Journal of Machine Learning Research 18 (2018) 1-26 Submitted 6/17; Revised 1/18; Published 4/18

Divide-and-Conquer for Debiased l_1-norm Support Vector Machine in Ultra-high Dimensions

Heng Lian HENGLIAN@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Kowloon Tong, Hong Kong

Zengyan Fan ZENGYANFAN@HOTMAIL.COM
Department of Statistics and Applied Probability
National University of Singapore
Singapore

Editor: Jie Peng

Abstract

1-norm support vector machine (SVM) generally has competitive performance compared to standard 2-norm support vector machine in classification problems, with the advantage of automatically selecting relevant features. We propose a divide-and-conquer approach in the large sample size and high-dimensional setting by splitting the data set across multiple machines, and then averaging the *debiased* estimators. Extension of existing theoretical studies to SVM is challenging in estimation of the inverse Hessian matrix that requires approximating the Dirac delta function via smoothing. We show that under appropriate conditions the aggregated estimator can obtain the same convergence rate as the central estimator utilizing all observations.

Keywords: classification, debiased estimator, distributed estimator, divide and conquer, sparsity

1. Introduction

The support vector machine (SVM) is a widely used tool for classification (Vapnik, 2013; Scholkopf and Smola, 2001; Cristianini and Shawe-Taylor, 2000). Although the original motivation of Cortes and Vapnik (1995) is in terms of finding a maximum-margin hyperplane, its equivalent formulation as a regularized functional optimization problem is perhaps more easily understood by statisticians and more amenable for statistical asymptotic analysis. In the standard formulation the penalized functional is a sum of the hinge loss plus an l_2-norm regularization term. Statistical properties of the SVM, especially its nonlinear version using general kernels, has been studied in a lot of works recently including but not limited to Bartlett et al. (2006); Blanchard et al. (2008); Lin (2000, 2004); Steinwart and Scovel (2007); Steinwart (2005); Zhang (2004). In this work, we focus on penalized linear SVM with large sample size and large dimension, with particular emphasis on dealing with distributed estimation in such contexts.

Data sets with thousands of features have become increasingly common recently in many real-world applications. For example, a microarray data set typically contains more than

10,000 genes. A drawback of standard SVM based on l_2-norm penalty is that it can be adversely affected if many redundant variables are included in building the decision rule. A modern approach to feature selection is based on the idea of shrinkage. This approach involves fitting a model involving all p predictors. However, the estimated coefficients are shrunk towards zero. In particular with appropriate choice of penalty some of the coefficients may be estimated to be exactly zero. Automatic variable selection using penalized estimation that can shrink some coefficients to be exactly zero was pioneered in Tibshirani (1996) using an l_1-norm penalty (or called lasso penalty). Other penalties proposed include those in Fan and Li (2001), Zou (2006) and Zhang (2010).

The idea of using l_1 norm to automatically select variables has been extended to classification problems. van de Geer (2008) analyzed lasso penalized estimator for generalized linear models which include logistic regression as a special case. Zhu et al. (2003) proposed the l_1-norm support vector machine and oracle properties of SCAD-penalized support vector machines were established in Park et al. (2012), based on the Bahadur representation of Koo et al. (2008). See also the earlier work of Bradley and Mangasarjan (1998); Song et al. (2002). When feature dimension is larger than the sample size, Peng et al. (2016); Zhang et al. (2016) obtained the convergence rate of the SCAD and lasso-penalized estimators for SVM, respectively. Our work follows the lead of these works on understanding the statistical properties of the estimated SVM coefficients, instead of on generalization error rates or empirical risk.

In this paper, we focus on distributed estimation of l_1 penalized linear SVM coefficients using multiple computing machines. The simplest and most popular approach in data parallelism is averaging: each machine uses a part of the data and obtains a local estimator using the standard estimation methods and sends it back to the master machine which combines the local estimators by simple averaging into an aggregated estimator. In the classical regime concerning fixed dimensional problems, this has been advocated in McDonald et al. (2009), and was also studied by Zinkevich et al. (2010); Zhang et al. (2013, 2015); Balcan et al. (2015); Zhao et al. (2016). In all these studies, the typical outcome of asymptotic analysis is that under suitable assumptions, in particular that the number of machines are not excessive compared to the sample size, the aggregated estimator enjoys the same or similar statistical properties as the centralized estimator obtained by a single machine using all observations (if the centralized estimator can be feasibly obtained). Such results convincingly illustrate that the divide-and-conquer strategy works in the big data world. In the high dimensional regime, for lasso penalized estimators, there is a well-known bias-variance trade-off. When the tuning parameter in the penalty is chosen optimally in each local machine, the size of bias and standard deviation are of the same order. Aggregation can decrease the variance thanks to the magic of central limit theorem or some related finite-sample bounds for averages of mean zero random variables, but it cannot decrease the bias in general. Thus debiasing becomes crucial to reduce the bias to a smaller order before aggregation. This is done for sparse linear regression in Lee et al. (2017), which shows the debiased estimator in van de Geer et al. (2014) works satisfactorily for parameter inferences. Lee et al. (2017) studied both the least squares estimator and the more general M-estimator with *smooth* loss functions using an l_1 (LASSO) penalty and applied it to distributed estimation. Our study of linear SVM coefficients differs significantly from Lee et al. (2017) due to the unsmooth nature of the hinge loss function. In particular, estimation of the

Hessian matrix which involves Dirac delta function requires a smoothing procedure, which is nontrivial to analyze.

The rest of the paper is organized as follow. After a brief introduction of some notations below, we consider debiased l_1-norm SVM in Section 2.1. Although the main focus is on distributed estimation, we need to first consider statistical properties of debiased estimator on a single machine, which requires a lengthy and detailed analysis. Once this is done, properties of the aggregated estimator are relatively easy to establish, as is done in Section 2.2. In terms of l_∞ norm, the aggregated estimator has the convergence rate $O_p(\sqrt{\log p/N})$ when the number of features is p and the total sample size is N under appropriate assumptions. However, its convergence rate in l_1 or l_2 norm is unacceptably larger, which motivated a further thresholding step in Section 2.3. Section 3 report some numerical results to demonstrate the finite sample performance of the proposed estimators. Finally, we conclude this paper with a discussion in Section 4.

Notations. For a vector $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$, $\|\mathbf{a}\|_\infty = \max_j |a_j|$, $\|\mathbf{a}\|_1 = \sum_j |a_j|$, $\|\mathbf{a}\| = (\sum_j a_j^2)^{1/2}$ and $\|\mathbf{a}\|_0$ is the number of nonzero components of $\mathbf{a}$. For a matrix $\mathbf{A} = \{a_{ij}\}_{i,j=1}^n$, $\|\mathbf{A}\|_\infty = \max_{i,j} |a_{ij}|$, $\|\mathbf{A}\|_1 = \sum_{i,j} |a_{ij}|$ and $\|\mathbf{A}\|_{L_1} = \max_i \sum_j |a_{ij}|$. Throughout the paper, C denotes a generic constant that may assume different values even on the same line.

2. Divide-and-conquer for l_1-SVM

2.1 Debiased l_1-SVM

We begin with the basic setup of SVM for binary classification. We observe a simple random sample $(\mathbf{x}_i, y_i), i = 1, \ldots, N$, from an unknown distribution $P(\mathbf{x}, y)$. Here $y_i \in \{-1, 1\}$ is the class label and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the p-dimensional features. For simplicity of presentation, we do not use any special treatment for the intercept, although the intercept term is typically not shrunk in l_1-SVM. The standard linear SVM estimates the parameters by solving

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} N^{-1} \sum_{i=1}^N L(y_i, \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2,$$

where L is the hinge loss function $L(y, t) = \max\{0, 1 - yt\}$ and λ is the regularization parameter which changes with N (typically converging to zero as N goes to infinity), but we suppress its dependence on N in our notation. Throughout the paper we make the mild assumption that

$$\log N = O(\log p).$$

This does not mean $p \geq N$, but exclude the case that p is fixed. This restriction is mainly to make the notation slightly simpler. Without this restriction, Theorem 1 below still hold with $\log p$ replaced by $\log(\max\{p, N\})$, and the probability $1 - p^{-C}$ replaced by $1 - (\max\{p, N\})^{-C}$.

Variable selection is of particular interest when p is large compared to N, due to its ability to avoid overfitting as well as to enhance interpretation. The l_1-SVM (Zhu et al., 2003) estimates the parameter by solving

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} N^{-1} \sum_{i=1}^N L(y_i, \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1. \tag{1}$$

The l_1 penalty here encourages sparsity of the solution (Tibshirani, 1996, 1997).

Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$ be the true parameter, which is defined as the minimizer of the population hinge loss,

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} E[L(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})]. \tag{2}$$

We assume $\boldsymbol{\beta}_0$ exists and is unique. Koo et al. (2008) provided some regularity conditions under which $\boldsymbol{\beta}_0$ is unique and $\boldsymbol{\beta}_0 \neq \mathbf{0}$. Towards variable selection in SVM, it is natural to assume $\boldsymbol{\beta}_0$ is sparse. Let $A = \{1 \leq j \leq p : \beta_{0j} \neq 0\}$ be the support set of $\boldsymbol{\beta}_0$ with $s = |A|$ the cardinality of A.

As calculated rigorously in Koo et al. (2008), the gradient vector and the Hessian matrix of the population hinge loss in Equation 2 is given by

$$S(\boldsymbol{\beta}) \quad = \quad -E[I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \leq 1\}\mathbf{x}y]$$

and

$$\mathbf{H}(\boldsymbol{\beta}) = E[\delta(1 - y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\mathbf{x}^{\mathrm{T}}],$$

respectively, where $\delta(.)$ is the Dirac delta function. Let f and g be the conditional density of $\mathbf{x}$ given $y = 1$ and $y = -1$, respectively.

