Model Averaging for Support Vector Machine by $J$-fold Cross-Validation*

Jiahui Zou,† Xinyu Zhang,‡ Guohua Zou,§ Alan T. K. Wan¶

June 1, 2021

Abstract

Support vector machine (SVM) is a classical tool to deal with classification problems, which is widely used in biology, statistics and machine learning and good at small sample size and high-dimensional situation. This paper proposes a model averaging method, called SVMMA, to address the uncertainty from deciding which covariates should be included for SVM and to promote its prediction ability. We offer a criterion to search the weights to combine many candidate models that are composed of different parts from the total covariates. To build up the candidate model set, we suggest to use a screening-averaging form in practice. Especially, the model averaging estimator is proved to be asymptotically optimal in the sense of achieving the lowest hinge risk among all possible combination. Finally, we do some simulation to compare the proposed model averaging method with several other model selection/averaging methods and apply to four real datasets.

Keywords: Cross-validation; model selection; model averaging; prediction; SVM.

1 Introduction

Support vector machine (SVM) [Vapnik, 1995, Wahba, 1999, Schölkopf and Smola, 2001]) is a statistical technique widely used for classification problems in biology, medicine, machine learning and other fields. One important problem in SVM is how to select the right covariates in the model.

*This paper generates from a chapter of Zou [2021].
†School of Statistics, Capital University of Economics and Business, Beijing 100070, China
‡Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
§School of Mathematical Sciences, Capital Normal University, Beijing 100048, China
¶Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
There exist many studies on covariate selection for SVM. Weston et al. [2000] proposed a scaling method and Guyon et al. [2002] suggested a recursive feature elimination procedure. Some authors have also considered regularisation methods within then context of SVM. For example, Bradley and Mangasarian [1998], Zhu et al. [2004] and Wegkamp and Yuan [2011] investigated the properties of a $L_1$ penalized SVM. Wang et al. [2006] considered SVMs with $L_1$ and $L_2$ penalties. Zou and Yuan [2008] considered the $L_{\infty}$ penalized SVM in the presence of prior knowledge on the grouping information of features. Zhang et al. [2006] and Becker et al. [2011] suggested SVM with a non-convex penalty in the application of gene selection. Park et al. [2012] and Zhang et al. [2014] investigated the oracle property of the SCAD-penalized SVM with a fixed number of predictors and under a class of non-convex penalties. Zhang et al. [2016a] developed a consistent information criterion for SVM with divergent-dimensional covariates. Further, this work also offers the proof of consistency for SVMIC$_L$ proposed by Claeskens et al. [2008] in finite dimension situation.

Model averaging is an alternative method for handling variable selection that is receiving increasing attention. Unlike model selection that chooses one champion model from the pool and discounts all others in subsequent analysis, model averaging combines different candidate models by an appropriate weighting scheme. The aggregate model then forms the basis of subsequent analysis including inference. A major takeaway from the growing literature is that model averaging is advantageous from the point of view of estimator’s risk and a more robust strategy than model selection [Hansen, 2007, Wan et al., 2010]. Although Bayesian model averaging (BMA) has been a well-known technique, the involvement of a subjective prior in the weight assignment is a major critique of BMA. See Hoeting et al. [1999] for a review of the BMA methodology. Frequentist model averaging (FMA), which is a more recent vintage, avoids the above issue of BMA. Many data-driven weight choice methods within the FMA paradigm, including the smoothed information criteria [Buckland et al., 1997, Claeskens et al., 2006], optimal weighting [Hansen, 2007, Zhang et al., 2016b, 2020], adaptive weighting [Yuan and Yang, 2005, Zhang et al., 2013] and others. So far, the SVM literature has paid only scant attention to model averaging. Das and Bhattacharya [2015] provides a approach, called DWPA, to combine the coefficient estimators from different notes in distributed computation, but only offers the proof of robust. Therefore, this paper plans to use model averaging method to address the uncertainty from variable selection and promote the prediction ability for SVM.

The contribution of this paper mainly lies in three aspects. First, we provide a weight choice criterion and the proof of asymptotically optimal for the proposed model averaging estimator, i.e., the estimator will get a smallest hinge risk among all feasible combination in the asymptotic sense. Second, we put forward a model screening procedure before model averaging to reduce the computational burden. Third, we develop optimal model averaging theories for hinge loss which is
non-symmetrical and non-smoothed.

The rest of the paper is organized as follows. Section 2 introduces the model averaging method for SVMs. Section 3 provides the conditions and theoretical results for the proposed model averaging method. Section 4 presents the simulation studies and Appendix provides the proofs of lemmas and theorems.

2 Model setup and estimation

2.1 Model averaging for SVM

Consider a random sample \( D_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \), where \( y_i \in \{-1, +1\} \) and each of \( (x_i, y_i) \) is independently drawn from an identical distribution. Denote \( x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^{p+1} \), and \( x_i^+ = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \). Conformably, write \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1} \) and \( \beta^+ = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^{(p+1)} \), where \( \beta^+ \) is the coefficient vector corresponding to \( x^+_i \). Linear SVM aims to determine a hyperplane, defined by \( x^T \beta = 0 \), to draw a decision boundary between \( y = 1 \) and \( y = -1 \). This hyperplane is commonly estimated by solving the following optimisation problem [Hastie et al., 2001]:

\[
\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta)_+ + \frac{\lambda_n}{2} \| \beta^+ \|^2 \right\},
\]

where \( \| a \| \) is the Euclidean norm operator of the vector \( a \), \( (1 - t)_+ = \max(1 - t, 0) \) is the hinge loss function, and \( \lambda_n \) is a tuning parameter. One usually tackles the problem about the uncertainty in \( x \) by model selection, as discussed in Section 1. Here, we consider the alternative strategy of model averaging that combines models with different covariates. It is assumed that each model contains a minimum of one covariate in addition to the intercept term. Under this setup, there exist a maximum of \( 2^p - 1 \) potential candidate models. Considering some models are lack of information and the amount is too large to calculate directly, we can discard some ones before estimation in practical application. Without loss of generality, assume there are \( S_n \) ponderable candidate models to use. Clearly, it is required that \( S_n \leq 2^p - 1 \). For \( s \in \{1, \ldots, S_n\} \), denote \( M_s = \{j_1, \ldots, j_{p_s}\} \subset \{1, \ldots, p\} \) as the set consisting of the indices of elements of \( x_{(s)i} \). For the \( s^{th} \) model, \( x_{(s)i} = (1, x_{(s)i,j_1}, \ldots, x_{(s)i,j_{p_s}})^T \) and \( \beta_{(s)}^T = (\beta_{(s)0}, \beta_{(s)j_1}, \ldots, \beta_{(s)j_{p_s}}) \). The estimator of \( \beta_{(s)} \) is obtained by solving the optimisation problem described in (1), replacing \( \beta \) by \( \beta_{(s)} \) and \( x_i \) by \( x_{(s)i} \) everywhere, yielding

\[
\hat{\beta}_{(s)} = \arg \min_{\beta_{(s)}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_{(s)i}^T \beta_{(s)})_+ + \frac{\lambda_n}{2} \| \beta_{(s)}^+ \|^2 \right\}. \tag{2}
\]
Following Koo et al. [2008], we denote $\beta^*_s$ as the “quasi-true” parameter\footnote{When the forms between work model and data generating process are not coincident, the limitation of parameter estimation is “quasi-true” parameter. If the forms are coincident, the “quasi-true” parameter and true parameter are the same.} that minimises the population hinge loss. That is,

$$
\beta^*_s = \arg \min_{\beta_s} \mathbb{E}(1 - y x^T \beta_s) + .
$$

Model averaging combines the estimates obtained from the different models via a weighted average. Now, to unify the dimensions of the estimators from different models, let $\Pi_s$ be a $(p + 1) \times (p_s + 1)$ dimensional selection matrix consisting of 1 or 0 and write $\hat{\beta}_s = \Pi_s \hat{\beta}_s$. For example, if the covariate vector of model $s$ is $x_{(s),i} = (1, x_{(s),i3})^T$, and $\hat{\beta}_s = (1, 2)^T$, then $\Pi_s = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^T$ and $\hat{\beta}_s = (1, 0, 0, 2)^T$. The model average estimator of $\beta$ is a weighted sum of $\hat{\beta}_s$'s, $s = 1, \cdots, S_n$, i.e.,

$$
\hat{\beta}(w) = \sum_{s=1}^{S_n} w_s \hat{\beta}_s,
$$

where $w = (w_1, \ldots, w_{S_n})^T$ is the weight vector belonging to the set $W = \{ w \in [0, 1]^{S_n} : \sum_{s=1}^{S_n} w_s = 1 \}$. We label $\hat{\beta}(w)$ as the SVM model average (SVMMA) estimator.

