Spring 2015

Stability of machine learning algorithms

Wei Sun
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Wei Sun

Entitled
STABILITY OF MACHINE LEARNING ALGORITHMS

For the degree of Doctor of Philosophy ▼

Is approved by the final examining committee:

Guang Cheng
Chair

Jayanta K. Ghosh

Xiao Wang

Mark Daniel Ward

Lingsong Zhang

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Guang Cheng

Approved by: Jun Xie 4/8/2015
Head of the Departmental Graduate Program Date

STABILITY OF MACHINE LEARNING ALGORITHMS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Wei Sun

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to extend my sincerest thank to my advisor, Professor Guang Cheng for his brilliant guidance and inspirational advice. It has been the most valuable and rewarding experience working with him. As an advisor, Guang gives me enough freedom to pursue my research interests in machine learning. He has also provided numerous opportunities for me to attend meetings and collaborate with faculties from other Universities. He has been not only my advisor, but my role model as a diligent researcher to pursue important and deep topics. As a friend, he has been listening to my heart and helping me. He has been and will continue to be a source of wisdom in my life! Thanks for being a fantastic advisor and friend.

I would also like to express my great gratitude to my collaborators. I am especially indebted to Professor Junhui Wang for opening the door for me to the world of machine learning. It was a great pleasure to work with Professor Yufeng Liu at UNC, who has strongly supported every step I took in the graduate school. I thank Professor Xingye Qiao from Binghamton University for his valuable suggestions and helpful discussions on my thesis. I was very lucky to work with extremely intelligent and hard-working people at Princeton University, namely Zhaoran Wang, Junwei Lu, and Professor Han Liu. Thank Professor Yixin Fang at NYU for many valuable discussions. I also give many thanks to Pengyuan Wang and Dawei Yin at Yahoo! labs for the enjoyable collaborations during my summer internship.

On the other hand, I deeply appreciate the guidance I have received from professors at Purdue University. Especially, I wish to thank Professor Jayanta K. Ghosh for his helpful comments on teaching during the period when I was a TA for his STAT528 course. Many thanks go to Professor Xiao Wang for the fruitful discussions on deep learning and Professor Lingsong Zhang, Professor Mark Ward for serving on my committee and giving me invaluable comments to improve the thesis. Special thanks go

to Professor Rebecca W. Doerge for her numerous supports on my academic travels and various award applications. Furthermore, I thank Professors William S. Cleveland, Jose Figueroa-Lopez, Sergey Kirshner, Chuanhai Liu, Yuan (Alan) Qi, Thomas Sellke for inspirational lectures that help a lot in my daily research.

I would like to acknowledge group members of Professor Guang Cheng's research group, including Professor Shenchun Kong, Professor Qifan Song, Dr. Zuofeng Shang, Zhuqing Yu, Ching-Wei Cheng, Meimei Liu, and Botao Hao for many valuable discussions on research problems over the past four years.

I also deeply appreciate generous helps from friends at Purdue. Fishing with Qiming Huang and Whitney Huang was a lot of fun. I also enjoyed a lot when we played cards during fun nights with Longjie Cheng, Xian He, Cheng Li, Chao Pan, Qiming Huang, and Bingrou (Alice) Zhou. I would also like to thank Yongheng Zhang and Xia Huang. It was great time to have fun with Terrence. Without the happiness brought to me by my friends, my life as a PhD student at West Lafayette would be miserable without a doubt.

Finally, I would like to express my heartfelt gratitude to my family, especially to my wife, whose love and support has been the driving force of my journey.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

BNN	Bagged Nearest Neighbor Classifier
CIS	Classification Instability
DBI	Decision Boundary Instability
GE	Generalization Error
kNN	k Nearest Neighbor Classifier
OWNN	Optimal Weighted Nearest Neighbor Classifier
SNN	Stabilized Nearest Neighbor Classifier
WNN	Weighted Nearest Neighbor Classifier

ABSTRACT

Sun, Wei PhD, Purdue University, May 2015. Stability of Machine Learning Algorithms. Major Professor: Guang Cheng.

In the literature, the predictive accuracy is often the primary criterion for evaluating a learning algorithm. In this thesis, I will introduce novel concepts of stability into the machine learning community. A learning algorithm is said to be stable if it produces consistent predictions with respect to small perturbation of training samples. Stability is an important aspect of a learning procedure because unstable predictions can potentially reduce users' trust in the system and also harm the reproducibility of scientific conclusions. As a prototypical example, stability of the classification procedure will be discussed extensively. In particular, I will present two new concepts of classification stability.