(A1) The densities f and g are bounded and continuously differentiable with bounded partial derivatives, with compact support. x_j's are bounded random variables. Without loss of generality, we assume the distribution of $\mathbf{x}$ has a support contained in $[0, 1]^p$.

Under assumption (A1), $\mathbf{H}(\boldsymbol{\beta})$ is well-defined and continuous in $\boldsymbol{\beta}$.

Due to the penalty term, the penalized estimator is generally biased (i.e. shrunk towards zero). Conceptually, λ controls the trade-off between bias and standard deviation of the estimator. While averaging will reduce the standard deviation of the estimator, it generally cannot reduce the bias. Thus it is important to apply a debiasing mechanism before we aggregate estimators from different machines. For simplicity of presentation, for now we focus on the properties of the debiased estimator using all observations and later we will argue (almost trivially) these properties hold for the local estimates, uniformly over M machines. In this subsection, with $M = 1$, we have $N = n$ where n is the sample size on a single machine.

Let $\widehat{\boldsymbol{\beta}}$ be the penalized estimator obtained from Equation 1. It is known that $\widehat{\boldsymbol{\beta}}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) + \lambda\kappa = 0 \tag{3}$$

where $L_t(y, t)$ is a sub-derivative of $L(y, t)$ with respect to t, and $\kappa = (\kappa_1, \ldots, \kappa_p)^{\mathrm{T}}$ with $\kappa_j = \mathrm{sign}(\widehat{\beta}_j)$ if $\widehat{\beta}_j \neq 0$ and $\kappa_j \in [-1, 1]$ if $\widehat{\beta}_j = 0$.

When the loss is twice differentiable, a simple Taylor's expansion can be used to expand $L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$ at $\boldsymbol{\beta}_0$ as in van de Geer et al. (2014). For the nonsmooth loss function here, we need to use empirical processes techniques. Let $G_n = \sqrt{n}(P_n - P)$ be the empirical process, where P is the population distribution of $(\mathbf{x}, y)$ and P_n is the empirical distribution of the observations. Informally, when $\widehat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}_0$, $G_n(\mathbf{x}\{L_t(y, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - L_t(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\})$ is small, and thus

$$\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) \\
\approx\ & \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + E\mathbf{x}L_t(y, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - E\mathbf{x}L_t(y, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \\
\approx\ & \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \qquad (4)
\end{aligned}$$

Then

$$\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) + \lambda\kappa \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \lambda\kappa \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}),
\end{aligned}$$

where we used Equation 3 in both the first and the last inequality, and this leads to

$$\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0 + [\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}) - [\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0),$$

if $\mathbf{H}(\boldsymbol{\beta}_0)$ is invertible. Since the last term in the right hand side above has mean zero, we are motivated to define the debiased estimator as

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - [\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i L_t(y_i, \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}), \qquad (5)$$

and we set $L_t(y, t) = -yI\{yt \le 1\}$. However, $\mathbf{H}(\boldsymbol{\beta}_0)$ is unknown in two aspects. On one hand, the true parameter $\boldsymbol{\beta}_0$ is unknown and should be replaced by its estimator, say $\widehat{\boldsymbol{\beta}}$. On the other hand, $\mathbf{H}(\boldsymbol{\beta}) = E[\delta(1 - y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ is an expectation which should be approximated by samples. Although expectations are usually easily estimated by a simple moment estimator, this is not the case here, since there may not even be a single sample that satisfies exactly $y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = 1$ and $\delta(.)$ as a generalized function should be treated carefully. Finally, after $\mathbf{H}(\boldsymbol{\beta})$ is approximated by samples, high-dimensionality means that the usual algebraic inverse of the estimator may not be well-defined and some approximate inverse must be used.

Thus it seems one major component of the estimation is the approximation of $[\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}$ and we deal with this problem first. To motivate an estimator of $\mathbf{H}(\boldsymbol{\beta})$, we can start from its antiderivative $S(\boldsymbol{\beta}) = E[-I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \leq 1\}\mathbf{x}y]$. For any given $\boldsymbol{\beta}$, $S(\boldsymbol{\beta})$ can be approximated by $-N^{-1}\sum_{i=1}^{N} I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \leq 1\}\mathbf{x}_i\mathbf{y}_i$. If this were differentiable, we could use its derivative as an estimator of $\mathbf{H}(\boldsymbol{\beta})$. This observation motivates us to smooth the indicator function using some cumulative distribution function, say Q, and approximates $S(\boldsymbol{\beta})$ by $-N^{-1}\sum_{i=1}^{N} Q((1 - y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/h)\mathbf{x}_iy_i$. When the bandwidth parameter h is sufficiently small, $Q(./h)$ will approximate $I\{. \geq 0\}$ well. Assuming Q is differentiable, then it is natural to approximate $\mathbf{H}(\boldsymbol{\beta})$ by

$$\widehat{\mathbf{H}}(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{N}(1/h)q((1 - y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})/h)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},$$

where $q(.)$ is the density of the distribution Q (derivative of Q), and thus $\mathbf{H}(\boldsymbol{\beta}_0)$ is estimated by $\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}}$ is the l_1-SVM estimator.

Since the rank of $\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})$ is at most N, $\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})$ is singular when p is larger than N. Even when p is smaller than N and $\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})$ is non-singular, the standard inverse $[\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})]^{-1}$ is often not a good estimator of $\mathbf{H}(\boldsymbol{\beta}_0)$ when p is diverging with N. An approximate inverse of $\mathbf{H}(\boldsymbol{\beta}_0)$ can be found via an approach similar to that used in Cai et al. (2011) as

$$\begin{aligned} &\widehat{\boldsymbol{\Theta}} = \arg\min \|\boldsymbol{\Theta}\|_1 \\ &\text{subject to } \|\boldsymbol{\Theta}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{I}\|_\infty \leq Cb_N, \end{aligned} \tag{6}$$

for some tuning parameter $b_N \to 0$. Note that Cai et al. (2011) would use a slightly different constraint on $\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Theta} - \mathbf{I}\|_\infty$, while our constraint is more convenient in the current context since we *pre-multiply* the gradient of the loss by $[\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}$ in Equation 5. We note that the constraint $\|\boldsymbol{\Theta}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{I}\|_\infty \leq Cb_N$ is obviously the same as $\|\boldsymbol{\Theta}_{j.}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{e}_j^{\mathrm{T}}\|_\infty \leq Cb_N, \forall j$, where $\boldsymbol{\Theta}_{j.}$ is the j-th row of $\boldsymbol{\Theta}$ (as a row vector) and $\mathbf{e}_j$ is the unit vector with j-th componenet 1. Thus the optimization problem can be solved row by row. This actually was noted in Cai et al. (2011) in their Lemma 1 (since their constraint is $\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})\boldsymbol{\Theta} - \mathbf{I}\|_\infty \leq Cb_N$, their problem can be solved column by column).

Before proceeding, we impose some additional assumptions.

(A2) $\boldsymbol{\beta}_0 \neq \mathbf{0}$ and without loss of generality we assume $\beta_{01} = \max_{1\leq j\leq p}|\beta_{0j}|$.

(A3) $\|\boldsymbol{\Theta}_0\|_{L_1} \leq C_N$, where $\boldsymbol{\Theta}_0 = [\mathbf{H}(\boldsymbol{\beta}_0)]^{-1}$.

(A4) $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq Cs\sqrt{\frac{\log p}{N}}$ with probability at least $1 - p^{-C}$, where $s = |\mathrm{supp}\{\boldsymbol{\beta}_0\}|$ is the number of nonzero entries in $\boldsymbol{\beta}_0$.

(A5) The density q is an even function, twice continuously differentiable, with $\int x^2q(x)dx < \infty$, $\int q^2(x)dx < \infty$, $\int(q')^2(x)dx < \infty$, $\sup_x q'(x) < \infty$, $\sup_x q''(x) < \infty$, where q' and q'' are the first two derivatives of q.

(A6) $\|\widehat{\boldsymbol{\beta}}\|_0 \leq K$ with probability at least $1 - p^{-C}$.

Assumption (A2) is mild. Koo et al. (2008) gives sufficient conditions that guarantee $\boldsymbol{\beta}_0 \neq \mathbf{0}$. To simplify the bounds below one can think of β_{01} as bounded away from zero so that it will disappear from the bounds. Note that β_{01} is the largest nonzero coefficient. In the literature of sparse regression, it is often assumed that the *smallest* nonzero coefficient is large enough so that it can be distinguished from zero coefficients, which is a totally different assumption. Cai et al. (2011) assumed that their inverse Hessian matrix has a bounded L_1 norm (such an assumption is obviously related to sparsity of the matrix), which motivated our assumption (A3). We allow C_N in (A3) to be diverging for slightly greater generality. In particular, this mean we need to have a control on the l_1 norm of the rows of the inverse Hessian matrix. Due to that l_1 norm is a convex relaxation of the l_0 norm, we call such matrix as approximately sparse. Again, it is probably easier for the reader to regard C_N as a fixed constant. We further discuss (A3) in detail in Appendix B. Theorem 4 of Peng et al. (2016) showed $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| = O_p(\sqrt{s\log p/N})$ which together with their Lemma 2 implies $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 = O_p(s\sqrt{\log p/N})$ as in (A4). It is also easy to choose a density that satisfies (A5), such as the standard normal density, which will be used in our numerical studies. In (A6), we assume the estimator is sufficiently sparse. This can be guaranteed in several different ways. First, we conjecture it could be proved that $\|\widehat{\boldsymbol{\beta}}\|_0 = O_p(s)$ for SVM coefficient using a similar strategy as for Theorem 3 of Belloni and Chernozhukov (2011), although the details looks quite lengthy. Second, one could add a thresholding step similar to what we will use in subsection 2.3 later to get a sparse estimator. Finally, we could add an constraint $\|\boldsymbol{\beta}\| \le K$ to the lasso problem. Such a constrained penalized problem was also proposed in Fan and Lv (2013); Zheng et al. (2014). In any case, one could expect that K is of the same order as s, the sparsity of $\boldsymbol{\beta}_0$.