### 2.2 Weight choice criterion

As discussed above, the standard SVM approach derives the coefficient estimates by minimising the hinge loss associated with a given model. Analogously, when more than one model is involved, the hinge loss may be modified as

$$
\frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \hat{\beta}(w) \right) + .
$$

The purpose is to find a weight vector $w$ to be used in (4) to yield an optimal property for the resultant SVMMA. Clearly, the hinge loss in (5) favours bigger models, thus creating a high chance of over-fitting. For this reason we do not calculate $w$ by minimising (5) directly and choose to minimise the out-of-sample risk, i.e.,

$$
R_n(w) = \mathbb{E} \left\{ \left( 1 - \tilde{y} \tilde{x}^T \hat{\beta}(w) \right)_+ | D_n \right\},
$$

where $(\tilde{y}, \tilde{x})$ is an independent copy from the same distribution of $D_n$, the expectation calculates for $(\tilde{y}, \tilde{x})$ but $\hat{\beta}(w)$ is estimated based on $D_n$. As we do not know the real distribution of $D_n$, one way
to handle this problem is to minimize an estimate of this risk, namely, the out-of-sample prediction error. It is defined as

\[ Q_m(w) = \frac{1}{m} \sum_{i=1}^{m} \left( 1 - \tilde{y}_i \tilde{x}_i^T \tilde{\beta}(w) \right)_+. \]  \tag{7} \]

where \( \{(\tilde{y}_i, \tilde{x}_i)\}_{i=1}^{m} \) are independent copies from the same distribution of \( D_n \). In practical implement, we split the sample into training data and testing data to calculate (7) approximatively. This is the idea of cross-validation (CV) approach, which has the additional advantage of enabling the testing of the model on new data. Let \( J \) as the number of folds in CV such that the number of observations in each block is \( \lfloor n/J \rfloor \), where \( \lfloor g \rfloor \) is the integer portion of \( g \). Denote \( A(j) = \{(j-1)M_n + 1, (j-1)M_n + 2, \ldots, jM_n\} \), \( |A(j)| \) the cardinality of \( A(j) \), and \( B(j) = \{1, 2, \ldots, n\} \cap A(j)^c \). The CV approach is based on a minimisation of

\[ CV_n(w) = \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in B(j)} \left( 1 - y_i x_i^T \tilde{\beta}^{[-j]}(w) \right)_+ , \]  \tag{8} \]

where

\[ \tilde{\beta}^{[-j]}(w) = \sum_{s=1}^{S_n} w_s \Pi_s \tilde{\beta}^{[-j]}_{(s)} \quad \text{and} \quad \tilde{\beta}^{[-j]}_{(s)} = \arg \min_{\beta_{(s)}} \frac{1}{|A(j)|} \sum_{i \in A(j)} (1 - y_i x_i^T \beta_{(s)})_+ . \]

In the above formulas, \( \tilde{\beta}^{[-j]}(w) \) is the coefficient estimate trained without the \( j \)th fold data. Considering the number of training data is \( |A(j)| \), we achieve it by minimizing the average hinge loss. Though \( \tilde{\beta}^{[-j]}_{(s)} \) will work well in the training data, we do not know how the integrated estimate \( \tilde{\beta}^{[-j]}(w) \) performs in the testing data so that the hinge loss is calculated again in the \( j \)th fold data which contains \( |B(j)| \) observations. Therewith, we iterate these \( J \) folds to get the average loss and obtain the weights by minimizing this loss.

Although model averaging often has the advantages of better precision and reduced bias compared to the estimate from a single "best" model, it also imposes a heavy burden on computation, especially with high-dimensional data. With \( p \) covariates, there are \( 2^p - 1 \) potential candidate models, and when \( p \) is large, it is virtually impossible to take all models into account. We reconcile this difficulty by using a screening procedure to screen out the uninformative covariates prior to combining models. Our model screening procedure entails sorting the covariates by SVM under \( L_1 \) penalty. The idea of model screening based on solution path is also adopted in Zhang et al. [2020]. The details are described in Algorithm 1.

With the tuning parameter becoming larger, the covariates are sorted in the inverse order as the coefficients being zero. We then construct the candidate models by adding one at a time an extra
covariate that survives the screening procedure to form a sequence of nested models. Lastly, we
compute the weights by the CV criterion in (8) and combine the candidate models using these
weights as in (4). We find that the results are generally insensitive to the number of folds, and we
suggest choosing \( J = 5 \) or \( J = 10 \). After computing the model average we follow the steps of
Algorithm 2 to compute the prediction of \( y \).

**Algorithm 1 Calculate \( \hat{w} \)**

Require: \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \)  #The first element of \( x_i \) is 1.

**Step 1: Pre-screening**

Require: \( a, b, L, C_0 = \emptyset, I = \text{list}() \)  # The \( l_1 \) penalty parameter \( \lambda \) is in \([a, b]\) and \( a \) is recommended to be 0.001.

\( J, S_n, M_n = \lfloor n/J \rfloor \)  # \( J \) is the number of folds of CV and \( S_n \) is the number of candidate models.

for \( l = 0 \) to \( L \)

\( \lambda = a + l(b - a) / L \)

\( \tilde{\beta} = \operatorname{arg min}_{\beta} n^{-1} \sum_{i=1}^{n} (1 - y_i x_i^T \beta)_+ + \lambda \| \beta^+ \|_1 \)

\( C_{l+1} = \{ i | \tilde{\beta}^+_i = 0, i = 1, 2, \ldots, p \} \)

\( I = I.extend(\text{list}(C_{l+1} \setminus C_l)) \)

end for

\( I = I.extend(0) \)

\( I = I.reverse() \)

**Step 2: Solve candidate models**

for \( j = 1 \) to \( J \)

\( A(j) = \{(j-1)M_n + 1, (j-1)M_n + 2, \ldots, jM_n\} \)

\( B(j) = \{1, 2, \ldots, n\} \cap \overline{A(j)} \)

for \( s = 1 \) to \( S_n \)

\( \tilde{\beta}_{(s)} = \operatorname{arg min}_{\beta(s)} \left\{ (n - M_n)^{-1} \sum_{i \in A(j)} (1 - y_i x_{i, I[0:s]}^T \beta(s))_+ + 2(n - M_n))^{-1} \| \beta^+_s \|_2^2 \right\} \)

\( \hat{\beta}_{(s)}^{(j)} = \Pi_s \tilde{\beta}_{(s)} \)  #\( \Pi_s \) is the selection matrix

end for

end for

**Step 3: Calculate \( \hat{w} \)**

\( \hat{w} = \operatorname{arg min}_w n^{-1} \sum_{j=1}^{J} \sum_{i \in B(j)} \left( 1 - y_i x_i^T \sum_{s=1}^{S_n} w_s \hat{\beta}_{(s)}^{(j)} \right)_+ \)

return \( \hat{w} \)
Algorithm 2 Prediction

Require: \( \hat{w}, x_{new}, I, S_n, D = \{ (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \} \)

Solve all the candidate models based on \( I, S_n \) and \( D \) to obtain \( \{ \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{S_n} \} \)

Return \( \text{sign}(x_{new}^T \sum_{s=1}^{S_n} \hat{w}_s \hat{\beta}_s) \)

3 Theoretical justification

3.1 Notions and conditions

This section is devoted to an investigation of the theoretical properties of the proposed model averaging strategy. First, let us introduce some notations. Denote \( L_s(\beta) = \mathbb{E}(1 - yx^T_s \beta)_{+} \), \( J_s(\beta) = -\mathbb{E}\left( (1 - yx^T_s \beta)_{+} \right) \) and \( H_s(\beta) = \mathbb{E}\{\delta(1 - yx^T_s \beta) x_s x_s^T \} \), where \( \delta(\cdot) \) is the indicator function and \( \delta(\cdot) \) is the Dirac delta function, \( s = 1, 2, ..., S_n \). Koo et al. [2008] showed that under some regularity conditions, \( J_s(\beta) \) and \( H_s(\beta) \) behave like the gradient and Hessian matrix of \( L_s(\beta) \), respectively. In addition, we let \( f_+ \) and \( f_- \) be the densities of \( x^+ \in \mathbb{R}^p \) conditioning on \( y = 1 \) and \( y = -1 \) respectively. Note that the dimension of the covariates \( x_{(s):i} \) used in the \( s^{th} \) candidate model is \( p_s + 1 \). We therefore write \( p_{\max} = \max_{1 \leq s \leq S_n} p_s + 1 \) as the dimension of the largest candidate model.

Our proofs of results require the following conditions:

Condition 1

\((C1)\) \( f_+ \) and \( f_- \) are continuous and have the same common support in \( \mathbb{R}^p \).

\((C2)\) There is a constant \( C_1 > 0 \) such that \( \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |x_{ij}| < C_1 \).

\((C3)\) For \( s = 1, 2, ..., S_n \), the \( s^{th} \) candidate model has a unique “quasi-true” parameter \( \beta^*_s \) and there exists a constant \( C_2 > 0 \) such that \( \|\beta^*_s\| \leq C_2 \sqrt{p_n} \).

\((C4)\) The densities of \( x^T_s \beta^*_s \) conditioning on \( y = 1 \) and \( y = -1 \) are uniformly bounded away from zero and have a uniform upper bound \( C_3 \) (a positive constant) at the neighborhood of \( x^T_s \beta^*_s = 1 \) and \( x^T_s \beta^*_s = -1 \), respectively.

\((C5)\) \( p_{\max} = O(n^\kappa) \) for some constant \( \kappa \in (0, 1/5) \).

\((C6)\) \( S_n = O(n^\tau) \) for some constant \( \tau \in (0, 1 - 2\kappa) \).

\((C7)\) For \( s = 1, 2, ..., S_n \), there exists a positive constants \( c_0 \) such that \( \min_{1 \leq s \leq S_n} \lambda_{\min}\{H_s(\beta^*_s)\} \geq c_0 \) where \( \lambda_{\min}(\cdot) \) is the smallest eigenvalue of a matrix.
These conditions are similar to Conditions (A1)–(A8) of Zhang et al. [2016a], but there are some differences. Condition (C1) ensures that $J_s(\beta_{(s)})$ and $H_s(\beta_{(s)})$ are well-defined as in Koo et al. [2008]. Condition (C2) is required to measure the volume of $x_{(s),i}$ which is commonly used in studies of high-dimensional regression [e.g., Wang et al., 2012, Lee et al., 2014]. Condition (C3) is a mild condition to guarantee the “quasi-true” parameter is existed and similar conditions can be seen in Zhang et al. [2016b]. Condition (C4) assumes that as the sample size increases, there is sufficient information This condition is also required for model selection consistency of non-convex penalized SVM in high dimensions [Zhang et al., 2014]. Conditions (C5) and (C6) allow the dimension of variables of all candidate models and the amount of candidate models to diverge with sample size but the volume is restrictive. Especially, $p_{\text{max}} \leq p$ and we do not put any restrictions on $p$, so the Algorithm 1 is able deal with the scenario that $p > n$ and $p/n \to \infty$. Condition (C7) requires the Hessian matrix is well-behaved and is nonsingular when $\beta_{(s)}$ is around the “quasi-true” parameter.