The first one is the decision boundary instability (DBI) which measures the variability of linear decision boundaries generated from homogenous training samples. Incorporating DBI with the generalization error (GE), we propose a two-stage algorithm for selecting the *most accurate and stable* classifier. The proposed classifier selection method introduces the statistical inference thinking into the machine learning society. Our selection method is shown to be consistent in the sense that the optimal classifier simultaneously achieves the minimal GE and the minimal DBI. Various simulations and real examples further demonstrate the superiority of our method over several alternative approaches.

The second one is the classification instability (CIS). CIS is a general measure of stability and generalizes DBI to nonlinear classifiers. This allows us to establish a sharp convergence rate of CIS for general plug-in classifiers under a low-noise condition. As one of the simplest plug-in classifiers, the nearest neighbor classifier is extensively studied. Motivated by an asymptotic expansion formula of the CIS of the weighted nearest neighbor classifier, we propose a new classifier called stabilized nearest neighbor (SNN) classifier. Our theoretical developments further push the frontier of statistical theory in machine learning. In particular, we prove that SNN attains the minimax optimal convergence rate in the risk, and the established sharp convergence rate in CIS. Extensive simulation and real experiments demonstrate that SNN achieves a considerable improvement in stability over existing classifiers with no sacrifice of predictive accuracy.

1. INTRODUCTION

The predictive accuracy is often the primary criterion for evaluating a machine learning algorithm. Recently, researchers have started to explore alternative measures to evaluate the performance of a learning algorithm. For instance, besides prediction accuracy, computational complexity, robustness, interpretability, and variable selection performance have been considered in the literature. Our work follows this research line since we believe there are other critical properties (other than accuracy) of a machine learning algorithm that have been overlooked in the research community. In this thesis, I will introduce novel concepts of stability into the machine learning community. A learning algorithm is said to be stable if it produces consistent predictions with respect to small perturbation of training samples.

Stability is an important aspect of a learning algorithm. Data analyses have become a driving force for much scientific research work. As datasets get bigger and analysis methods become more complex, the need for reproducibility has increased significantly [1]. Many experiments are being conducted and conclusions are being made with the aid of statistical analyses. Those with great potential impacts must be scrutinized before being accepted. An initial scrutiny involves reproducing the result. A minimal requirement is that one can reach the same conclusion by applying the described analyses to the same data, a notion some refer to as *replicability*. A more general requirement is that one can reach a similar result based on independently generated datasets. The issue of reproducibility has drawn much attention in statistics [2], biostatistics [3, 4], computational science [5] and other scientific communities [6]. Recent discussions can be found in a special issue of *Nature*[1]). Moreover, Marcia McNutt, the Editor-in-Chief of *Science*, pointed out that "reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions."

[1] at http://www.nature.com/nature/focus/reproducibility/

That is, if conclusions can not be reproduced, the credit of the researchers, along with the scientific conclusions themselves, will be in jeopardy.

Throughout the whole scientific research process, there are many ways statistics as a subject can help improve reproducibility. One particular aspect we stress in this thesis is the stability of the statistical procedure used in the analysis. According to [2], scientific conclusions should be stable with respect to small perturbation of data. The danger of an unstable statistical method is that a correct scientific conclusion may not be recognized and could be falsely discredited, simply because an unstable statistical method was used.

Moreover, stability can be very important in some practical domains. Customers often evaluate a service based on their experience for a small sample, where the accuracy is either hard to perceive (due to the lack of ground truth), or does not appear to differ between different services (due to data inadequacy); on the other hand, stability is often more perceptible and hence can be an important criterion. For example, Internet streaming service provider Netflix has a movie recommendation system based on complex learning algorithms. Viewers either can not promptly perceive the inaccuracy because they themselves do not know which film is the best for them, or are quite tolerable even if a sub-optimal recommendation is given. However, if two consecutively recommended movies are from two totally different genres, the customer can immediately perceive such instability, and have a bad user experience with the service. Furthermore, providing a stable prediction plays a crucial role on users' trust of the classification system. In the psychology literature, it has been shown that advice-giving agents with a lager variability in past opinions are considered less informative and less helpful than those with a more consistent pattern of opinions [7, 8]. Therefore, a machine learning system may be distrusted by users if it generates highly unstable predictions simply due to the randomness of training samples.