We first state several propositions whose proof is left to Appendix A. The first proposition considers the accuracy bound of $\widehat{\boldsymbol{\Theta}}$ as an approximation of the inverse of $\mathbf{H}(\boldsymbol{\beta}_0)$. The second proposition shows that the first approximation in Equation 4 is sufficiently accurate based on the empirical processes results. Finally, the third proposition establishes a Lipschitz property of the Hessian matrix which implies that the second approximation in Equation 4 is sufficiently accurate.

Proposition 1 *Under assumptions (A1)-(A5), with probability at least* $1-p^{-C}$,

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{I}\|_\infty \le Cb_N,$$
$$\|\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0) - \mathbf{I}\|_\infty \le Cb_N,$$
and
$$\|\widehat{\boldsymbol{\Theta}}\|_{L_1} \le C_N,$$

when we set $b_N = C_N\left(\left(\frac{1}{\beta_{01}} + \sqrt{\frac{\log p}{Nh^3\beta_{01}}} + \frac{\log p}{Nh^2}\right)s\sqrt{\frac{\log p}{N}} + \frac{s^2\log p}{Nh^3} + \frac{h}{\beta_{01}^2} + \sqrt{\frac{\log p}{Nh\beta_{01}}}\right)$.

Proposition 2 *Under assumptions (A1)-(A6), with probability at least* $1-p^{-C}$,

$$\left\|\frac{1}{N}\sum_i y_i\mathbf{x}_i(I\{y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \le 1\} - I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \le 1\}) - Ey\mathbf{x}(I\{y\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \le 1\} - I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \le 1\})\right\|_\infty$$
$$\le \quad Ca_N,$$

where $a_N = \left(\left(\frac{s}{\beta_{01}}\sqrt{\frac{\log p}{N}}\right)^{1/2}\sqrt{\frac{K\log p}{N}} + \frac{K\log p}{N}\right)$.

Proposition 3 *(Local Lipschitz property of $\mathbf{H}(\boldsymbol{\beta})$) Under assumptions (A1) and (A2) and in addition $2\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1 \le \beta_{01} := \max_j |\beta_{0j}|$, we have*

$$\|\mathbf{H}(\boldsymbol{\beta})-\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty \le \frac{C}{\beta_{01}^3}\|\boldsymbol{\beta}_0\|_1\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1.$$

Now we derive a finite-sample bound for $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty$.

Theorem 1 *Under assumptions (A1)-(A6) and that $s\sqrt{\log p/N} = o(\beta_{01})$, we have*

$$\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty \le C\left(C_N\left(a_N+\sqrt{\frac{\log p}{N}}+\frac{\|\boldsymbol{\beta}_0\|_1}{\beta_{01}^3}\frac{s^2\log p}{N}\right)+b_N s\sqrt{\frac{\log p}{N}}\right)$$

with probability at least $1-p^{-C}$.

Proof of Theorem 1. We have

$$\begin{aligned}
& \widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \\
= \; & \widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0+\widehat{\boldsymbol{\Theta}}\left\{\frac{1}{N}\sum_i y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}\right\} \\
= \; & (\mathbf{I}-\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\widehat{\boldsymbol{\Theta}}\left\{\frac{1}{N}\sum_i y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}\right\} \\
= \; & (\mathbf{I}-\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
& +\widehat{\boldsymbol{\Theta}}\Big\{\frac{1}{N}\sum_i y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\}+E[y\mathbf{x}I\{y\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}]-E[y\mathbf{x}I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\}] \\
& \qquad +\mathbf{a}_N+\mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\Big\}, \qquad (7)
\end{aligned}$$

where $\mathbf{a}_N = \frac{1}{N}\sum_i y_i\mathbf{x}_i(I\{y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}-I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\})-Ey\mathbf{x}(I\{y\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}-I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\})$ with $\|\mathbf{a}_N\|_\infty \le a_N$ with probability $1-p^{-C}$.

Using Proposition 1, the first term above is bounded by $\|\mathbf{I}-\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \le b_N s\sqrt{\log p/N}$ with probability at least $1-p^{-C}$. We also have

$$\begin{aligned}
& \left\|E[y\mathbf{x}I\{y\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\le 1\}]-E[y\mathbf{x}I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\}]+\mathbf{H}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right\|_\infty \\
= \; & \|(\mathbf{H}(\boldsymbol{\beta}^*)-\mathbf{H}(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_\infty \\
\le \; & \|\mathbf{H}(\boldsymbol{\beta}^*)-\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1 \\
\le \; & \frac{C}{\beta_{01}^3}\|\boldsymbol{\beta}_0\|_1\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1^2,
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. Furthermore, using Hoeffding's inequality and the union bound,

$$P\left(\left\|\frac{1}{N}\sum_i y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0\le 1\}\right\|_\infty > t\right) \le 2p\exp\{-CNt^2\}.$$

and thus

$$\left\| \frac{1}{N} \sum_i y_i x_i I\{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 \leq 1\} \right\|_\infty \leq C\sqrt{\frac{\log p}{N}},$$

with probability at least $1 - p^{-C}$.

Finally, combining the various bounds above and using that for any vector $\mathbf{a}$, $\|\widehat{\boldsymbol{\Theta}}\mathbf{a}\|_\infty \leq \|\widehat{\boldsymbol{\Theta}}\|_{L_1}\|\mathbf{a}\|_\infty$, we can get that the second term of Equation 7 is bounded by a constant multiple of $C_n\left(a_N + \sqrt{\frac{\log p}{N}} + \frac{\|\boldsymbol{\beta}_0\|_1}{\beta_{01}^3}\frac{s^2 \log p}{N}\right)$ with probability at least $1 - p^{-C}$. ■

In l_∞ norm, based on Theorem 1, we can see that under reasonable assumptions (see for example corollary 1) the debiased estimator has the convergence rate $\sqrt{\log p/N}$, which is the dominating term in the bound. However, in terms of l_1 or l_2 norm, since $\widetilde{\boldsymbol{\beta}}$ is non-sparse, we generally have $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_p(p\sqrt{\log p/N})$ and $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{p \log p/N})$, which is much larger than the bounds for the centralized estimator $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_p(s\sqrt{\log p/N})$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{s \log p/N})$ (Peng et al., 2016), where s is the number of nonzero components in $\boldsymbol{\beta}_0$. Post-processing using thresholding can be used to address this problem, which we will consider after we discuss distributed estimation next.

2.2 Distributed estimation

We now consider distributed estimation, in which the whole data set is evenly distributed to M machines, with M possibly diverging with N. The size of the data at each local machine is $n = N/M$, assumed to be an integer for simplicity. On each machine m, $1 \leq m \leq M$, we use the local data to obtain $\widehat{\boldsymbol{\beta}}^{(m)}$, $\widehat{\boldsymbol{\Theta}}^{(m)}$, and the debiased estimator $\widetilde{\boldsymbol{\beta}}^{(m)}$. Finally, the aggregated estimator is defined by

$$\bar{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M} \widetilde{\boldsymbol{\beta}}^{(m)}.$$

Theorem 2 *Under assumptions (A1)-(A6) (with N replaced by n, and in (A4) and (A6) $\widehat{\boldsymbol{\beta}}$ replaced by $\widehat{\boldsymbol{\beta}}^{(m)}$, $m = 1, \ldots, M$), and that $s\sqrt{\log p/n} = o(\beta_{01})$, we have*

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leq C\left(C_n\left(a_n + \sqrt{\frac{\log p}{N}} + \frac{\|\boldsymbol{\beta}_0\|_1}{\beta_{01}^3}\frac{s^2 \log p}{n}\right) + b_n s\sqrt{\frac{\log p}{n}}\right),$$

with probability at least $1 - p^{-C}$, where $a_n = \left(\left(\frac{s}{\beta_{01}}\sqrt{\frac{\log p}{n}}\right)^{1/2}\sqrt{\frac{K \log p}{n}} + \frac{K \log p}{n}\right)$ and $b_n = C_n\left(\left(\frac{1}{\beta_{01}} + \sqrt{\frac{\log p}{nh^3\beta_{01}}} + \frac{\log p}{nh^2}\right)s\sqrt{\frac{\log p}{n}} + \frac{s^2 \log p}{nh^3} + \frac{h}{\beta_{01}^2} + \sqrt{\frac{\log p}{nh\beta_{01}}}\right)$

Proof of Theorem 2. Let $D_m \subset \{1, \ldots, N\}$ with cardinality $|D_m| = n$ be the indices of the sub-data-set distributed to machine m. We have

$$\begin{aligned} & \bar{\boldsymbol\beta} - \boldsymbol\beta_0 \\ = \; & \frac{1}{M}\sum_{m=1}^{M}(\mathbf{I} - \widehat{\boldsymbol\Theta}^{(m)}\mathbf{H}(\boldsymbol\beta_0))(\widehat{\boldsymbol\beta}^{(m)} - \boldsymbol\beta_0) \\ & + \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol\Theta}^{(m)}\left\{\frac{1}{n}\sum_{i\in D_m} y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm T}\boldsymbol\beta \le 1\}\right\} \\ & - \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol\Theta}^{(m)}\left\{E\left[y\mathbf{x}\left(I\{y\mathbf{x}^{\mathrm T}\widehat{\boldsymbol\beta}^{(m)} \le 1\} - I\{y\mathbf{x}^{\mathrm T}\boldsymbol\beta_0 \le 1\}\right)\right] + \mathbf{H}(\boldsymbol\beta_0)(\widehat{\boldsymbol\beta}^{(m)} - \boldsymbol\beta_0)\right\} \\ & - \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol\Theta}^{(m)}\mathbf{a}_n^{(m)}, \end{aligned}$$

where $\mathbf{a}_n^{(m)} = \frac{1}{n}\sum_{i\in D_m} y_i\mathbf{x}_i(I\{y_i\mathbf{x}_i^{\mathrm T}\widehat{\boldsymbol\beta} \le 1\} - I\{y_i\mathbf{x}_i^{\mathrm T}\boldsymbol\beta_0 \le 1\}) - Ey\mathbf{x}(I\{y\mathbf{x}^{\mathrm T}\widehat{\boldsymbol\beta} \le 1\} - I\{y\mathbf{x}^{\mathrm T}\boldsymbol\beta_0 \le 1\})$ with $\|\mathbf{a}_n\|_\infty \le a_n$ with probability at least $1 - p^{-C}$.