### 3.2 Theoretical results

**Lemma 1** Under Conditions (C1)-(C7) and $\lambda_n = O(\sqrt{\log(p_{\text{max}})/n})$, we have

$$\max_{1 \leq s \leq S_n} \| \hat{\beta}_{(s)} - \beta^*_s \| = O_p \left( \sqrt{\frac{p_{\text{max}} \log(p_{\text{max}})}{n}} \right).$$

(9)

Further, we have

$$\max_{1 \leq j \leq J} \max_{1 \leq s \leq S_n} \| \tilde{\beta}_{(s)}^{[-j]} - \beta^*_s \| = O_p \left( \sqrt{\frac{p_{\text{max}} \log(p_{\text{max}})}{n}} \right).$$

(10)

This lemma is similar to Zhang et al. [2016a] and the both convergence about $\hat{\beta}_{(s)}$ are uniformly for $s$, but the conditions we used are different such as Conditions (C5) and (C6). Besides, the results shown in Zhang et al. [2014] are not converging in uniformly. In Zhang et al. [2016a], the right term of (9) is $O_p(\sqrt{n^{-1}p_s \log(p)})$ but we introduce the largest dimension among all candidate models, $p_{\text{max}}(\leq p)$. Mainly, we split the explanatory variables into different candidate models and estimate the parameters independently, so we just need to explore the relation between $n$ and $p_{\text{max}}$ rather than $p$.

**Condition 2** There exists a constant $\xi_0$ such that

$$\liminf_{n \to \infty} \xi_n \geq \xi_0 > 0,$$

(11)

where $\xi_n = \inf_{w \in W} R_n(w)$. 

8
This condition can be satisfied easily because the hinge loss is nonnegative and usually the data we
used cannot be separated clearly by a linear hyperplane. The similar conditions are often used in
papers about model averaging, such as Condition (A.6) of Hansen and Racine [2012] and Condition
(A3) of Ando and Li [2017].

**Theorem 1**  If Conditions 1 and 2 hold, then

\[
\frac{R_n(\hat{w})}{\inf_{w \in \mathcal{W}} R_n(w)} \to 1
\]  \hspace{1cm} (12)

in probability, where \( \hat{w} \) is the optimal solution from (8).

This theorem shows the SVMMA estimator is asymptotically optimal in the sense that its hinge
risk is asymptotically identical to that obtained from the infeasible best possible model averaging
estimator.

## 4 A simulation study

The purpose of this section is to examine, via a simulation study, the performance of the SVMMA
estimator under sample sizes commonly encountered in practice. The following estimators are in-
cluded in the comparison:

- The SVM information criterion (SVMICL) and a modified information criterion for high di-
dimensional case (SVMICH) estimators introduced by Zhang et al. [2016a]. They consist of a
hinge loss and a penalty on the dimension of candidate model, which are the derivatives of
BIC.

- The smoothed SVMICL (SCL) and smoothed SVMICH (SCH) model averaging methods.
They are generated from smoothing SVMICL and SVMICH by exponential function, respec-
tively.

- The estimators based on the Bagging [Breiman, 1996] and Adaboosting [Freund and Schapire,
1997] ensembling learning methods. They are the popular methods in machine learning, but
the weight choice criteria are different from model averaging.

- Uniform weight form. Assigning all the candidate models with a equal weight is a benchmark
in comparison.
The SVMICL and SVMICH are model selection criteria developed by Zhang et al. [2016a], defined as

\[
\text{SVMICL}_s = \sum_{i=1}^{n} (1 - y_i \mathbf{x}_{(s)},_i \hat{\beta}_{(s)})_+ + p_s \log(n)
\]

and

\[
\text{SVMICH}_s = \sum_{i=1}^{n} (1 - y_i \mathbf{x}_{(s)},_i \hat{\beta}_{(s)})_+ + \sqrt{\log(n)} p_s \log(n)
\]

respectively. By the SVMICL or SVMICH criterion, the model with the smallest value of the criterion is selected. The SCL and SCH are model averaging methods that use the smoothed versions of the SVMICH and SVMICH as weights, constructed analogously to the SAIC and SBIC model weighting methods commonly in the model averaging literature. Specifically, the SCL and SCH weights assigned to the \(s\)th model are given by

\[
\text{SCL}_s = \exp \left( \frac{-\text{SVMICL}_s}{n} \right) \sum_{s=1}^{S_n} \exp \left( \frac{-\text{SVMICL}_s}{n} \right), \quad s = 1, 2, \ldots, S_n.
\]

and

\[
\text{SCH}_s = \exp \left( \frac{-\text{SVMICH}_s}{n} \right) \sum_{s=1}^{S_n} \exp \left( \frac{-\text{SVMICH}_s}{n} \right), \quad s = 1, 2, \ldots, S_n.
\]

respectively. Bagging aims to combine the prediction from base learners by averaging but voting rule is often used for classification problem. Adaboosting is a type of boosting [Schapire, 1990] which is also developed for classification. Different from model averaging, there is no constraints to restrict the weights that used to aggregate base learners for adaboosting. Overall, bagging focuses on reducing the variance but adaboosting aims at reducing the bias.

We evaluate the performance of the methods using the following normalised hinge loss (NHL) and error rate on prediction (ER):

\[
\text{NHL} = \frac{1}{D} \sum_{d=1}^{D} \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left( 1 - \hat{y}_i^{(d)} \mathbf{x}_i^{(d)^T} (\hat{\beta}^{(d)}(\hat{w}^{(d)}))_+ \right),
\]

\[
\text{ER} = \frac{1}{Dn_{\text{test}}} \sum_{d=1}^{D} \sum_{i=1}^{n_{\text{test}}} 1(\hat{y}_i^{(d)} \neq \hat{y}_i^{(d)}),
\]

where \(\hat{\beta}^{(d)}(\hat{w}^{(d)}) = \sum_{s=1}^{S_n} \hat{w}_s^{(d)} \hat{\beta}_s^{(d)}\) is the coefficient estimate in the \(d\)th repetition and \(\hat{w}_s^{(d)}\) takes 1 or 0 for selection methods, \(\hat{\beta}_s^{(d)}\) are calculated based on training set, but \(\hat{y}_i\) and \(\hat{x}_i\) are from testing set and \(n_{\text{test}}\) is the size of testing set, \(\hat{y}_i^{(d)}\) represents the predicted value based of different methods. bagging and adaboosting just provided the outcome rather than the estimation of \(\beta\), they are omitted in NHL comparison.
4.1 Simulation designs

We consider two different data generating processes (DGPs) used in Zhang et al. [2016a]. The first is adopted from Fisher’s linear discriminant analysis and the second is related to Probit regression.

**DGP 1:** \[ \Pr(Y = 1) = \Pr(Y = -1) = 0.5, \quad x|(Y = 1) \sim N(\mu, \Sigma), \quad x|(Y = -1) \sim N(-\mu, \Sigma), \quad \mu = (0.6, \cdots, 0.6, 0, \ldots, 0)^T \in \mathbb{R}^p, \quad \Sigma = (\sigma_{ij})_{p \times p}, \quad \sigma_{ii} = 1 \text{ for } i = 1, 2, \ldots, p, \sigma_{ij} = 0.2 \text{ for } 1 \leq i \neq j \leq p \text{ and } q \text{ is a tuning parameter to measure the sparsity.} \]

**DGP 2:** \[ \Pr(Y = 1) = \Phi(x^T \beta), \quad x \sim N(0_p, \Sigma), \quad \beta = (2, \cdots, 2, 0, \ldots, 0)^T, \quad \Sigma = (\sigma_{ij})_{p \times p} \text{ with } \sigma_{ij} = 1 \text{ for } i = 1, 2, \ldots, p \text{ and } \sigma_{ij} = 0.4^{|i-j|} \text{ for } 1 \leq i \neq j \leq p, \text{ where } \Phi(\cdot) \text{ is the cumulative distribution function (CDF) of standard normal distribution and } q \text{ is a tuning parameter to measure the sparsity.} \]

To evaluate the method’s robustness to model misspecification due to missing covariates, we consider the following two scenarios:

- **Scenario 1:** the model correctly represents the underlying DGP and no relevant covariate has been omitted from the model;
- **Scenario 2:** some relevant covariates are omitted from the DGP

To generate these two situations, we set \( q = 4 \) and use all the covariates to training for the former and we set \( q = 5 \) and omit the \( q^{th} \) variable before training for the latter. Our implementation of the proposed model averaging method is based on 5-fold CV. We consider sample sizes of \( n = 100, 200, 300 \) and 400 for the training data, and \( n_{\text{test}} = 10000 \) for the testing data. All results reported are based on \( D = 200 \) replications and set the dimension \( p = 1000 \). In each case, we generate 100 candidate models according to Algorithm 1. These models are also the base learners for bagging and adaboosting on equal footing. To eliminate the influence of base learners’ order, they are randomly chosen one by one in each iteration for Adaboosting.

4.2 Simulation results

The simulation results corresponding to Scenarios 1 and 2 are shown in Figures 1-2 and Figures 3-4 respectively. Our results show that in terms of NHL, the SVMMA approach delivers by far the best estimates, often by a large margin, in all parts of the parameter space. On the other hand, neither the SVMICH nor the SVMICL model selection methods yield very good estimates, both

---

11

---

\(^{2}\)5-fold and 10-fold CV all work well and the number of folds has small influence on the performs according to the experiments
being dominated by their corresponding model averaging counterparts, the SCH and SCL estimators, of which the SCL is always the better estimator of the two. Although the UNIF method is a distant second best compared to the SVMMA estimator, it produces superior estimates to the SCL estimator over a large part of the parameter space. Exceptions occur when \( n \) is very large, where the SCL can sometimes yield slightly better results. Generally speaking, these findings also hold when the methods are assessed in terms of ER, in which the comparisons also include the ensembling learning methods of Adaboosting and Bagging. In all cases, Adaboosting performs rather poorly, at a level comparable to the SVMICH estimator. On the other hand, Bagging can sometimes deliver marginally better estimates than the SVMMA estimator when \( n \) is small. That being said, bagging is an inferior strategy compared to most other methods when \( n \) is moderate to large, while SVMMA offers more stable performance and is the preferred estimator in the overwhelming majority of cases. The rather erratic performance of bagging can be explained by noting that a good bagging learner depends on many good and independent based learners [Zhou, 2012]. However, in our case, every two candidate models have at least one common covariate, and as the training sample size grows the diversity of their outcomes will decrease. On the other hand, in the case of the SVMMA estimator, the weights being constrained by \( \sum_{s=1}^{S} w_s = 1 \) reflect the strength of the different models, and hence the better base learners will be given higher weights, yielding improved accuracy. As well, model average estimators with a weight constraint may be interpreted as a shrinkage estimator that balances between bias and variance [Jagannathan and Ma, 2003]. Figures 5 and 6 show that as the number of base learners increases, bagging’s performance will improve, but for the SVMMA method to achieve the same accuracy, SVMMA needs fewer base learners (i.e., candidate models) than bagging. This latter feature is another advantage of the proposed model averaging approach.