It is worth mentioning that stability has indeed received much attention in statistics. For example, in clustering problems, [9] introduced the clustering instability to assess the quality of a clustering algorithm; [10] used the clustering instability as a criterion to select the number of clusters. In high-dimensional regression, [11] and [12] proposed stability selection procedures for variable selection, and [13] and [14] applied stability for tuning parameter selection. For more applications, see the use of stability in model selection [15], analyzing the effect of bagging [16], and deriving the generalization error bound [17, 18]. However, many of these works view stability as a tool for other purposes. In literature, few work has emphasized the importance of stability itself.

As a prototypical example, in this thesis we will discuss extensively the stability of a classification procedure. Classification aims to identify the class label of a new subject using a classifier constructed from training data whose class memberships are given. It has been widely used in diverse fields, e.g., medical diagnosis, fraud detection, and computer vision. In the literature, much of the research focuses on improving the accuracy of classifiers. Recently, alternative criteria have been explored, such as computational complexity and training time [19], the robustness [20], among others. Our work focuses on another critical property of classifiers, namely stability, that has been somewhat overlooked. A classification procedure with more stable prediction performance is preferred when researchers aim to reproduce the reported results from randomly generated samples. Consequently, aside from high prediction accuracy, high stability is another crucial factor to consider when evaluating the performance of a classification procedure. Our work tries to fill this gap by presenting two new concepts of classification stability.

1.1 Decision Boundary Instability (DBI)

In Section 2, I will introduce the decision boundary instability (DBI) to capture the variability of decision boundaries arose from homogenous training samples. Incorporating DBI with the generalization error (GE), we propose a two-stage algorithm for selecting the *most accurate and stable* classifier: Stage (i) eliminate the classifiers whose GEs are significantly larger than the minimal one among all the candidate classifiers; Stage (ii) select the optimal classifier as that with the most stable decision boundary, i.e., the minimal DBI, among the remaining classifiers. Our selection method is shown to be consistent in the sense that the optimal classifier simultaneously achieves the minimal GE and the minimal DBI. Various simulations and real examples further demonstrate the superiority of our method over several alternative approaches.

1.2 Classification Instability (CIS)

In Section 3, I will introduce the classification instability (CIS) which characterizes the sampling variability of the yielded prediction. CIS is a general measure of stability for both linear and nonlinear classifiers. This allows us to establish a sharp convergence rate of CIS for general plug-in classifiers under a low-noise condition. This sharp rate is slower than but approaching n^{-1}, which is shown by adapting the theoretical framework of [21]. As one of the simplest plug-in classifiers, the nearest neighbor classifier is extensively studied. An important result we find is that the CIS of a weighted nearest neighbor (WNN) classifier procedure is asymptotically proportional to the Euclidean norm of the weight vector. This rather concise form allows us to propose a new classifier called stabilized nearest neighbor (SNN) classifier, which is the optimal solution by minimizing the CIS of a WNN procedure over an acceptable region of the weight where the regret is small. In theory, we prove that SNN attains the minimax optimal convergence rate in the risk, and the established sharp convergence rate in CIS. Extensive simulation and real experiments demonstrate that SNN achieves a considerable improvement in stability over existing classifiers with no sacrifice of predictive accuracy.

2. DECISION BOUNDARY INSTABILITY

Classification aims to identify the class label of a new subject using a classifier constructed from training data whose class memberships are given. It has been widely used in diverse fields, e.g., medical diagnosis, fraud detection, and natural language processing. Numerous classification methods have been successfully developed with classical approaches such as Fisher's linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression [22], and modern approaches such as support vector machine (SVM) [23], boosting [24], distance weighted discrimination (DWD) [25], classification based on the reject option [26], and optimal weighted nearest neighbor classifiers [27]. In a recent paper, [28] proposed a platform, large-margin unified machines (LUM), for unifying various large margin classifiers ranging from soft to hard.

In the literature, much of the research has focused on improving the predictive accuracy of classifiers and hence generalization error (GE) is often the primary criterion for selecting the optimal one from the rich pool of existing classifiers; see [29] and [30]. Recently, researchers have started to explore alternative measures to evaluate the performance of classifiers. For instance, besides prediction accuracy, computational complexity and training time of classifiers are considered in [19]. Moreover, [20] proposed the robust truncated hinge loss SVM to improve the robustness of the standard SVM. [31] and [32] investigated several measures of cost-sensitive weighted generalization errors for highly unbalanced classification tasks since, in this case, GE itself is not very informative. Our work follows this research line since we believe there are other critical properties (other than accuracy) of classifiers that have been overlooked in the research community. In this article, we introduce a notion of decision boundary instability (DBI) to assess the stability [15] of a classification procedure arising from the randomness of training samples. Aside from high prediction accuracy, high stability is another crucial factor to consider in the classifier selection.