For terms other than the second one above, the proof is exactly the same as for the proof of Theorem 1. Note that for each machine m, all derived inequalities hold with probability at least $1 - p^{-C}$ (note C can be chosen to be arbitrarily large) and thus they hold with probability at least $1 - Mp^{-C}$ simultaneously for all M machines. Since we assumed $\log M \le \log N = O(\log p)$, $1 - Mp^{-C}$ can again be written as $1 - p^{-C}$ (with a different C). For the second term above, the difference from the calculations in Theorem 1 is that here $\widehat{\boldsymbol\Theta}^{(m)}$ is different for different m. Let $\mathbf{e}_j \in \mathbb{R}^p$ be the unit vector with a single one for the j-th entry and let $a_{ij} = y_i\mathbf{e}_j^{\mathrm T}\widehat{\boldsymbol\Theta}^{(m)}\mathbf{x}_i$ if $i \in D_m$. Note $|a_{ij}| = |\mathbf{e}_j^{\mathrm T}\widehat{\boldsymbol\Theta}^{(m)}\mathbf{x}_i| \le \|\mathbf{e}_j\|_\infty\|\widehat{\boldsymbol\Theta}^{(m)}\|_{L_1}\|\mathbf{x}_i\|_\infty \le CC_n$ with probability at least $1 - p^{-C}$. By Hoeffding's inequality, we have

$$\begin{aligned} & P\left(\left|\frac{1}{M}\sum_{m=1}^{M}\mathbf{e}_j^{\mathrm T}\widehat{\boldsymbol\Theta}^{(m)}\left\{\frac{1}{n}\sum_{i\in D_m} y_i\mathbf{x}_i I\{y_i\mathbf{x}_i^{\mathrm T}\boldsymbol\beta_0 \le 1\}\right\}\right| > t\right) \\ = \; & P\left(\left|\frac{1}{N}\sum_{m\le M, i\in D_m} a_{ij} I\{y_i\mathbf{x}_i^{\mathrm T}\boldsymbol\beta_0 \le 1\}\right| > t\right) \\ \le \; & 2\exp\left\{-CC_n^{-2}Nt^2\right\}. \end{aligned} \tag{8}$$

Thus

$$\left\|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol\Theta}^{(m)}\left\{\frac{1}{n}\sum_{i\in D_m} y_i\mathbf{x}_i I_{\{y_i\mathbf{x}_i^{\mathrm T}\boldsymbol\beta_0 \le 1\}}\right\}\right\|_\infty \le CC_n\sqrt{\frac{\log p}{N}},$$

with probability at least $1 - p^{-C}$. ■

Under reasonable assumptions, the dominating term in the bounds in Theorem 2 is $\sqrt{\log p/N}$. One version of such assumptions is presented below without proof, since it is based on simple algebra.

Corollary 1 *Assume the same conditions as in Theorem 2. In addition, we assume $s, K, \|\boldsymbol{\beta}_0\|$ are bounded, β_{01} is bounded away from zero, $h \sim n^{-1/5}$, $\log p/n^{2/5} \to 0$ and $M^3 = O(N/\log p)$, then*

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leq C\sqrt{\frac{\log p}{N}},$$

with probability at least $1 - p^{-C}$.

Remark 1 *In our case, M can scale like $(N/\log p)^{1/3}$ while for the smooth loss consider in Lee et al. (2017), M can scale as $(N/\log p)^{1/2}$. This is mainly due to that for the unsmooth loss function, the empirical process as in Proposition 2 has a slower rate. In particular, in Proposition 2 the derived bound in terms of N scales as $N^{-3/4}$. For smooth loss, this term would have been N^{-1}, which eventually leads to the constraint that M should scale like $N^{1/3}$ instead of $N^{1/2}$. Although we are not claiming the bound of Proposition 2 is optimal, it is common to see that for unsmooth functions the empirical process converges slower than that for smooth functions (Belloni and Chernozhukov, 2011).*

2.3 Thresholding aggregated estimator

As mentioned in subsection 2.1, the l_2 norm of $\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0$ is generally unfavorably large compared to the centralized estimator using all observations. This is also illustrated in our simulations. To improve performance, thresholding can be used as a post-processing step which produced a sparse aggregate estimator. Let c be a threshold level. We define $\bar{\boldsymbol{\beta}}^c = (\bar{\beta}_1^c, \ldots, \bar{\beta}_p^c)^{\mathrm{T}}$ where $\bar{\beta}_j^c = \bar{\beta}_j I\{|\bar{\beta}_j| > c\}$. Here for illustration we used hard thresholding and similar results hold for soft thresholding. Under appropriate choice of the threshold, the thresholded estimator has the same convergence rate as the centralized estimator in l_1 and l_2 norm, when choosing $c \asymp \sqrt{\log p/N}$.

Theorem 3 *On the event $c > \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty$, we have $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_\infty \leq 2c$, $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_1 \leq 2sc$ and $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\| \leq 2\sqrt{s}c$.*

Proof of Theorem 3. Using $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_\infty \leq \|\bar{\boldsymbol{\beta}}^c - \bar{\boldsymbol{\beta}}\|_\infty + \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty \leq 2c$ giving the first result. Since $c > \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_\infty$, we have $\bar{\boldsymbol{\beta}}_j^c = 0$ if $\boldsymbol{\beta}_{0j} = 0$ and thus the support of $\bar{\boldsymbol{\beta}}^c$ is contained in that of $\boldsymbol{\beta}_0$. This implies $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_1 \leq s\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_\infty \leq 2sc$ and $\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\| \leq \sqrt{s}\|\bar{\boldsymbol{\beta}}^c - \boldsymbol{\beta}_0\|_\infty \leq 2\sqrt{s}c$. ■

3. Simulations

We illustrate the performances of the distributed estimators of the linear SVM coefficients via simulations. We generate the data from the following model. First $y_i, i = 1, \ldots, N$ are generated from the binary distribution $P(y_i = 1) = P(y_i = -1) = 0.5$. Given $y_i = 1$, $\mathbf{x}_i$ is generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (0.2, -0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jj'})$ with $\sigma_{jj} = 1$ for all j, $\sigma_{jj'} = 0.2$ if $j \leq 5, j' \leq 5, j \neq j'$, $\sigma_{jj'} = 0$ otherwise. Given $y_i = -1$, $\mathbf{x}_i$ is generated from a multivariate normal distribution with mean $-\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. By the calculations in Appendix B of Peng et al. (2016), the true parameter can be found to be $\boldsymbol{\beta}_0 = (0.217, -0.503, 0.397, 0.573, 0.750, 0, \ldots, 0)^{\mathrm{T}}$.

The tuning parameters λ in the penalty and the bound Cb_n used in finding the matrix inverse are selected by 5-fold cross-validation in each local machine. For the threshhold c, we choose c such that the number of nonzero components of $\bar{\boldsymbol{\beta}}$ is equal to the maximum number of nonzero components in the M local estimates. The bandwidth h is another tuning parameter. Kato (2012) has derived the optimal bandwidth for quantile regression, but it is hard to see whether similar results can be obtained in the current setting. Thus we have used Silverman's rule of thumb for kernel density estimation $h = 1.06\widehat{\sigma}n^{-1/5}$ where $\widehat{\sigma}$ is the sample standard deviation of $1 - y_i\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. In our context this rule of thumb has no theoretical support, but seems to work well in practice. In our case, it seems we do not need to estimate the Hessian optimally since our purpose is not to perform inferences of $\boldsymbol{\beta}$. Finally, we use standard normal density as the smoothing kernel q.

We compute the centralized estimator (CE), the naive aggregated estimator without using bias correction (NAE), the aggregated estimator after debiasing (AE), and the final thresholded estimator (TE). The accuracy of the estimators are assessed by the l_∞ error ($\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_\infty$), the l_2 error ($\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$), as well as the prediction error based on independently generated $50,000$ observations.

First, we set $N = 20,000$, $M = 1, 5, 10, 15, 20, 25, 30$ ($M = 1$ is the centralized estimator) and $p = 5000$. Figure 1 shows errors of the estimators that change with M, based on 100 data sets generated in each scenario. The performances generally deteriorate with the increase of M. In terms of l_∞ error the effect of thresholding is very small if any, and both TE and AE (almost identical) are better than NAE. In terms of l_2 error, AE is much worse than NAE. This is due to that AE is non-sparse and the summation of small errors over p variables can be very large. On the other hand, although NAE may have large errors on the nonzero coefficients due to the large bias, the error is small on many zero coefficients which are estimated exactly as zero (NAE is sparse). Thresholding is effective in reducing the l_2 error as well as prediction error.