Overall, the performances of all methods in scenario 2 are similar as the results in scenario 1 but there are also some difference especially in DGP 2. We can also find the curves in Figure 2(b) is lower than those in Figure 4(b). This phenomenon is normal because the train data loses a vital covariate in scenario 2 so that the models cannot learn completely. This phenomenon mainly lies in overfitting because adaboosting aims at reduce the error rate in train data and some noisy information will be picked up. Fortunately, the weights used in SVMMA is calculated by CV criterion which will relieve the appearance of overfitting.

The rank of these lines has an obvious change in DGP 2 of scenario 1, but the line of SVMMA still stays below and UNIF follows, besides, the lines representing SVMICL and SVMICH perform poorest. This says SVMMA can keep stable in different model settings. Compared with the setting of DGP 1, the data from DGP 2 is nonlinear separable, so just using one linear SVM will lead to a bad result that is why SVMICL and SVMICH perform poor. On the other hand, the ability of the combined model to deal with non-linear separable problems is promoted by the form of weighted
combination, so the averaging methods usually perform very well.

Figure 1: The NHL in scenario 1.

Figure 2: The ER in scenario 1.
Figure 3: The NHL in scenario 2.

Figure 4: The ER in scenario 2.
5 Real data examples

In this section, we apply the proposed method to four real data sets. In all our analysis, we standardise the covariates and randomly split the data into a training subset and a testing subset, containing \( n_0 \) and \( n - n_0 \) observations respectively. The proportion of training set is around the ten uniform division points from 40% to 80%. Our computation of the SVMMA weights is based on five-fold CV. To calculate the NHL and ER, we repeat the process of splitting the data 200 times and each time we compute the SVMMA estimator. Our first example is based on the “Sonar, Mines versus Rocks” data of KEEL [Alcalá-Fdez et al., 2011], downloaded from https://sci2s.ugr.es/keel/dataset.php?cod=85. The data comprise signal infor-
mation obtained from different aspect angles, spanning 90 degrees for mines and 180 degrees for rocks. Each pattern is a set of 60 numbers in the range of 0.0 to 1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time.

The output attribute contains the letter R if the object is a rock and M if it is a mine (metal cylinder). In sum, it is a two classification data containing 60 features and 208 observations. More description can be found in Gorman and Sejnowski [1988]. We generate 20 candidate models as Algorithm 1 to implement these methods.

![The lines of NHL](image1)

![The lines of ER](image2)

Figure 7: The performance in Sonar.

![Learning curves](image3)

Figure 8: The learning curves in Sonar; the proportion of training set is 80% and testing set is 20%.

The second data set we used is from UCI[Du\ and Graff, 2017], called “Statlog(Heart)” which can be download in http://archive.ics.uci.edu/ml/datasets. The sample size is 270, with 13 features. It is a heart disease database similar to a the database, “Heart Disease
databases”, already present in the repository but in a slightly different form. The task is to detect the absence or presence of heart disease. We generate 8 candidate models as Algorithm 1 to implement these methods.

The third data set we used also is from UCI, called “Ionosphere”, which can be download in http://archive.ics.uci.edu/ml/datasets/Ionosphere. The sample size is 351, with 33 features. This radar data was collected by a system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. “Good” radar returns are those showing evidence of some type of structure in the ionosphere. “Bad” returns are those that do not; their signals pass through the ionosphere. The more description can be found in Sigillito et al. [1989]. We generate 30 candidate models as Algorithm 1 to implement these methods.
The fourth data set we used, also from CUI, is called “LSVT Voice Rehabilitation”, which can be downloaded at [http://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation](http://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation). The sample size is 126, with 309 features. This dataset is composed of a range of biomedical speech signal processing algorithms from 14 people who have been diagnosed with Parkinson’s disease undergoing LSVT (a program assisting voice rehabilitation). The aim is to use the 309 features to predict the binary response. This is also a binary classification and more description can be found in Tsanas et al. [2014]. We generate 30 candidate models as Algorithm 1 to implement these methods.
Figures 7– 14 show the SVMMA performs well. Besides, the phenomenon, that the curves of NHL of SVMMA is very closed to 1, is corresponding to Theorem 1. Usually, model averaging methods are better than selection methods and especially smoothed SVMICL and SVMICH are better than themselves. The reason is that model selection methods usually choose only one model to apply that will lead to a big risk to make mistakes, but model averaging methods utilize more models to predict and can reduce the mistaken risk well. In order to compare the learning ability of two special methodologies, i.e., model averaging and ensemble learning, we draw the learning curves in Figures 8, 10, 12 and 14. In the figures, the $x$-axis presents the amount of candidate models used for training integrate models and the $y$-axis presents the error rate of classification for training data or testing data. The lines of XX-Train is the error rate of XX method in training data and the lines of XX-Test is the error rate of XX method in testing data, where XX represent MA, Bag or
Ada. Usually, the model will get a smaller error rate in training data than that in testing data because it is trained by the training data. If the overfitting occurs, the lines of testing data will increase. We often use the stable error rate of testing data to express the learning ability of the model. So the model is better if it achieves the stable state earlier and a smaller error value. Comparing with Bagging and Adaboosting, SVMMA needs less base learners to achieve good classification results. So optimal model averaging is a good tool to promote the SVM’s performance.

6 Conclusion

This paper proposes an optimal model averaging technology to reduce the risk generating by mistakenly choosing only one bad candidate model for applying and to promote the prediction ability for SVM model. Not only the weights choice criterion is provided, but also the MA estimator is proved to be asymptotically optimal in the sense of achieving the lowest hinge loss among all possible combination. In implementation, we provide a manner to set up the candidate model set by $L_1$ penalty. Besides, the optimal model averaging estimator performs best, compared with several other selection or averaging methods, through the simulation and four real data sets. Last, it is also an interesting work in the future to develop the optimal model averaging method for multi-kernel SVM models since how to choosing the kernel is uncertain in practice.
7 Appendix

This section provides the proof of Theorem 1. All limiting processes below correspond to \( n \rightarrow \infty \) unless stated otherwise.

7.1 Proof of Lemma 1

Part of this proof follows from Zhang et al. [2016a], but there are some differences and the conclusion is also different from that of Zhang et al. [2016a].

We will prove (9) first. Recall that \( \hat{\beta}(s) = \arg \min_{\beta(s)} \{ n^{-1} \sum_{i=1}^{n} (1 - y_{i} x_{(s),i}^{T} \beta(s)) + 2^{-1} \lambda_n \| \beta(s) \|^2 \} \).

We will show that, for any \( 0 < \eta < 1 \), there exist large constants \( \Delta > 0 \) and \( N \), when \( n > N \),

\[
\Pr \left\{ \min_{1 \leq s \leq S_n} \inf_{\| u(s) \| = \Delta} \left\{ l_s \left( \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right) - l_s(\beta(s)) \right\} > 0 \right\} > 1 - \eta, \tag{19}
\]

where \( u(s) \in \mathcal{R}^{p_x} \) and \( l_s(\beta(s)) = n^{-1} \sum_{i=1}^{n} (1 - y_{i} x_{(s),i}^{T} \beta(s)) + 2^{-1} \lambda_n \| \beta(s) \|^2 \). As the hinge loss is convex, this implies that with probability \( 1 - \eta \), \( \max_{1 \leq s \leq S_n} \| \hat{\beta}(s) - \beta(s) \| \leq \Delta \sqrt{n^{-1} p_{\max} \log(p_{\max})} \).

Hence equation (9) in Lemma 1 holds.

Note that \( l_s \left( \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right) - l_s(\beta(s)) \) can be expressed as

\[
l_s \left( \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right) - l_s(\beta(s)) = n^{-1} \sum_{i=1}^{n} \left\{ \left( 1 - y_{i} x_{(s),i}^{T} \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right)_{+} - \left( 1 - y_{i} x_{(s),i}^{T} \beta(s) \right)_{+} \right\} + 2^{-1} \lambda_n \left\| \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right\|^2 - 2^{-1} \lambda_n \| \beta(s) \|^2. \tag{20}
\]

It is readily shown that

\[
\max_{1 \leq s \leq S_n} \sup_{\| u(s) \| = \Delta} \left\| \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \right\|^2 - \| \beta(s) \|^2 \leq \max_{1 \leq s \leq S_n} \sup_{\| u(s) \| = \Delta} \left( \| \beta(s) + \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \| + \| \beta(s) \| \right) \| \sqrt{n^{-1} p_{\max} \log(p_{\max})} u(s) \|
\leq 2\Delta C_2 p_{\max} \sqrt{n^{-1} \log(p_{\max})} + \Delta n^{-1} p_{\max} \log(p_{\max})
= O(\Delta p_{\max} \sqrt{n^{-1} \log(p_{\max})}), \tag{21}
\]

where the last inequality is obtained from Condition (C3). So the volume of difference of penalty terms in (20) is \( O(\Delta \lambda_n p_{\max} \sqrt{n^{-1} \log(p_{\max})}) \).

