) and \( \int_S I(X_k \geq V_k^-) g(X) dX = \min(1, \frac{\pi}{\pi_-}) \).
For an orthogonal transformation $A_j$ that maps $\frac{\beta^*}{||\beta||}$ to the $j$-th unit vector $e_j$ for some $j \in \{1, 2, 3, \ldots, p\}$, there exists rectangles

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

and

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

such that $f(x) \geq B_1 > 0$ on $D^+$, and $g(x) \geq B_2 > 0$ on $D^-$, where $M^+ = \{ x \in \mathbb{R}^p | x^T \beta^*_1 + \beta_{01}^* = 1 \}$ and $M^- = \{ x \in \mathbb{R}^p | x^T \beta^*_1 + \beta_{02}^* = -1 \}$.

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 $S \subset \mathbb{R}^p$ 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 $\mathbb{R}^p$. The marginal distribution of $Y$ is given by $\pi_+ = \pi_-=1/2$. Let $f$ and $g$ be the density functions of $X_-$ given $Y=1$ and $-1$, respectively. Here, we assume $f$ and $g$ are multivariate normal densities with different mean vectors $\mu$ and $\nu$ and a common covariance matrix $\Sigma$. 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 $\beta^*$ and $H(\beta^*)$, 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 $\phi$ and $\Phi$, respectively. Then we have $S(\beta^*) = 0$, where $S(\cdot)$ is defined in (6), that is

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

and

$$E_f(I(1 - X^T \beta^* \geq 0)X_-) = E_g(I(1 + X^T \beta^* \geq 0)X_-) \tag{22}$$

For left hand of equation (21), we have $X^T \beta^*_1 \sim N(\mu^T \hat{\beta}^*, \beta^*_1 \Sigma \beta^*_1)$, thus

$$E_f(I(1 - X^T \beta^* \geq 0)) = P_f(1 - \beta_{02}^* - X^T \beta^*_1 \geq 0) = \Phi(c_f), \tag{23}$$

where $c_f = \frac{1-\beta_{02}^*-\mu^T \hat{\beta}^*}{||\hat{\beta}^*||_2}$. Similarly, $E_g(I(1 + X^T \beta^* \geq 0)) = \Phi(c_g)$. where $c_g = \frac{1+\beta_{02}^*+\nu^T \hat{\beta}^*}{||\hat{\beta}^*||_2}$. 

21
To obtain an analytic expression of \( E_f(I(1 - X^T\beta^* \geq 0)) \), we consider an orthogonal matrix \( P \) that satisfies \( P\Sigma^{1/2}\beta^*_{-}\) = \((1, 0, 0, \ldots, 0)^T\). Such a matrix \( P \) can always be constructed. Actually, let \( P = (P_1, P_2, \ldots, P_p)^T \) and \( P_1 = \Sigma^{1/2}\beta^*_{-}\). By using Gram-Schmidt process, we can generate other orthogonal vectors \( P_i \) based on \( P_1 \) with \( i = 2, 3, \ldots, p \). Since \( P\Sigma^{-1/2}(X_\mu - \mu) = Z \), a standard multivariate normal random vector, we have \( I - X^T\beta^* = c_f\Sigma^{1/2}\beta^*_{-}\). Thus

\[
E_f(I(1 - X^T\beta^* \geq 0)|X) = E\phi(I(c_f - Z_1 \geq 0)(\Sigma^{1/2}P^TZ + \mu))
\]

\[
= E\phi(I(c_f - Z_1 \geq 0)\mu) + E\phi(I(c_f - Z_1 \geq 0)\Sigma^{1/2}P^TZ).
\]

where \( \phi \) 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

\[
E_f(I(1 - X^T\beta^* \geq 0)X_k) = \mu_k\Phi(c_f) + E\phi(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^{p} P_i Z_i)
\]

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

\[
E\phi(I(c_f - Z_1 \geq 0)\Lambda_k^T\sum_{i=1}^{p} (P_i Z_i)) = E\phi(I(c_f - Z_1 \geq 0)\Lambda_k^T P_1 Z_1)
\]

\[
= \Lambda_k^T \Sigma^{1/2}\beta^*_{-}\Lambda E\phi(I(c_f - Z_1 \geq 0)Z_1)
\]

\[
= \Lambda_k^T \Sigma^{1/2}\beta^*_{-}\Lambda \int_{-\infty}^{+\infty} I(c_f - x \geq 0)\phi(x)dx
\]

\[
= \Lambda_k^T \Sigma^{1/2}\beta^*_{-}\Lambda \int_{-\infty}^{c_f} \phi(x)dx
\]