In the second set of simulations, we still use $p = 5000$ and consider different sample sizes $N = 10000, 20000, 30000, 40000, 50000$, and fix the number of samples in each local machine to be $n = 5000$ (and thus the number of machines M increases with N from 2 to 10). For this simulation, as suggested by a reviewer, we also compute the thresholded estimator (after debiasing and aggregation) using the true Hessian (the true Hessian can be computed as explained in Appendix B). From the reported results in Figure 2, it is seen that the proposed estimator TE has errors decreasing with total sample size, while the errors of the naive aggregated estimator are much larger. For AE which is non-sparse, its performance in terms of l_2 and prediction error is the worst among different estimators. It is also seen that the thresholded estimator using the true Hessian has similar performances as TE.

The simulations are carried out on the computational cluster Katana in the University of New South Wales. For the second set of simulations for example, the central estimators require from 4 to 29 hours to compute depending on the sample size, to finish all 100 repetitions, while the distributed estimator requires about 2 hours for all sample sizes.

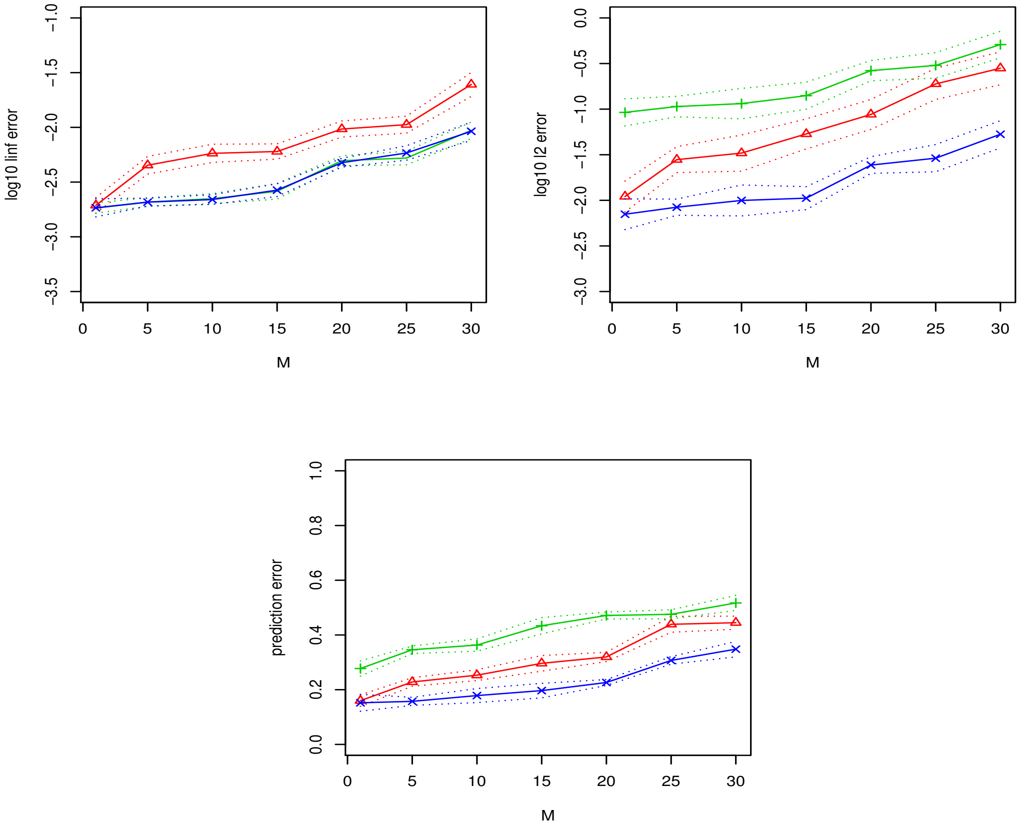

Figure 1: The l_∞, l_2 and prediction errors of estimates with $M \in \{1, 5, 10, 15, 20, 25, 30\}$ ($M = 1$ represents the centralized estimator). $\triangle$(red): naive aggregated estimator (NAE); $+$(green): the aggregated estimator after debiasing (AE); $\times$(blue): the thresholded estimator (TE). The l_∞ and l_2 errors are in the logarithmic scale with base 10. The dotted lines are computed based on the 100 repetitions showing 2 standard deviations of the estimated error.

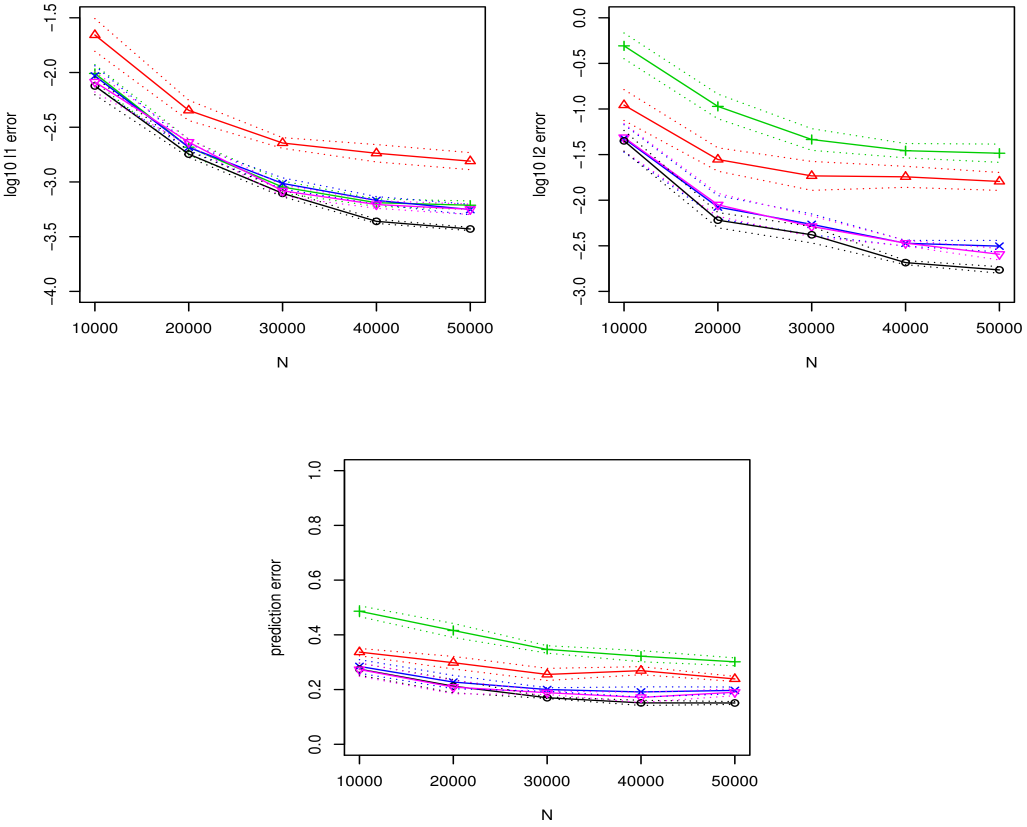

Figure 2: The l_∞ and l_2 errors of estimates with $p = 5000$ and $N \in \{10000, 20000, 30000, 40000, 50000\}$. $\circ$(black): centralized estimator (CE); $\triangle$(red): naive aggregated estimator (NAE); $+$(green): the aggregated estimator after debiasing (AE); $\times$(blue): the thresholded estimator (TE). $\triangledown$(purple): the thresholded estimator when the true Hessian is used in debiasing. The l_∞ and l_2 errors are in the logarithmic scale with base 10. The dotted lines are computed based on the 100 repetitions showing 2 standard deviations of the estimated error.

4. Conclusion

In this paper, we consider distributed estimation of l_1-penalized linear SVM. As long as the number of machines is not too large, the distributed estimator has the same convergence rate as the centralized estimator in l_∞, l_1 and l_2 norms, if the estimator is thresholded to retain sparsity.

We note that the optimization problem of Equation 6 can be solved row by row, and thus can also be done in a distributed way. Using local data, each local machine can obtain estimates for p/M rows of $\mathbf{\Theta}$ and then these estimates can be combined to obtain a *single* estimate of $\mathbf{\Theta}$ that satisfies $\|\widehat{\mathbf{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0)-\mathbf{I}\|_\infty \leq Cb_n$ with probability at least $1-p^{-C}$, where $b_n = C_n\left(\left(\frac{1}{\beta_{01}}+\sqrt{\frac{\log p}{nh^3\beta_{01}}}+\frac{\log p}{nh^2}\right)s\sqrt{\frac{\log p}{n}}+\frac{s^2\log p}{nh^3}+\frac{h}{\beta_{01}^2}+\sqrt{\frac{\log p}{nh\beta_{01}}}\right)$ and the same bound as in Theorem 2 for $\|\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty$ hold with minor modifications of the proofs. For ease of implementation, we do not investigate this alternative in our numerical studies.

Once an aggregated estimator of $\boldsymbol{\beta}$ is obtained, one can use this estimator in the evaluation of the inverse of $\mathbf{H}(\widehat{\boldsymbol{\beta}})$ in each local machine. This iterative approach requires further communications among the central machine and local machines, and we do not observe improved performances of the iterative estimator empirically.