21
Denote 
\[ g_{s,i}(u_{(s)}) = \left(1 - y_i x_{(s),i}^T \beta_{(s)}^* + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u_{(s)}\right)_{+} - \left(1 - y_i x_{(s),i}^T \beta_{(s)}^* \right)_{+} + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i} u_{(s)} 1_{(1 - y_i x_{(s),i}^T \beta_{(s)}^* \geq 0)} 
- \mathbb{E}\left[\left(1 - y_i x_{(s),i}^T \beta_{(s)}^* + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u_{(s)}\right)_{+}\right] + \mathbb{E}\left[\left(1 - y_i x_{(s),i}^T \beta_{(s)}^* \right)_{+}\right]. \]

It can be verified that \( \mathbb{E}[g_{s,i}(u)] = 0 \), \( s = 1, 2, \ldots, S_n \) by the definition of \( \beta_{(s)}^* \) and \( J_{(s)}(\beta_{(s)}^*) = 0 \). Note that (20) can be further decomposed as 
\[ n^{-1} \sum_{i=1}^{n} \left\{ \left(1 - y_i x_{(s),i}^T \beta_{(s)}^* + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u_{(s)}\right)_{+} - \left(1 - y_i x_{(s),i}^T \beta_{(s)}^* \right)_{+} \right\} = n^{-1}(A_{s,n} + B_{s,n}), \]
where 
\[ A_{s,n} = \sum_{i=1}^{n} g_{s,i}(u_{(s)}) \]
and 
\[ B_{s,n} = \sum_{i=1}^{n} \left[ - \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i} u_{(s)} 1_{(1 - y_i x_{(s),i}^T \beta_{(s)}^* \geq 0)} + \mathbb{E}\left\{\left(1 - y_i x_{(s),i}^T \beta_{(s)}^* + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u_{(s)}\right)_{+}\right\} - \mathbb{E}\left\{\left(1 - y_i x_{(s),i}^T \beta_{(s)}^* \right)_{+}\right\}\right]. \]

The remainder of the proof consists of three steps. In Step 1, we demonstrate that 
\[ \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \triangle} |A_{s,n}| = \triangle^{3/2} p_{\text{max}} o_p(1). \] (23)

In Step 2, it is shown that \( \min_{1 \leq s \leq S_n} \inf_{\|u_{(s)}\| = \triangle} B_{s,n} \) dominates the terms of order \( \triangle^{3/2} p_{\text{max}} o_p(1) \) and is larger than zero. In Step 3, we use the results from the previous steps to prove (19).

Step 1: We use the covering number introduced by Vaart and Wellner [1996] to prove the uniform rate in (23). It suffices to show, for any \( \epsilon > 0 \), that 
\[ \Pr\left( \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \triangle} \frac{1}{\sqrt{p_s}} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) \right| > \triangle^{3/2} \epsilon \right) \rightarrow 0. \] (24)

Note that the hinge loss satisfies the Lipschitz condition and \( \max_{1 \leq i \leq n} \|x_{(s),i}\| \leq C_1 \sqrt{p_s} \), \( \max_{1 \leq i \leq n} \mathbb{E}\|x_{(s),i}\| \leq C_1 \sqrt{p_s} \) from Condition (C2). It is readily shown that 
\[ |g_{s,i}(u_{(s)})| \leq 3\triangle \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x_{(s),i}\| \mathbb{E}\|x_{(s),i}\| \]

22
\[ \leq 3C_1 \triangle p_{\max} \sqrt{n^{-1} \log(p_{\max})} \]  

(25)  

and thus \( \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \triangle} p_s^{-1} g_{s,i}(u_{(s)}) = o(1) \) by Condition (C5). By Lemma 2.5 of van de Geer [2000], the ball \( \{u_{(s)} : \|u_{(s)}\| \leq \triangle\} \) in \( \mathbb{R}^{p_s} \) can be covered by \( N_s \) balls with radius \( \zeta_s \), where \( N_s \leq \{(4 \triangle + \zeta_s)/\zeta_s\}^{p_s+1} \). Denote \( u_{(s)}^1, \ldots, u_{(s)}^{N_s} \) as the centers of the \( N_s \) balls, let \( \zeta_s = (nM_1)^{-1} p_s \) (for some large constant \( M_1 > 0 \)) and denote \( U_{i}^{(k)} = \{u_{(s)} : \|u_{(s)} - u_{(s)}^k\| \leq \zeta_s \& \|u_{(s)}\| = \triangle\} \). For any \( \epsilon > 0 \), we have

\[
\begin{align*}
\max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} \sup_{u_{(s)} \in U_{i}^{(k)}} p_s^{-1} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) - \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) \right| \\
\leq \max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} \sup_{u_{(s)} \in U_{i}^{(k)}} p_s^{-1} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) - g_{s,i}(u_{(s)}^k) \right| \\
\leq \max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} \sup_{u_{(s)} \in U_{i}^{(k)}} n p_s^{-1} \left\{ 2 \sqrt{n^{-1} p_{\max} \log(p_{\max})} \max_{1 \leq i \leq n} \max_{1 \leq s \leq S_n} \|x_{(s),i}\| \|\|u_{(s)} - u_{(s)}^k\| \right\} \\
+ \sqrt{n^{-1} p_{\max} \log(p_{\max})} \|u_{(s)} - u_{(s)}^k\| \|\|x_{(s),i}\| \right\} \\
\leq 3 \triangle n p_s^{-1} \sqrt{n^{-1} p_{\max} \log(p_{\max})} \max_{1 \leq i \leq n} \max_{1 \leq s \leq S_n} \|x_{(s),i}\| \|\|u_{(s)} - u_{(s)}^k\| \\
\leq 3C_1 M_1^{-1} \triangle p_{\max} \sqrt{n^{-1} \log(p_{\max})} \\
= o(\triangle^{3/2} p_{\min}\epsilon/2), \\
\end{align*}
\]

(26)  

where the last inequality arises from Condition (C5). From (26), it can be shown that

\[
\begin{align*}
\Pr \left( \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \triangle} p_s^{-1} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) > \triangle^{3/2} \epsilon \right| \right) \\
\leq \Pr \left( \max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} \sup_{u_{(s)} \in U_{i}^{(k)}} p_s^{-1} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) - \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) \right| \\
+ \max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} p_s^{-1} \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) > \triangle^{3/2} \epsilon \right) \\
\leq \Pr \left( \max_{1 \leq s \leq S_n} \max_{1 \leq k \leq N_s} \sup_{u_{(s)} \in U_{i}^{(k)}} \left| \sum_{i=1}^{n} g_{s,i}(u_{(s)}) - \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) \right| > \triangle^{3/2} p_{\min}\epsilon/2 \right) \\
+ \sum_{s=1}^{S_n} \sum_{k=1}^{N_s} \Pr \left( \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) > \triangle^{3/2} p_{\min}\epsilon/2 \right) \\
= \sum_{s=1}^{S_n} \sum_{k=1}^{N_s} \Pr \left( \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) > \triangle^{3/2} p_{\min}\epsilon/2 \right) + o(1) \\
\end{align*}
\]

(27)  

and \( \sum_{i=1}^{n} g_{s,i}(u_{(s)}^k) \) is the sum of independent zero-mean random variables.
Note that when $1 - y_i x_{(s),i}^T \beta^*_s < \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s)$ and $\sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s) < 0$, or when $1 - y_i x_{(s),i}^T \beta^*_s > \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s)$ and $\sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s) > 0$,

$$
\left(1 - y_i x_{(s),i}^T \beta^*_s + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u(s)\right) - (1 - y_i x_{(s),i}^T \beta^*_s) + \\
+ \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s) 1(1 - y_i x_{(s),i}^T \beta^*_s) = 0.
$$

(28)

Furthermore, equation (28) holds when $|1 - y_i x_{(s),i}^T \beta^*_s| > \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta$ as $\sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta > \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u(s)$. Hence we can write

$$
\sum_{i=1}^{n} \mathbb{E}\{g^2_{i,j}(u_{(s)})\}
$$

(29)

$$
\leq \sum_{i=1}^{n} \mathbb{E}\left\{ \left|1 - y_i x_{(s),i}^T (\beta^*_s + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} u^k(s))\right| - (1 - y_i x_{(s),i}^T \beta^*_s) + \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u^k(s) 1(1 - y_i x_{(s),i}^T \beta^*_s) \right\}
\\
+ \left|\sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} y_i x_{(s),i}^T u^k(s)\right|^2 1\left(|1 - y_i x_{(s),i}^T \beta^*_s| \leq \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta\right)
\\
\leq \sum_{i=1}^{n} \mathbb{E}\left\{ \left|2 \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} x_{(s),i}^T u^k(s)\right|^2 1\left(|1 - y_i x_{(s),i}^T \beta^*_s| \leq \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta\right) \right\}.
$$

(30)

Now, for any $1 < m < \infty$, under Condition (C2), we have

$$
\left\{ \mathbb{E}\left( 2 \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} x_{(s),i}^T (u^k(s))^2 \right) \right\}^{1/m}
\\
\leq 4 \Delta^2 n^{-1} p_{\text{max}} \log(p_{\text{max}}) \left( \mathbb{E}\|x(s),i\|^2 \right)^{1/m}
\\
\leq 4 C_1^2 \Delta^2 n^{-1} p_s p_{\text{max}} \log(p_{\text{max}}).
$$

(31)

As well, by the bounded conditional density under Condition (C1), Condition (C4) and the fact that $\max_{1 \leq i \leq n} \|x(s),i\| \leq C_1 \sqrt{p_s}$, we have

$$
\text{Pr}\left(|1 - y_i x_{(s),i}^T \beta^*_s| \leq \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta\right)
\\
= \text{Pr}\left(\pm 1 - \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta \leq x_{(s),i}^T \beta^*_s \leq \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta + 1 | y_i = \pm 1\right)
\\
\leq 2 C_3 \sqrt{n^{-1} p_{\text{max}} \log(p_{\text{max}})} \max_{1 \leq i \leq n} \|x(s),i\|\Delta
\\
\leq 2 \Delta C_1 C_3 \sqrt{n^{-1} p_s p_{\text{max}} \log(p_{\text{max}})}.
$$

(32)
By (30)–(32), using the Holder’s inequality, for $1 < m, l < \infty$ and $1/m + 1/l = 1$, we can write
\[
\sum_{i=1}^{n} \mathbb{E}\{ g_{s,i}^2(u^k) \} \leq \sum_{i=1}^{n} \left\{ \mathbb{E}\left( 2\sqrt{n^{-1}p_{\max} \log(p_{\max})} y_i x_{(s),i}^T u^{(k)} \right)^{2m} \right\}^{1/m} \\
\times \left\{ \text{Pr}\left( |1 - y_i x_{(s),i}^T \beta^*_s| \leq \sqrt{n^{-1}p_{\max} \log(p_{\max}) \max_{1 \leq i \leq n} \| x_{(s),i} \| \triangle} \right) \right\}^{1/l} \leq 4C_1^2 \Delta^2 p_s p_{\max} \log(p_{\max}) \left( 2\Delta C_3 \sqrt{n^{-1}p_s p_{\max} \log(p_{\max})} \right)^{1/l}.
\]