In this paper, we attempt to select the *most accurate and stable* classifier by incorporating DBI into our selection process. Specifically, we suggest a two-stage selection procedure: (i) eliminate the classifiers whose GEs are significantly larger than the minimal one among all the candidate classifiers; (ii) select the optimal classifier as that with the most stable decision boundary, i.e., the minimal DBI, among the remaining classifiers. In the first stage, we show that the cross-validation estimator for the difference of GEs induced from two large-margin classifiers asymptotically follows a Gaussian distribution, which enables us to construct a confidence interval for the GE difference. If this confidence interval contains 0, these two classifiers are considered indistinguishable in terms of GE. By applying the above approach, we can obtain a collection of potentially good classifiers whose GEs are close enough to the minimal value. The uncertainty quantification of the cross-validation estimator is crucially important considering that only limited samples are available in practice. In fact, experiments indicate that for certain problems many classifiers do not significantly differ in their estimated GEs, and the corresponding absolute differences are mainly due to random noise.

A natural follow-up question is whether the collection of potentially good classifiers also perform well in terms of their stability. Interestingly, we observe that the decision boundary generated by the classifier with the minimal GE estimator sometimes has unstable behavior given a small perturbation of the training samples. This observation motivates us to propose a further selection criterion in the second stage: DBI. This new measure can precisely reflect the visual variability in the decision boundaries due to the perturbed training samples.

Our two-stage selection algorithm is shown to be consistent in the sense that the selected optimal classifier simultaneously achieves the minimal GE and the minimal DBI. The proof is nontrivial because of the stochastic nature of the two-stage algorithm. Note that our method is distinguished from the bias-variance analysis in classification since the latter focuses on the decomposition of GE, e.g., [33]. Our DBI is also conceptually different from the stability-oriented measure introduced in [17], which was defined as the maximal difference of the decision functions trained from the original datasets and the leave-one-out datasets. In addition, their variability measure suffers from the transformation variant issue since a scale transformation of the decision function coefficients will greatly affect their variability measure. Our DBI overcomes this problem via a rescaling scheme since DBI can be viewed as a weighted version of the asymptotic variance of the decision function. In the end, extensive experiments illustrate the advantage of our selection algorithm compared with the alternative approaches in terms of both classification accuracy and stability.

2.1 Large-Margin Classifiers

This section briefly reviews large-margin classifiers, which serve as prototypical examples to illustrate our two-stage classifier selection technique.

Let $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \{1, -1\}$ be random variables from an underlying distribution $\mathcal{P}(\boldsymbol{X}, Y)$. Denote the conditional probability of class $Y = 1$ given $\boldsymbol{X} = \boldsymbol{x}$ as $p(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$, where $p(\boldsymbol{x}) \in (0, 1)$ to exclude the degenerate case. Let the input variable be $\boldsymbol{x} = (x_1, \dots, x_d)^T$, $\tilde{\boldsymbol{x}} = (1, x_1, \dots, x_d)^T$, with coefficient $\boldsymbol{w} = (w_1, \dots, w_d)^T$ and parameter $\boldsymbol{\theta} = (b, \boldsymbol{w}^T)^T$. The linear decision function is defined as $f(\boldsymbol{x}; \boldsymbol{\theta}) = b + \boldsymbol{x}^T\boldsymbol{w} = \tilde{\boldsymbol{x}}^T\boldsymbol{\theta}$, and the decision boundary is $S(\boldsymbol{x}; \boldsymbol{\theta}) = \{\boldsymbol{x} : f(\boldsymbol{x}; \boldsymbol{\theta}) = 0\}$. The performance of the classifier $\text{sign}\{f(\boldsymbol{x}; \boldsymbol{\theta})\}$ is measured by the classification risk $E[I_{\{Y \neq \text{sign}\{f(\boldsymbol{X};\boldsymbol{\theta})\}\}}]$, where the expectation is with respect to $\mathcal{P}(\boldsymbol{X}, Y)$. Since the direct minimization of the above risk is NP hard [34], various convex surrogate loss functions $L(\cdot)$ have been proposed to deal with this computational issue. Denote the surrogate risk as $\mathcal{R}_L(\boldsymbol{\theta}) = E[L(Yf(\boldsymbol{X}; \boldsymbol{\theta}))]$, and assume that the minimizer of $\mathcal{R}_L(\boldsymbol{\theta})$ is obtained at $\boldsymbol{\theta}_{0L} = (b_{0L}, \boldsymbol{w}_{0L}^T)^T$. Here $\boldsymbol{\theta}_{0L}$ depends on the loss function L.