A few problems remain open in this study, among possibly many others. First, there is a gap in theory and simulation in that (A1) assumed that the covariates are bounded while the simulations are based on a multivariate normal distribution for covariates. We used the assumption of bounded covariates for at least two reasons. One reason is that we rely on the results of Peng et al. (2016), who assumed boundedness of predictors. The second reason is that we have used boundedness in our own derivation in various places, even though such assumption may be relaxed with more efforts and much messier proof. Another open question is whether the implied constraint on M mentioned at the end of Section 2.2 is tight. This appears to be a challenging open question that we are not currently able to answer.

Acknowlegements

The authors sincerely thank the Action Editor Professor Jie Peng, and two anonymous reviewers for their insightful comments and suggestions which greatly improved the paper. The research of Lian is partially supported by City University of Hong Kong Start Up Grant 7200521.

Appendix A. Proofs of Propositions.

Proof of Proposition 1. We first show that

$$\|\widehat{\mathbf{\Theta}}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{I}\|_\infty \leq Cb_N \text{ and } \|\widehat{\mathbf{\Theta}}\|_{L_1}\leq C_N. \tag{9}$$

First, we note these will be implied by that

$$\|\mathbf{\Theta}_0\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{I}\|_\infty \leq Cb_N. \tag{10}$$

In fact, Equation 10 means the constrained optimization problem is feasible and the feasibility of $\widehat{\mathbf{\Theta}}$ immediately implies the first equation of Equation 9. For the second equation of Equation 9, we only need to note that by Equation 10 and the definition of the constrained optimization problem, which can be solved for each row of $\mathbf{\Theta}$ separately, we have $\|\widehat{\mathbf{\Theta}}\|_{L_1} \leq \|\mathbf{\Theta}_0\|_{L_1}$.

To establish Equation 10, we bound $\|\mathbf{\Theta}_0\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{I}\|_\infty = \|\mathbf{\Theta}_0(\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{H}(\boldsymbol{\beta}_0))\|_\infty \leq \|\mathbf{\Theta}_0\|_{L_1}\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{H}(\boldsymbol{\beta}_0)\|_\infty$. Furthermore, we write

$$\begin{aligned}
& \|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \mathbf{H}(\boldsymbol{\beta}_0)\|_\infty \\
\leq\; & \|E[\widehat{\mathbf{H}}(\boldsymbol{\beta}_0)] - \mathbf{H}(\boldsymbol{\beta}_0)\|_\infty + \|\widehat{\mathbf{H}}(\boldsymbol{\beta}_0) - E[\widehat{\mathbf{H}}(\boldsymbol{\beta}_0)]\|_\infty + \|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}}) - \widehat{\mathbf{H}}(\boldsymbol{\beta}_0)\|_\infty \\
=:\; & I_1 + I_2 + I_3.
\end{aligned}$$

First we consider I_1 and show that for any bounded function $s(\mathbf{x})$ whose partial derivatives are also bounded, we have

$$\left| E\left[\frac{1}{h}q\left(\frac{1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right] - E\left[\delta(1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})\right]\right| \leq Ch/\beta_{01}^2,$$

and we will then obtain $I_1 \leq Ch/\beta_{01}^2$ by setting $s(\mathbf{x}) = x_j x_{j'}, 1 \leq j, j' \leq p$.

In fact, since, for example, $E[\delta(1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})] = P(y=1)E[\delta(1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})|y=1]+P(y=-1)E[\delta(1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})|y=-1]$, we only need to consider conditional expectation given $y=1$ (conditional expectation given $y=-1$ is similar). Write $\boldsymbol{\beta}_{0,-1} = (\beta_{02},\ldots,\beta_{0p})^{\mathrm{T}}$ and $\mathbf{x}_{-1} = (x_2,\ldots,x_p)^{\mathrm{T}}$. Then, by a change of variable $(x_1,\ldots,x_p) \to (z_1, x_2,\ldots,x_p)$ with $z_1 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$, we have

$$\begin{aligned}
& E[\frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})|y=1] \\
=\; & \int \frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\
=\; & \int \frac{1}{h}q\left(\frac{1-z_1}{h}\right)s\left(\frac{z_1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)f\left(\frac{z_1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}dz_1 d\mathbf{x}_{-1} \\
\overset{u=(1-z_1)/h}{=}\; & \int q(u)\,s\left(\frac{1-uh-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)f\left(\frac{1-uh-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}du d\mathbf{x}_{-1} \\
=\; & \int q(u)(sf)\left(\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}du d\mathbf{x}_{-1} \\
& -\int q(u)(sf)^{(1)}\left(\frac{1-*-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}^2}uhdud\mathbf{x}_{-1},
\end{aligned}$$

where $(sf)^{(1)}$ is the partial derivative of $(sf)(.)$ with respect to its first variable and $*$ represents a value between 0 and uh, and

$$\begin{aligned}
& E[\delta(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})|y=1] \\
=\; & \int \delta(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\
=\; & \int s\left(\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)f\left(\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}dz_1 d\mathbf{x}_{-1}.
\end{aligned}$$

Using $\int q(u)du = 1$, $\int |u|q(u)du < \infty$, and that $(sf)^{(1)}$ is bounded, we get

$$\begin{aligned}
&\left|E[\frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})|y=1] - E[\delta(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})|y=1]\right| \\
= &\left|\int q(u)(sf)^{(1)}\left(\frac{1-*-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}}, \mathbf{x}_{-1}\right)\frac{1}{\beta_{01}^2}uhdud\mathbf{x}_{-1}\right| \\
\leq\ & Ch/\beta_{01}^2,
\end{aligned}$$

and thus

$$I_1 \leq Ch/\beta_{01}^2. \tag{11}$$

Next we deal with I_2. Again with a bounded function s, we have $\left|\frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right| \leq C/h$ and

$$\begin{aligned}
&E\left[\left(\frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right)^2|y=1]\right] \\
= &\int \frac{1}{h^2}q^2\left(\frac{1-z_1}{h}\right)s^2\left(\frac{z_1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}}, \mathbf{x}_{-1}\right)f\left(\frac{z_1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}}, \mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}dz_1d\mathbf{x}_{-1} \\
= &\int \frac{1}{h}q^2(u)(s^2f)\left(\frac{1-uh-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}}, \mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}dud\mathbf{x}_{-1} \\
\leq\ & C/(h\beta_{01}),
\end{aligned} \tag{12}$$

since $\int q^2(u)du < \infty$. Thus

$$E\left[\left(\frac{1}{h}q\left(\frac{1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right)^r \middle| y=1\right] \leq (C/h)^{r-2}(1/(h\beta_{01})),\ r \geq 2.$$

By Bernstein's inequality (Lemma 2.2.11 in van der Vaart and Wellner (1996)),

$$\begin{aligned}
&P\left(\left|\frac{1}{N}\sum_i I_{\{y_i=1\}}\frac{1}{h}q\left(\frac{1-y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x}_i) - E\left[I_{\{y=1\}}\frac{1}{h}q\left(\frac{1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right]\right| > t\right) \\
\leq\ & 2\exp\left\{-C\frac{Nht^2}{t+\beta_{01}^{-1}}\right\},
\end{aligned}$$

and the same inequality holds with $y = 1$, $y_i = 1$ replaced by $y = -1$, $y_i = -1$, and thus

$$\left|\frac{1}{N}\sum_i \frac{1}{h}q\left(\frac{1-y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x}_i) - E\left[\frac{1}{h}q\left(\frac{1-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0}{h}\right)s(\mathbf{x})\right]\right| \leq C\left(\sqrt{\frac{\log p}{Nh\beta_{01}}} + \frac{\log p}{Nh}\right) \tag{13}$$

with probability at least $1-p^{-C}$ (note C here can be arbitrarily large as long as we are willing to set the constant C in the display above to be large enough). By choosing $s(\mathbf{x}) = x_jx_{j'}$, using the the union bound,

$$I_2 \leq C\left(\sqrt{\frac{\log p}{Nh\beta_{01}}} + \frac{\log p}{Nh}\right), \tag{14}$$

with probability at least $1-p^{-C}$.

Finally, we bound I_3. By Taylor's expansion, we have

$$\begin{aligned}
& \left|\frac{1}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q((1-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/h)/h-\frac{1}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q((1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h\right| \\
\leq & \left|\frac{C}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q'((1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2\cdot\mathbf{x}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\right| \\
& +\left|\frac{C}{2N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q''(*)/h^3\cdot(\mathbf{x}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))^2\right| \\
=: & \ J_1+J_2,
\end{aligned}$$

where $*$ denotes a value between $(1-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/h$ and $(1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h$. The second term above is easy to deal with. Since $q''(.)$ is bounded, we get

$$J_2\leq\frac{Cs^2\log p}{Nh^3},$$

with probability $1-p^{-C}$. Although J_1 could be bounded similar to J_2, a more careful calculation will yield a tighter bound. For this we write

$$\begin{aligned}
& \left|\frac{1}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q'(1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0/h)/h^2\right| \\
\leq & \left|\frac{1}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q'((1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2-E[I_{\{y=1\}}x_jx_{j'}q'(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0/h)/h^2]\right| \\
& +\left|E[I_{\{y=1\}}x_jx_{j'}q'(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0/h)/h^2]\right|.
\end{aligned}$$