Let $l \to 1$, we obtain
\[
\sum_{i=1}^{n} \mathbb{E}\{ g_{s,i}^2(u^k) \} \leq 8\Delta^3 C_3 n^{-1/2} p_s^{3/2} p_{\max}^{3/2} \log^{3/2}(p_{\max}). \tag{33}
\]

Finally, by Bernstein’s inequality and recognising (25) and (33), we can write
\[
\sum_{s=1}^{S_n} \sum_{k=1}^{N_s} \text{Pr}\left( \left| \sum_{i=1}^{n} g_{s,i}(u^k) \right| > \Delta^3/2 p_s \epsilon/2 \right) \\
\leq \sum_{s=1}^{S_n} \sum_{k=1}^{N_s} 2 \exp \left( -\frac{\Delta^3 p_s^2 \epsilon^2/4}{\sum_{i=1}^{n} \mathbb{E}\{ g_{s,i}^2(u^k) \} + 3\Delta^{5/2} C_1 p_s p_{\max} \sqrt{n^{-1} \log(p_{\max})} \epsilon/2} \right) \\
\leq \sum_{s=1}^{S_n} \left( \frac{4\Delta + (nM_1) p_s}{(nM_1) p_s} \right)^{p_s+1} \\
\times \exp \left( -\frac{\Delta^3 p_s^2 \epsilon^2/4}{8\Delta^3 C_3 n^{-1/2} p_s^{3/2} p_{\max}^{3/2} \log^{3/2}(p_{\max}) + 3\Delta^{5/2} C_1 p_s p_{\max} \sqrt{n^{-1} \log(p_{\max})} \epsilon/2} \right) = O(1) \exp \left\{ \log(S_n) + (p_{\max} + 1) \log(4\Delta n M_1 / p_{\min} + 1) - 32^{-1} C_1^{-3} C_3^{-1} \epsilon^2 n^{1/2} p_{\max}^{1/2} p_{\max}^{-3/2} \log^{-3/2}(p_{\max}) \right\} = o(1), \tag{34}
\]

where the limit is due to Conditions (C5) and (C6). The proof of (24) is completed by combining (27) and (34).

Step 2: Let us rewrite $B_{s,n}$ as $B_{s,n} \equiv B_{s,n1} + B_{s,n2}$, where
\[
B_{s,n1} = -\sum_{i=1}^{n} \sqrt{n^{-1}p_{\max} \log(p_{\max})} y_i x_{(s),i}^T u_s \mathbf{1} \left( 1 - y_i x_{(s),i}^T \beta^*_s \geq 0 \right),
\]
and
\[
B_{s,n2} = \mathbb{E}\left\{ (1 - y_i x_{(s),i}^T \beta^*_s) + \sqrt{n^{-1}p_{\max} \log(p_{\max})} u_s \right\}_+ - \mathbb{E}\left\{ (1 - y_i x_{(s),i}^T \beta^*_s) \right\}_+.
\]
To analyse $B_{s,n1}$, we observe that

$$
\left| \sum_{i=1}^{n} y_i x_{(s),i}^T u_{(s)} 1 \left( 1 - y_i x_{(s),i}^T \beta^*_s \geq 0 \right) \right| \\
= \left| \sum_{j=0}^{p_s} \sum_{i=1}^{n} y_i x_{(s),ij} u_{(s),j} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right| \\
\leq \left| \sum_{j=0}^{p_s} \max_{0 \leq j \leq p_s} \left| \sum_{i=1}^{n} y_i x_{(s),ij} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right| \right| \\
\leq \sqrt{p_s + 1} \Delta \max_{0 \leq j \leq p_s} \left| \sum_{i=1}^{n} y_i x_{(s),ij} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right|.
$$

By the definition of $J_s(\beta^*_s)$, note that $\mathbb{E} \left[ y_i x_{(s),ij} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right] = 0$ for $0 \leq j \leq p_s$. By Lemma 14.24 in Bühlmann and van de Geer [2011] (the Nemirovski moment inequality),

$$
\mathbb{E} \left\{ \max_{0 \leq j \leq p_s} \left| \sum_{i=1}^{n} y_i x_{(s),ij} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right| \right\} \\
\leq \sqrt{8 \log(2p_s + 2)} \mathbb{E} \left( \max_{1 \leq j \leq p_s+1} \sum_{i=1}^{n} y_i^2 x_{(s),ij}^2 \right)^{1/2} \\
\leq \sqrt{8 \log(2p_s + 2)} \sqrt{nC_1^2} \\
= O(\sqrt{n \log(p_s)}),
$$

where the last inequality is established by Condition (C2). Additionally, using Markov’s inequality and by (36), we obtain

$$
\max_{0 \leq j \leq p_s} \left| \sum_{i=1}^{n} y_i x_{(s),ij} 1 \left( 1 - y_i x_{(s),ij}^T \beta^*_s \geq 0 \right) \right| = O_p(\sqrt{n \log(p_s)}).
$$

Combining (35) and (37), we have

$$
\max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \Delta} |B_{s,n1}| \\
= \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \Delta} \left| \sum_{i=1}^{n} y_i x_{(s),i}^T u_{(s)} 1 \left( 1 - y_i x_{(s),i}^T \beta^*_s \geq 0 \right) \right| \\
= \sqrt{n^{-1} p_{\max} \log(p_{\max})} \max_{1 \leq s \leq S_n} \sup_{\|u_{(s)}\| = \Delta} \left| \sum_{i=1}^{n} y_i x_{(s),i}^T u_{(s)} 1 \left( 1 - y_i x_{(s),i}^T \beta^*_s \geq 0 \right) \right|.
$$
\[ \leq \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} \Delta O_p \left\{ \sqrt{p_{\text{max}}} + 1 \sqrt{n \log(p_{\text{max}})} \right\} \]
= \(O_p(\Delta p_{\text{max}} \log(p_{\text{max}}))\). \quad \text{(38)}

Turning to \(B_{s,n^2}\), by Taylor expansion of the hinge loss at \(\beta^*_s\) and Condition (C7), we have

\[
\min_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} B_{s,n^2} = \min_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} \sum_{i=1}^{n} \left[ \mathbb{E} \left\{ (1 - y_i \mathbf{x}_i^T (\beta^*_s + \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u_s)) \right\} - \mathbb{E} \left\{ (1 - y_i \mathbf{x}_i^T \beta^*_s) \right\} \right] \geq 2^{-1} \Delta^2 c_0 p_{\text{max}} \log(p_{\text{max}}), \quad \text{(39)}
\]

for some \(0 < t < 1\). From Koo et al. [2008], under Conditions (C5) and (C6), \(H_s(\beta_s)\) is element-wise continuous at \(\beta^*_s\). Hence

\[ H_s(\beta^*_s + t \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u_s) = H_s(\beta^*_s) + o(1). \]

It can be readily shown by (22), (38), (39) and Condition (C5) that when \(\Delta\) is sufficiently large, \(2^{-1} \Delta^2 c_0 p_{\text{max}} \log(p_{\text{max}})(> 0)\) dominates other terms in \(B_{s,n}\). This completes the proof of Step 2.

**Step 3:** Combining (21), (24), (38) and (39), when \(n\) and \(\Delta\) are sufficiently large, we have

\[
\max_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} \left\{ l_s (\beta^*_s + \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u_s) - l_s(\beta^*_s) \right\} 
\]
= \[
\max_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} \left\{ n^{-1}(A_{s,n} + B_{s,n}) + 2^{-1} \lambda_n \left\| \beta^*_s + \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u^+_s \right\|^2 - 2^{-1} \lambda_n \left\| \beta^*_s \right\|^2 \right\} 
\]
\[
\geq \max_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} \left\{ n^{-1}B_{s,n^2} - n^{-1}|A_{s,n}| - 2^{-1} \lambda_n \left\| \beta^*_s + \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u^+_s \right\|^2 - \left\| \beta^*_s \right\|^2 \right\} 
\]
\[
\geq \min_{1 \leq s \leq S_n} \inf_{\|u_s\| = \Delta} \left\{ n^{-1}B_{s,n^2} - \max_{1 \leq s \leq S_n} \sup_{\|u_s\| = \Delta} n^{-1}|B_{s,n^1}| - \max_{1 \leq s \leq S_n} \sup_{\|u_s\| = \Delta} n^{-1}|A_{s,n}| - 2^{-1} \lambda_n \max_{1 \leq s \leq S_n} \sup_{\|u_s\| = \Delta} \left\| \beta^*_s + \sqrt{n^{-1}p_{\text{max}} \log(p_{\text{max}})} u^+_s \right\|^2 - \left\| \beta^*_s \right\|^2 \right\} 
\]
\[
= 2^{-1} n^{-1} \Delta^2 c_0 p_{\text{max}} \log(p_{\text{max}}) - O_p(\Delta n^{-1}p_{\text{max}} \log(p_{\text{max}})) - \Delta^3/2 n^{-1}p_{\text{max}} o_p(1) 
\]
\[
= 2^{-1} n^{-1} \Delta^2 c_0 p_{\text{max}} \log(p_{\text{max}}) \geq 0, \quad \text{(40)}
\]

where the last inequality is obtained from Conditions (C5) and (C6) and the fact that \(\lambda_n = O(\sqrt{\log(p_{\text{max}})/n}).\) This completes the proof of (19).

Equation (10) can be proved in a similar way. Note that \(n - \lfloor n/J \rfloor \sim n\) and each sample from \(D_n\) is drawn independently from an identical distribution. Hence \(\hat{\beta}_{(s)}^{[n]}\) converges to \(\beta^*_s\) in the same
order as $\hat{\beta}_{(s)}$ for each $j = 1, 2, \ldots, J$, i.e.,

$$
\max_{1 \leq j \leq J} \max_{1 \leq s \leq S_n} \left\| \hat{\beta}_{[j]}^{(s)} - \beta_{(s)}^* \right\| = O_p \left( \sqrt{\frac{p_{\max} \log(p_{\max})}{n}} \right). \tag{41}
$$

7.2 Proof of Theorem 1

Let us introduce Lemma 2 which is useful for completing the proof of Theorem 1.