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

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

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

\[
E_f(I(1 - X^T\beta^* \geq 0)X_\mu) = \mu\Phi(c_f) - \phi(c_f)\Sigma^{1/2}P_1.
\]

Similarly,

\[
E_g(I(1 + X^T\beta^* \geq 0)X_\nu) = \nu\Phi(c_g) + \phi(c_g)\Sigma^{1/2}P_1
\]
Then, we have
\[
\Phi(c_f) = \Phi(c_g) 
\] (24)
and
\[
\mu \Phi(c_f) - \phi(c_f) \Sigma^{1/2} P_1 = \nu \Phi(c_g) + \phi(c_g) \Sigma^{1/2} P_1
\] (25)
From (24), we have \( \tilde{c} = c_f = c_g \), which implies
\[
\beta^T_0 (\mu + \nu) = -2 \beta^*_0
\] (26)
From (25),
\[
\beta^*_0 \frac{\| \Sigma^{1/2} \beta^* \|_2}{2 \phi(\tilde{c})} = \Phi(\tilde{c}) \Sigma^{-1} (\mu - \nu)
\] (27)
Let \( d_\Sigma(\mu, \nu) = ((\mu - \nu)^T \Sigma^{-1} (\mu - \nu))^{1/2} \) be the Mahalanobis distance between \( \mu \) and \( \nu \) and \( R(x) = \frac{\phi(x)}{\Phi(x)} \). As \( \Sigma^{1/2} \beta^* \frac{\| \Sigma^{1/2} \beta^* \|_2}{2 \phi(\tilde{c})} \) has \( l_2 \) norm equal to 1, we have \( \| \Phi(\tilde{c}) \Sigma^{-1/2} (\mu - \nu) \|_2 = 1 \), i.e., \( R(\tilde{c}) = \frac{d_\Sigma(\mu, \nu)}{2} \). \( R(x) \) is a monotonically decreasing function, thus we have \( \tilde{c} = R^{-1} \left( \frac{d_\Sigma(\mu, \nu)}{2} \right) \). Meanwhile, \( \tilde{c} = c_f = \frac{1 - \beta^*_0 - \mu^T \beta^*}{\| \Sigma^{1/2} \beta^* \|_2} \), we can solve the problem based on (26) and (27),
\[
\beta^*_0 = \frac{\| \mu - \nu \|}{2 \tilde{c} d_\Sigma(\mu, \nu) + d^2_\Sigma(\mu, \nu)} (\mu - \nu)^T \Sigma^{-1} (\mu + \nu)
\] (28)
From (25),
\[
\beta^*_0 = \frac{2 \Sigma^{-1} (\mu - \nu)}{2 \tilde{c} d_\Sigma(\mu, \nu) + d^2_\Sigma(\mu, \nu)}
\] (29)
By plugging (28) and (29) into (7), we can calculate \( H(\beta^*) \) as
\[
H(\beta^*) = \frac{\phi(\tilde{c})}{4} (2 \tilde{c} + d_\Sigma(\mu, \nu)) \left( \frac{2}{\mu + \nu} \begin{pmatrix} \mu + \nu \\ H_{22}(\beta^*) \end{pmatrix} \right)
\] (30)
where
\[
H_{22}(\beta^*) = \mu \mu^T + \nu \nu^T + 2 \Sigma + 2 \left( \frac{\tilde{c}}{d_\Sigma(\mu, \nu)} \right)^2 + \tilde{c} \frac{1}{d^2_\Sigma(\mu, \nu)} \right) (\mu - \nu)(\mu - \nu)^T
\]
As we have obtained the analytic form of \( H(\beta^*) \), we consider Model 1 in Section 5.1 as an example. In Model 1, \( q = 5 \), \( \mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^T \) and \( \nu = (-0.1, -0.2, -0.3, -0.4, -0.5, 0, \ldots, 0)^T \) and \( \pi^+ = \pi^- = 1/2 \). The covariance matrix \( \Sigma = (\sigma_{ij}) \) 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 \( \beta^* = (0, 1.39, 1.47, 1.56, 1.65, 1.74, 0, \ldots, 0)^T \). Based on (30), we derived \( H(\beta^*) \) and numerically validated its positive-definiteness.
References