Similar to Equation 12, we have $|q'((1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2|\leq C/h^2$ and $E[(q'((1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2)^2|y=1]\leq C/(h^3\beta_{01})$, and by the same arguments as those that lead to Equation 13, we get

$$\begin{aligned}
& \max_{j,j'}\left|\frac{1}{N}\sum_i I_{\{y_i=1\}}x_{ij}x_{ij'}q'((1-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2-E[I_{\{y=1\}}x_jx_{j'}q'(1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0/h)/h^2]\right| \\
\leq & \ C\left(\sqrt{\frac{\log p}{Nh^3\beta_{01}}}+\frac{\log p}{Nh^2}\right),
\end{aligned}$$

with probability at least $1-p^{-C}$. Furthermore, for any bounded $s(\mathbf{x})$ whose partial derivatives are also bounded, we have

$$\begin{aligned}
& \left|E[s(\mathbf{x})q'((1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/h)/h^2|y=1]\right| \\
= & \left|\int q'((1-z_1)/h)/h^2\cdot(sf)\left(\frac{z_1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{\beta_{01}}dz_1d\mathbf{x}_{-1}\right| \\
= & \left|\int q'(u)\cdot(sf)\left(\frac{1-uh-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)\frac{1}{h\beta_{01}}dud\mathbf{x}_{-1}\right| \\
= & \left|\int(sf)\left(\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_{01}},\mathbf{x}_{-1}\right)(\int q'(u)du)\frac{1}{h\beta_{01}}d\mathbf{x}_{-1}\right|+\left|\int q'(u)u(sf)^{(1)}(*)\frac{1}{\beta_{01}^2}dud\mathbf{x}_{-1}\right| \\
\leq & \ C/\beta_{01}^2,
\end{aligned}$$

using that $\int q'(u)du=0$. Thus

$$J_1\leq C\left(\frac{1}{\beta_{01}^2}+\sqrt{\frac{\log p}{Nh^3\beta_{01}}}+\frac{\log p}{Nh^2}\right)\cdot s\sqrt{\frac{\log p}{N}}$$

with probability at least $1-p^{-C}$ and then

$$I_3\leq C\left(\left(\frac{1}{\beta_{01}^2}+\sqrt{\frac{\log p}{Nh^3\beta_{01}}}+\frac{\log p}{Nh^2}\right)\cdot s\sqrt{\frac{\log p}{N}}+\frac{s^2\log p}{Nh^3}\right), \tag{15}$$

with probability at least $1-p^{-C}$. Combining bounds in Equation 11, Equation 14 and Equation 15, we get

$$\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty\leq C\left(\left(\frac{1}{\beta_{01}^2}+\sqrt{\frac{\log p}{Nh^3\beta_{01}}}+\frac{\log p}{Nh^2}\right)\cdot s\sqrt{\frac{\log p}{N}}+\frac{s^2\log p}{Nh^3}+\frac{h}{\beta_{01}^2}+\sqrt{\frac{\log p}{Nh\beta_{01}}}\right),$$

and thus

$$\|\boldsymbol{\Theta}_0\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{I}\|_\infty\leq Cb_N,$$

with probability at least $1-p^{-C}$.

Finally, we also have

$$\begin{aligned}
& \|\widehat{\boldsymbol{\Theta}}\mathbf{H}(\boldsymbol{\beta}_0)-\mathbf{I}\|_\infty \\
\leq & \ \|\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{I}\|_\infty+\|\widehat{\boldsymbol{\Theta}}(\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{H}(\boldsymbol{\beta}_0))\|_\infty \\
\leq & \ Cb_N+\|\widehat{\boldsymbol{\Theta}}\|_{L_1}\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty \\
\leq & \ Cb_N,
\end{aligned}$$

using the bound for $\|\widehat{\mathbf{H}}(\widehat{\boldsymbol{\beta}})-\mathbf{H}(\boldsymbol{\beta}_0)\|_\infty$ above. ■

Proof of Proposition 2. We define $\Omega=\{\boldsymbol{\beta}\in\mathbb{R}^p:\|\boldsymbol{\beta}\|_0\leq K,\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1\leq Cs\sqrt{\log p/N}\}$ and we have $\widehat{\boldsymbol{\beta}}\in\Omega$ with probability at least $1-p^{-C}$. Define the class of functions

$$\mathcal{G}_j=\{yx_j(I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\leq 1\}-I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0\leq 1\}):\boldsymbol{\beta}\in\Omega\},$$

with squared integrable envelope function $F(\mathbf{x}, y) = |x_j|$.

We decompose Ω as $\Omega = \cup_{T \subset \{1,\ldots,p\}, |T| \leq K} \Omega(T)$ with $\Omega(T) = \{\boldsymbol{\beta} : \text{ support of } \boldsymbol{\beta} \subset T\} \cap \Omega$. We also define $\mathcal{G}_j(T) = \{y x_j (I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \leq 1\} - I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1\}) : \boldsymbol{\beta} \in \Omega(T)\}$.

By Lemma 2.6.15, Lemma 2.6.18 (vi) and (viii) (actually by the proof of Lemma 2.6.18 (viii)) in van der Vaart and Wellner (1996), for each fixed $T \subset \{1, \ldots, p\}$ with $|T| \leq K$, $\mathcal{G}_j(T)$ is a VC-subgraph with index bounded by $K + 2$ and by Theorem 2.6.7 of van der Vaart and Wellner (1996), we have

$$N(\epsilon, \mathcal{G}_j(T), L_2(P_n)) \leq \left(\frac{C\|F\|_{L_2(P_n)}}{\epsilon}\right)^{CK} \leq \left(\frac{C}{\epsilon}\right)^{CK}.$$

Since there are at most $\binom{p}{K} \leq (ep/K)^K$ different such T, we have

$$N(\epsilon, \mathcal{G}_j, L_2(P_n)) \leq \left(\frac{C}{\epsilon}\right)^{CK} \left(\frac{ep}{K}\right)^{K} \leq \left(\frac{Cp}{\epsilon}\right)^{CK},$$

and thus

$$N(\epsilon, \cup_{j=1}^{p}\mathcal{G}_j, L_2(P_n)) \leq p\left(\frac{Cp}{\epsilon}\right)^{CK}.$$

Let $\sigma^2 = \sup_{f \in \cup_j \mathcal{G}_j} Pf^2$. Then by Theorem 3.12 of Koltchinskii (2011), we have

$$E\|R_n\|_{\cup_j \mathcal{G}_j} \leq C\left(\sigma\sqrt{\frac{K\log p}{N}} + \frac{K \log p}{N}\right),$$

where $\|R_n\|_{\cup_j \mathcal{G}_j} = \sup_{f \in \cup_j \mathcal{G}_j} N^{-1}\sum_{i=1}^{N} \varepsilon_i f(\mathbf{x}_i, y_i)$ with ε_i being i.i.d. Rademacher random variables. Using the symmetrization inequality which states that $E\|P_n - P\|_{\cup_j \mathcal{G}_j} \leq 2E\|R_n\|_{\cup_j \mathcal{G}_j}$, where $\|P_n - P\|_{\cup_j \mathcal{G}_j} = \sup_{f \in \cup_j \mathcal{G}_j} N^{-1}\sum_i f(\mathbf{x}_i, y_i) - Ef(\mathbf{x}, y)$, Talagrand's inequality (page 24 of Koltchinskii (2011)) gives

$$P\left(\|P_n - P\|_{\cup_j \mathcal{G}_j} \geq C\left(\sigma\sqrt{\frac{K\log p}{N}} + \frac{K\log p}{N} + \sqrt{\frac{\sigma^2 t}{N}} + \frac{t}{N}\right)\right) \leq e^{-t},$$

that is, with probability at least $1 - p^{-C}$,

$$\begin{aligned} & \left\|\frac{1}{N}\sum_i y_i \mathbf{x}_i (I\{y_i \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \leq 1\} - I\{y_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1\}) - Ey\mathbf{x}(I\{y\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \leq 1\} - I\{y\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1\})\right\|_\infty \\ \leq \; & C\left(\sigma\sqrt{\frac{K\log p}{N}} + \frac{K\log p}{N}\right). \end{aligned}$$

Finally, we need to decide the size of σ^2. For $\boldsymbol{\beta} \in \Omega$, we have that

$$\begin{aligned} & E[(I\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \leq 1\} - I\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1\})^2 | y = 1] \\ \leq \; & P(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \leq 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \geq 1 | y = 1) + P(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \geq 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1 | y = 1) \\ \leq \; & P(1 \leq \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1 + Cs\sqrt{\log p/N} | y = 1) + P(1 - Cs\sqrt{\log p/N} \leq \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \leq 1 | y = 1) \\ \leq \; & Cs\sqrt{\log p/N}/\beta_{01}, \end{aligned}$$

where we used in the second inequality the fact that $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \le 1 \le \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ implies $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \le 1 + |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| \le 1 + C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1$ and similarly that $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 \ge 1 - C\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1$, and in the last inequality we used that the density of $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ conditional on $y = 1$ is bounded by $1/\beta_{01}$. This last observation follows easily from that, by change of variable $z_1 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$, the joint density of $(z_1, \mathbf{x}_{-1})$ conditional on $y = 1$ is given by

$$f((z_1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1})/\beta_{01}, \mathbf{x}_{-1})/\beta_{01}.$$

Thus we have $\sigma^2 \le Cs\sqrt{\log p/N}/\beta_{01}$ which proved the proposition. ■

Proof of Proposition 3. By integrating over x_1 first, we have for $s(\mathbf{x}) = x_j x_{j'}$,

$$\begin{aligned}
&\int \delta(1 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})s(\mathbf{x})f(\mathbf{x})d\mathbf{x} - \int \delta(1 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\
= \;& \int \frac{1}{\beta_1}(sf)\left(\frac{1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1}\right) d\mathbf{x}_{-1} - \int \frac{1}{\beta_{01}}(sf)\left(\frac{1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, \mathbf{x}_{-1}\right) d\mathbf{x}_{-1} \\
= \;& \frac{\beta_{01} - \beta_1}{\beta_1\beta_{01}} \int (sf)\left(\frac{1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_1}, \mathbf{x}_{-1}\right) d\mathbf{x}_{-1} + \frac{1}{\beta_{01}} \int (sf)^{(1)}(*, \mathbf{x}_{-1})\, \mathbf{x}_{-1}^{\mathrm{T}}\left(\frac{\boldsymbol{\beta}_{0,-1}}{\beta_{01}} - \frac{\boldsymbol{\beta}_{-1}}{\beta_1}\right) d\mathbf{x}_{-1},
\end{aligned}$$

where $*$ represents a value between $\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}$ and $\frac{1-\mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{-1}}{\beta_1}$. Using $|\beta_1 - \beta_{01}| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le \beta_{01}/2$, $\beta_1 \ge \beta_{01} - |\beta_1 - \beta_{01}| \ge (1/2)\beta_{01}$, and $\|\frac{\boldsymbol{\beta}_{0,-1}}{\beta_{01}} - \frac{\boldsymbol{\beta}_{-1}}{\beta_1}\|_1 \le \frac{\|\boldsymbol{\beta}_0\|_1 \cdot \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\|_1}{|\beta_{01}\beta_1|} \le \frac{C}{\beta_{01}^2}\|\boldsymbol{\beta}_0\|_1\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\|_1$, the lemma is proved. ■

Appendix B. Discussions of Assumption (A3).