**Lemma 2** Assume that Condition 2 and

$$
\sup_{w \in \mathcal{W}} \left| \frac{\text{CV}(w) - R_n(w)}{R_n(w)} \right| = o_p(1) \tag{42}
$$

hold. Then

$$
\frac{R_n(\hat{w})}{\inf_{w \in \mathcal{W}} R_n(w)} \to 1 \tag{43}
$$

in probability, where $\hat{w}$ is the optimal solution from (8).

**Proof:**

By the definition of infimum, there exist a sequence $\vartheta_n$ and a vector sequence $w_n \in \mathcal{W}$ such that as $n \to \infty$, $\vartheta_n \to 0$ and

$$
\inf_{w \in \mathcal{W}} R_n(w) = R_n(w_n) - \vartheta_n. \tag{44}
$$

Depend on Condition 2, we have

$$
\frac{R_n(w_n)}{\inf_{w \in \mathcal{W}} R_n(w)} > \frac{\inf_{w \in \mathcal{W}} R_n(w)}{\inf_{w \in \mathcal{W}} R_n(w)} = 1, \tag{45}
$$

and

$$
\frac{\vartheta_n}{\inf_{w \in \mathcal{W}} R_n(w)} = o_p(1). \tag{46}
$$

Taking (42), (45) and (46) together, for any $\delta > 0$,

$$
\Pr \left\{ \left| \frac{\inf_{w \in \mathcal{W}} R_n(w)}{R_n(\hat{w})} - 1 \right| > \delta \right\} = \Pr \left\{ \frac{R_n(\hat{w}) - \inf_{w \in \mathcal{W}} R_n(w)}{R_n(\hat{w})} - 1 > \delta \right\} = \Pr \left\{ \frac{R_n(\hat{w}) - \text{CV}(\hat{w}) + \text{CV}(\hat{w}) - R_n(w_n) + \vartheta_n}{R_n(\hat{w})} > \delta \right\}
$$
By Lemma 2 and the triangle inequality, it suffices to verify that
\[
\Pr\left\{ \frac{R_n(\hat{w}) - CV(\hat{w}) + CV(w_n) - R_n(w_n) + \vartheta_n}{R_n(\hat{w})} > \delta \right\}
\]
\[
\leq \Pr\left\{ \frac{|R_n(\hat{w}) - CV(\hat{w})|}{R_n(\hat{w})} + \frac{|CV(w_n) - R_n(w_n)|}{\inf_{w \in W} R_n(w)} + \frac{\vartheta_n}{R_n(\hat{w})} > \delta \right\}
\]
\[
\leq \Pr\left\{ \sup_{w \in W} \left| \frac{R_n(w) - CV(w)}{R_n(w)} \right| + \frac{|CV(w_n) - R_n(w_n)|}{\inf_{w \in W} R_n(w)} + \frac{\vartheta_n}{R_n(\hat{w})} > \delta \right\}
\]
\[
\rightarrow 0,
\]
(47)
in sum (43) is valid.

Let
\[
T_n = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \tilde{\beta}(w) \right)_+
\]
(48)
By Lemma 2 and the triangle inequality, it suffices to verify that
\[
\sup_{w \in W} \left| \frac{CV(w) - T_n(w)}{R_n(w)} \right| = o_p(1),
\]
(49)
and
\[
\sup_{w \in W} \left| \frac{T_n(w) - R_n(w)}{R_n(w)} \right| = o_p(1).
\]
(50)
For (49),
\[
|CV(w) - T_n(w)| = \left| \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in A(j)} \left\{ (1 - y_i x_i^T \tilde{\beta}(w))_+ - (1 - y_i x_i^T \hat{\beta}(w))_+ \right\} \right|
\]
\[
\leq \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in A(j)} \left| \int_{y_i x_i^T \tilde{\beta}(w)}^{y_i x_i^T \hat{\beta}(w)} I(t \leq 1) dt \right|
\]
\[
\leq \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in A(j)} \left| y_i x_i^T \left( \tilde{\beta}^{[-j]}(w) - \hat{\beta}(w) \right) \right|
\]
\[
\leq \frac{1}{n} \sum_{j=1}^{J} \sum_{i \in A(j)} \sum_{s=1}^{S_n} w_s \| x_{i(s)} \| \| \tilde{\beta}^{[-j]}(s) - \hat{\beta}(s) \|
\]
\[
\leq \frac{1}{n} \sum_{i=1}^{n} \max_{1 \leq j \leq S_n} \| x_{(s),i} \| \max_{1 \leq j \leq S_n} \frac{\max_{1 \leq j \leq S_n} \| \tilde{\beta}^{[-j]}(s) - \hat{\beta}(s) \|}{\| \tilde{\beta}(s) - \hat{\beta}(s) \|}
\]
29
\[
\leq C_1 \sqrt{p_{\text{max}}} \max_{1 \leq j \leq J} \max_{1 \leq s \leq S_n} \left\| \beta^{[-j]} - \widehat{\beta}_{(s)} \right\|
= O_p \left( \frac{p_{\text{max}} \sqrt{\log(p_{\text{max}})}}{\sqrt{n}} \right)
= o_p(1),
\]

where the second last equality is established based on Lemma 1, and the last equality is based on Conditions (C5)-(C6). Coupled with Condition 2 and (51), we obtain (49).

To prove (50), note that

\[
|T_n(w) - R_n(w)| = \left| \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \widehat{\beta}(w) \right) + - \mathbb{E} \left\{ (1 - y x^T \widehat{\beta}(w))_+ | D_n \right\} \right|
\leq \left| \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \widehat{\beta}(w) \right) + - \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta^*(w))_+ \right|
+ \left| \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta^*(w))_+ - \mathbb{E} (1 - y x^T \beta^*(w))_+ \right|
+ \mathbb{E} (1 - y x^T \beta^*(w))_+ - \mathbb{E} \left\{ (1 - y x^T \widehat{\beta}(w))_+ | D_n \right\}
\equiv |\Omega_1(w)| + |\Omega_2(w)| + |\Omega_3(w)|.
\]

(52)

Recognising the above, Lemma 1 and Conditions (C3), (C5) and (C6), it can be shown that

\[
\sup_{w \in W} |\Omega_1(w)| \leq \sup_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \left| \left( 1 - y_i x_i^T \widehat{\beta}(w) \right)_+ - (1 - y_i x_i^T \beta^*(w))_+ \right|
\leq \sup_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \left| y_i x_i^T \left( \beta^*(w) - \widehat{\beta}(w) \right) \right|
\leq \sup_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{S_n} w_s \|x_{(s),i}\| \left\| \beta^*_s - \widehat{\beta}_s \right\|
\leq \max_{1 \leq s \leq S_n} \left\| \beta^*_s - \widehat{\beta}_s \right\| \max_{1 \leq i \leq n} \max_{1 \leq s \leq S_n} |x_{(s),i}|\]
= O_p \left( \frac{p_{\text{max}} \sqrt{\log(p_{\text{max}})}}{\sqrt{n}} \right)
= o_p(1).
\]

(53)

Define the metric $| \cdot |_1$ on $W$, where

\[
|w - w'|_1 = \sum_{s=1}^{S_n} |w_s - w'_s|.
\]

(54)
for any $w = (w_1, ..., w_{S_n}) \in \mathcal{W}$ and $w' = (w'_1, ..., w'_{S_n}) \in \mathcal{W}$. Let $h_n = 1/(p_{\max} \log n)$ and create grids using regions of the form $\mathcal{W}^{(l)} = \{w : |w - w^{(l)}|_1 \leq h_n\}$. By the notion of the $\epsilon$–covering number introduced by Vaart and Wellner [1996], $\mathcal{W}$ can be covered with $N = O(1/h_n^{S_n-1})$ regions $\mathcal{W}^{(l)}, l = 1, ..., N$.

Note that

$$
\sup_{w \in \mathcal{W}^{(l)}} |\Omega_2(w) - \Omega_2(w^{(l)})| \\
\leq \sup_{w \in \mathcal{W}^{(l)}} \left| \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta^*(w))_+ - \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta^*(w^{(l)}))_+ \right| \\
+ \sup_{w \in \mathcal{W}^{(l)}} \left| \mathbb{E} \left( 1 - y x^T \beta^*(w) \right)_+ - \mathbb{E} \left( 1 - y x^T \beta^*(w^{(l)}) \right)_+ \right|
$$

$$
\leq \sup_{w \in \mathcal{W}^{(l)}} \frac{1}{n} \sum_{i=1}^{n} \left| y_i x_i^T \left\{ \beta^*(w^{(l)}) - \beta^*(w) \right\} \right| + \sup_{w \in \mathcal{W}^{(l)}} \mathbb{E} \left| y x^T \left\{ \beta^*(w^{(l)}) - \beta^*(w) \right\} \right|
$$

$$
\leq \sup_{w \in \mathcal{W}^{(l)}} \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{S_n} |w_s - w_s^{(l)}| \left\| x_i^T \beta^*(s) \right\| + \sup_{w \in \mathcal{W}^{(l)}} \sum_{s=1}^{S_n} |w_s - w_s^{(l)}| \mathbb{E} \left\| x_i^T \beta^*(s) \right\|
$$

$$
= \sup_{w \in \mathcal{W}^{(l)}} |w - w^{(l)}|_1 \max_{1 \leq s \leq S_n} \left\| \beta^*_s \right\| \left( \max_{1 \leq i \leq n} \max_{1 \leq s \leq S_n} \| x_i \| \right) + \max_{1 \leq i \leq n} \max_{1 \leq s \leq S_n} \mathbb{E} \left\| x_i \right\|
$$

$$
\leq \frac{C_2 \sqrt{p_{\max}}}{p_{\max} \log(n)} \frac{2C_1 \sqrt{p_{\max}}}{p_{\max} \log(n)}
$$

$$
= O_p \left( \log^{-1}(n) \right)
$$

$$
= o_p(1),
$$

(55)

where the result holds uniformly for $j$. Hence we have

$$
\sup_{w \in \mathcal{W}} |\Omega_2(w)| = \max_{1 \leq l \leq N} \sup_{w \in \mathcal{W}^{(l)}} |\Omega_2(w)|
$$

$$
\leq \max_{1 \leq l \leq N} |\Omega_2(w^{(l)})| + \max_{1 \leq l \leq N} \sup_{w \in \mathcal{W}^{(l)}} |\Omega_2(w) - \Omega_2(w^{(l)})|
$$

$$
= \max_{1 \leq l \leq N} |\Omega_2(w^{(l)})| + o_p(1).
$$

(56)

Furthermore, for any $\epsilon > 0$,

$$
\Pr \left\{ \max_{1 \leq l \leq N} |\Omega_2(w^{(l)})| > 3\epsilon \right\}
$$

$$
= \Pr \left[ \max_{1 \leq l \leq N} \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \beta^*(w^{(l)}) \right)_+ 1 \left( |1 - y_i x_i^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right)
$$

$$
+ \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i x_i^T \beta^*(w^{(l)}) \right)_+ 1 \left( |1 - y_i x_i^T \beta^*(w^{(l)})| \geq p_{\max} n^{0.1} \right)
$$

31
Clearly,

\[
- \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)})) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right) \right\} \\
- \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)})) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| \geq p_{\max} n^{0.1} \right) \right\} \leq 3 \epsilon \\
\leq \Pr \left[ \max_{1 \leq i \leq N} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})) + 1 \left( |1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right) \right] \\
- \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)})) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right) \right\} \geq \epsilon \\
+ \Pr \left[ \max_{1 \leq i \leq N} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})) + 1 \left( |1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})| \geq p_{\max} n^{0.1} \right) > \epsilon \right] \\
+ \Pr \left[ \max_{1 \leq i \leq N} \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)})) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| \geq p_{\max} n^{0.1} \right) \right\} > \epsilon \right] \\
\equiv \Xi_1 + \Xi_2 + \Xi_3. \quad (57)
\]

Using Boole’s and Bernstein’s inequalities and by (58),

\[
\Xi_1 \leq \sum_{j=1}^{N} \Pr \left[ \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})) + 1 \left( |1 - y_i \tilde{x}_i^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right) \right] \\
- \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)})) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| < p_{\max} n^{0.1} \right) \right\} \geq \epsilon \\
\leq N \exp \left( - \frac{n^2 \epsilon^2}{4 C_1^2 C_2^2 n p_{\max}^2 + \epsilon p_{\max} n^{0.1}/3} \right) \\
\leq (p_{\max} \log n)^{S_n-1} \exp \left( - \frac{n^2 \epsilon^2}{4 C_1^2 C_2^2 n p_{\max}^2 + \epsilon p_{\max} n^{0.1}/3} \right) \\
= O \left\{ \exp \left( - \epsilon^2 n p_{\max}^{-2} + S_n \log(p_{\max}) + S_n \log \log(n) \right) \right\}
\]

32
where the last equality is established from Condition (C5) and (C6). Additionally, we can write

\[ \Xi_2 = \Pr \left\{ \max_{1 \leq i \leq N} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \beta^*(w^{(l)}) + 1 \left( |1 - y_i x_i^T \beta^*(w^{(l)})| \geq p_{\text{max}} n^{0.1} \right) > \epsilon \right\} \]

\leq \Pr \left\{ \max_{1 \leq i \leq N} \max_{1 \leq i \leq n} |1 - y_i x_i^T \beta^*(w^{(l)})| \geq p_{\text{max}} n^{0.1} \right\}

\leq \Pr \left\{ \max_{1 \leq i \leq n} \max_{1 \leq i \leq n} \sum_{s=1}^{S_n} w_s^{(l)} \left( 1 + \|x_s,i\| \|\beta^*_s\| \right) \geq p_{\text{max}} n^{0.1} \right\}

\leq \Pr \left\{ \left( 1 + \max_{1 \leq i \leq n} \|x_s,i\| \max_{1 \leq s \leq S_n} \|\beta^*_s\| \right) \geq p_{\text{max}} n^{0.1} \right\}

= o(1), \quad (59)

where the last inequality holds because of Conditions (C2) and (C3). Similarly,

\[ \Xi_3 = \Pr \left\{ \max_{1 \leq i \leq N} \mathbb{E} \left\{ (1 - y x^T \beta^*(w^{(l)}) + 1 \left( |1 - y x^T \beta^*(w^{(l)})| \geq p_{\text{max}} n^{0.1} \right) \right\} > \epsilon \right\} \]

\leq \Pr \left\{ \max_{1 \leq i \leq N} \mathbb{E} |1 - y x^T \beta^*(w^{(l)})| \geq p_{\text{max}} n^{0.1} \right\}

\leq \Pr \left\{ \left( 1 + \max_{1 \leq s \leq S_n} \mathbb{E} \|x_s\| \max_{1 \leq s \leq S_n} \|\beta^*_s\| \right) \geq p_{\text{max}} n^{0.1} \right\}

= o(1). \quad (60)

Together with (57), (59)--(61), we obtain \( \max_{1 \leq i \leq N} |\Omega_2(w^{(l)})| = o_P(1) \). As well, by (56), we have

\[ \sup_{w \in \mathcal{W}} |\Omega_2(w)| = o_P(1). \quad (62) \]

Finally, note that \( (y, x) \) and \( (\tilde{y}, \tilde{x}) \) are independently and identically distributed, and under Lemma 1, we have

\[
\sup_{w \in \mathcal{W}} |\Omega_3(w)| = \sup_{w \in \mathcal{W}} \left| \mathbb{E} \left\{ (1 - y x^T \beta^*(w)) + \mathbb{E} \left\{ (1 - \tilde{y} \tilde{x}^T \hat{\beta}(w)) \right\} | D_n \right\} \right|
\]

\leq \sup_{w \in \mathcal{W}} \left| \mathbb{E} \left\{ \int_{0}^{\infty} \tilde{y} \tilde{x}^T \hat{\beta}(w) I(t \leq 1) dt \right| D_n \right|

\leq \sup_{w \in \mathcal{W}} \left| \mathbb{E} \left\{ \left\| \tilde{y} \tilde{x}^T \hat{\beta}(w) - \beta^*(w) \right\| \right| D_n \right|

\leq \sup_{w \in \mathcal{W}} \sum_{s=1}^{S_n} w_s \mathbb{E} \left\{ \left\| \tilde{x}_s^T \left( \hat{\beta}(s) - \beta^*_s \right) \right\| \right| D_n \right|
\[
\leq \max_{1 \leq s \leq S_n} \left\| \hat{\beta}_{(s)} - \beta^*_s \right\| \max_{1 \leq s \leq S_n} \mathbb{E}\left\| \tilde{x}_{(s),i} \right\|
\]

\[
= O_p \left( p_{\max} \sqrt{\log(p_{\max})} \sqrt{\frac{n}{t}} \right)
\]

\[
= o_p(1),
\]

where the last inequality holds due to Condition (C6). Putting (52), (53), (62) and (63) together, we complete the proof of (50).

References

J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287, 2011.

T. Ando and K.-C. Li. A weight-relaxed model averaging approach for high-dimensional generalized linear models. *The Annals of Statistics*, 45:2654–2679, 2017.

N. Becker, G. Toedt, P. Lichter, and A. Benner. Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC bioinformatics*, 12(1):138–151, 2011.

P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. *In ICML*, 98:82–90, 1998.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: An integral part of inference. *Biometrics*, 53(2):603–618, 1997.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

Claeskens, Gerda, Croux, Christophe, V. Kerckhoven, and Johan. An information criterion for variable selection in support vector machines. *Journal of Machine Learning Research*, 9:541–558, 2008.

G. Claeskens, C. Croux, and J. van Kerckhoven. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62:972–979, 2006.

A. Das and S. Bhattacharya. Distributed weighted parameter averaging for svm training on big data. *arXiv*, 2015.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.
Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

R. P. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89, 1988.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

B. E. Hansen and J. Racine. Jackknife model averaging. *Journal of Econometrics*, 167:38–46, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–417, 1999.

R. Jagannathan and T. Ma. Risk reduction in large portfolios why imposing the wrong constraints helps. *The Journal of Finance*, 58:1651–1683, 2003.

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

E. R. Lee, H. Noh, and B. U. Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

C. Park, K. R. Kim, R. Myung, and J. Y. Koo. Oracle properties of scad-penalized support vector machine. *Journal of Statistical Planning and Inference*, 142(8):2257–2270, 2012.

R. E. Schapire. The strength of week learnability. *Machine Learning*, 5(2):197–227, 1990.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10:262–266, 1989.

A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig. Objective automatic assessment of rehabilitative speech treatment in parkinson’s disease. *IEEE transactions on Neural Systems and Rehabilitation Engineering*, 22:181–190, 2014.

A. W. v. d. Vaart and J. A. Wellner. *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer, 1996.

S. van de Geer. *Empirical processes in m-estimation*. Cambridge series in statistical and probabilistic mathematics, 2000.
V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

G. Wahba. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, pages 69–87, 1999.

A. T. K. Wan, X. Zhang, and G. Zou. Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156:277–283, 2010.

L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16 (2):589–615, 2006.

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

M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17:1368–1385, 2011.

J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. *In NIPS*, 12:668–674, 2000.

Z. Yuan and Y. Yang. Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472):1202–1214, 2005.

H. H. Zhang, J. Ahn, X. Lin, and C. Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1):88–95, 2006.

X. Zhang, Z. Lu, and G. Zou. Adaptively combined forecasting for discrete response time series. *Journal of Econometrics*, 176:80–91, 2013.

X. Zhang, Y. Wu, L. Wang, and R. Li. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society, Series B*, 2014.

X. Zhang, Y. Wu, L. Wang, and R. Li. A consistent information criterion for support vector machines in diverging model spaces. *Journal of Machine Learning Research*, 17:1–26, 2016a.

X. Zhang, D. Yu, G. Zou, and H. Liang. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111:1775–1790, 2016b.

X. Zhang, G. Zou, H. Liang, and R. J. Carroll. Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115:972–984, 2020.

Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC Press, New York, 2012.

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

H. Zou and M. Yuan. The $f_\infty$-norm support vector machine. *Statistica Sinica*, 18:379–398, 2008.

J. Zou. *Weighted Average Estimation of Some Complex Models in Divergent Dimensions*. PhD thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2021.