First we show that $\mathbf{H}(\boldsymbol{\beta}_0)$ can also be expressed as

$$c_1 E[\mathbf{x}\mathbf{x}^{\mathrm{T}} | y = 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] + c_2 E[\mathbf{x}\mathbf{x}^{\mathrm{T}} | y = -1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = -1], \tag{16}$$

for two positive constants c_1, c_2. Indeed, let $h(z, x_2, \ldots, x_p)$ be the joint density of $(z = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, x_2, \ldots, x_p)^{\mathrm{T}}$. We have $h(z, x_2, \ldots, x_p) = f(\frac{z - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, x_2, \ldots, x_p)\frac{1}{\beta_{01}}$. Then, for any function $s(\mathbf{x})$, we have

$$\begin{aligned}
&E[\delta(1 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x}) | y = 1] \\
= \;& \int \delta(1 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})f(\mathbf{x})d\mathbf{x} \\
= \;& \int s(\frac{1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, x_2, \ldots, x_p) f(\frac{z - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, x_2, \ldots, x_p)\frac{1}{\beta_{01}} d\mathbf{x}_{-1},
\end{aligned}$$

and

$$\begin{aligned}
&E[s(\mathbf{x}) | y = 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] \\
= \;& E[s(\frac{z - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, x_2, \ldots, x_p) | y = 1, z = 1] \\
= \;& \int s(\frac{1 - \mathbf{x}_{-1}^{\mathrm{T}}\boldsymbol{\beta}_{0,-1}}{\beta_{01}}, x_2, \ldots, x_p)\frac{h(1, x_2, \ldots, x_p)}{h_z(1)} d\mathbf{x}_{-1},
\end{aligned}$$

where h_z is the marginal density of $z = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$ conditional on $y = 1$. Thus we see that

$$E[\delta(1 - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0)s(\mathbf{x})|y = 1] = h_z(1)E[s(\mathbf{x})|y = 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1],$$

which implies Equation 16

Let's further assume that y is independent of $\mathbf{x}$ given $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0$. This is a natural sufficient dimension reduction type of assumption. This is the case for example if the class label is generated as in our simulations, or if the data follows the popular logistic regression model. Also assume $\mathbf{x}$ is multivariate normal. Then $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|y = 1, \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] = E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]$ by the conditional independence. Since $\mathbf{x}$ has a symmetric distribution, $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] = E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = -1]$ and thus the Hessian is equal to a constant multiple of $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]$. Now we examine the inverse of $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]$.

Assume $E[\mathbf{x}] = \mathbf{0}$ and $Cov(\mathbf{x}) = E[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \mathbf{S}$. The literature on high-dimensional precision matrix estimation typically assume that $\mathbf{S}^{-1}$ is sparse or approximately sparse to make the estimation feasible. We first see how the inverse of $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]$ is related to $\mathbf{S}$ under normality. By the normality of $\mathbf{x}$, $(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0, x_1, \ldots, x_p)$ is again (degenerate) normal with covariance matrix

$$\begin{pmatrix} \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}_0 & \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S} \\ \mathbf{S}\boldsymbol{\beta}_0 & \mathbf{S} \end{pmatrix}.$$

Then by the property of multivariate normal distribution,

$$\begin{aligned} E[\mathbf{x}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] &= \mathbf{S}\boldsymbol{\beta}_0/(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}_0), \\ Cov(\mathbf{x}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1) &= \mathbf{S} - \frac{\mathbf{S}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S}}{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}_0}, \end{aligned}$$

and thus

$$E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] = E[\mathbf{x}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]E^{\mathrm{T}}[\mathbf{x}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1] + Cov(\mathbf{x}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1) = \mathbf{S} + a\mathbf{S}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S},$$

where a is a scalar. By the Sherman-Morrison formula,

$$\begin{aligned} &(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1])^{-1} \\ = \; &\mathbf{S}^{-1} - \frac{a}{1 + a\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{S}\boldsymbol{\beta}_0}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}. \end{aligned}$$

Thus the Hessian matrix is sparse if both $\mathbf{S}^{-1}$ and $\boldsymbol{\beta}_0$ are sparse.

Now we consider several popular and concrete cases.

Case 1. Consider the autoregressive correlation matrix where the (i, j) entry of $\mathbf{S}$ is $s_{ij} = \rho^{|i-j|}$, $|\rho| < 1$. In this case, it is known that

$$\mathbf{S}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & & & & \\ -\rho & 1+\rho^2 & -\rho & & & \\ & -\rho & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & 1+\rho^2 & -\rho \\ & & & & -\rho & 1 \end{pmatrix}.$$

In particular we can see $\|\mathbf{S}^{-1}\|_{L_1} = 1/(1-|\rho|)$ independent of the size of the matrix.

Case 2. Assume $\mathbf{S}$ is a banded matrix with a fixed bandwidth, then by Theorem 2.2 of Demko (1977), the (i,j) entry of $\mathbf{S}^{-1}$ is bounded by $C\gamma^{|i-j|}$ for some constants $C > 0$, $0 < \gamma < 1$. Thus $\mathbf{S}^{-1}$ is approximately sparse in the sense that $\|\mathbf{S}^{-1}\|_{L_1}$ is bounded.

Case 3. Consider the exchangeable correlation matrix where all the non-diagonal entries of $\mathbf{S}$ are equal to ρ. Here we are not able to give theoretical properties of $\mathbf{S}^{-1}$ but will numerically compute the L_1 norm of the inverse of $(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1])^{-1}$.

For all three cases, we set $\boldsymbol{\beta}_0 = (1,1,1,1,1,0,\ldots,0)^{\mathrm{T}}$ and $\rho = 0.3$ and report the numerical value of the L_1 norm of the inverse of $(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1])^{-1}$ in Table 1. We see that in case 1 and 2 the L_1 norm does not change with p. For case 1, this can be theoretically shown easily. It seems to be an extremely cumbersome exercise to show this for case 2, however, and thus we do not try to establish this theoretically. For case 3, we see that numerically the norm increases very slowly with p.

Table 1: L_1 norm of the inverse of $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0 = 1]$.

	$p = 50$	$p = 100$	$p = 200$	$p = 500$	$p = 1000$	$p = 5000$
case 1	5.578	5.578	5.578	5.578	5.578	5.578
case 2	5.889	5.889	5.889	5.889	5.889	5.889
case 3	7.066	7.230	7.316	7.368	7.385	7.399

References

Maria-Florina Balcan, Yingyu Liang, Le Song, David Woodruff, and Bo Xie. Communication efficient distributed kernel principal component analysis. *arXiv:1503.06858*, mar 2015.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

A Belloni and V Chernozhukov. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 36:489–531, 2008.

P. S. Bradley and O. L. Mangasarjan. Feature selection via concave minimization and support vector machines. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, number 98, pages 82–90, 1998.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, 2000.

S. Demko. Inverses of band matrices and local convergence of spline projection. *SIAM Journal on Numerical Analysis*, 14:616–619, 1977.

J Q Fan and R Z Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Yingying Fan and Jinchi Lv. Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, 108(503):1044–1061, 2013.

Kengo Kato. Asymptotic normality of Powell's kernel estimator. *Annals of the Institute of Statistical Mathematics*, 64(2):255–273, 2012.

V Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer, New York, 2011.

J Y Koo, Y Lee, Y Kim, and C Park. A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, 9:1343–1368, 2008.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18:1–30, 2017.

Y Lin. Some asymptotic properties of the support vector machine. *TR1029, University of Wisconsin, Madison*, 2000.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73–82, 2004.

Ryan McDonald, Gideon Mann, and Nathan Silberman. Efficient large-scale distributed training of conditional maximum entropy models. *Proceedings of Advances in Neural Information Processing Systems*, pages 1231–1239, 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.

Bo Peng, Lan Wang, and Yichao Wu. An error bound for l1-norm support vector machine coefficients in ultra-high dimension. *Journal of Machine Learning Research*, 17(236):1–26, 2016.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, 2001.

Minghu Song, Curt M. Breneman, Jinbo Bi, N. 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.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35:575–607, 2007.

R Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996.

R Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

Sara van de Geer, Peter Buhlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

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

A W van der Vaart and J A Wellner. *Weak convergence and empirical processes*. Springer Verlag, 1996.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer, New York, 2013.

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

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–134, 2004.

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, 2016.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of Statistics*, 44(4):1400–1437, 2016.

Zemin Zheng, Yingying Fan, and Jinchi Lv. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):627–649, 2014.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems*, 2003.

Martin a Zinkevich, Alex Smola, and Markus Weimer. Parallelized stochastic gradient descent. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.

H Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.