Given the training sample $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ drawn from $\mathcal{P}(\boldsymbol{X}, Y)$, a large-margin classifier minimizes the empirical risk $O_{nL}(\boldsymbol{\theta})$ defined as

$$O_{nL}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)\Big) + \frac{\lambda_n}{2} \boldsymbol{w}^T \boldsymbol{w}, \tag{2.1}$$

where λ_n is some positive tuning parameter. The estimator minimizing $O_{nL}(\boldsymbol{\theta})$ is denoted as $\widehat{\boldsymbol{\theta}}_L = (\widehat{b}_L, \widehat{\boldsymbol{w}}_L^T)^T$. Common large-margin classifiers include the squared loss $L(u) = (1-u)^2$, the exponential loss $L(u) = e^{-u}$, the logistic loss $L(u) = \log(1+e^{-u})$, and the hinge loss $L(u) = (1-u)_+$. Unfortunately, there seems to be no general guideline for selecting these loss functions in practice except the cross validation error. Ideally if we had access to an arbitrarily large test set, we would just choose the classifier for which the test error is the smallest. However, in reality where only limited samples are available, the commonly used cross validation error may not be able to accurately approximate the testing error. The main goal of this paper is to establish a practically useful selection criterion by incorporating DBI with the cross validation error.

2.2 Classifier Selection Algorithm

In this section, we propose a two-stage classifier selection algorithm: (i) we select candidate classifiers whose estimated GEs are relatively small; (ii) the optimal classifier is that with the minimal DBI among those selected in Stage (i).

2.2.1 Stage 1: Initial Screening via GE

In this subsection, we show that the difference of the cross-validation errors obtained from two large-margin classifiers asymptotically follows a Gaussian distribution, which enables us to construct a confidence interval for their GE difference. We further propose a perturbation-based resampling approach to construct this confidence interval.

Given a new input $(\boldsymbol{X}_0, Y_0)$ from $\mathcal{P}(\boldsymbol{X}, Y)$, we define the GE induced by the loss function L as

$$D_{0L} = \frac{1}{2} E|Y_0 - \operatorname{sign}\{f(\boldsymbol{X}_0; \widehat{\boldsymbol{\theta}}_L)\}|, \tag{2.2}$$

where $\widehat{\boldsymbol{\theta}}_L$ is based on the training sample $\mathcal{D}_n$, and the expectation is with respect to both $\mathcal{D}_n$ and $(\boldsymbol{X}_0, Y_0)$. In practice, the GE, which depends on the underlying distribution $\mathcal{P}(\boldsymbol{X}, Y)$, needs to be estimated using $\mathcal{D}_n$. One possible estimate is the empirical generalization error defined as $\widehat{D}_L \equiv \widehat{D}(\widehat{\boldsymbol{\theta}}_L)$, where $\widehat{D}(\boldsymbol{\theta}) = (2n)^{-1} \sum_{i=1}^{n} |y_i - \operatorname{sign}\{f(\boldsymbol{x}_i; \boldsymbol{\theta})\}|$. However, the above estimate suffers from the problem of overfitting [35]. Hence, one can use the K-fold cross-validation procedure to estimate the GE; this can significantly reduce the bias [36]. Specifically, we randomly split $\mathcal{D}_n$ into K disjoint subgroups and denote the kth subgroup as I_k. For $k = 1, \ldots, K$, we obtain the estimator $\widehat{\boldsymbol{\theta}}_{L(-k)}$ from all the data except those in I_k, and calculate the empirical average $\widehat{D}(\widehat{\boldsymbol{\theta}}_{L(-k)})$ based only on I_k, i.e., $\widehat{D}(\widehat{\boldsymbol{\theta}}_{L(-k)}) = (2|I_k|)^{-1} \sum_{i \in I_k} |y_i - \operatorname{sign}\{f(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{L(-k)})\}|$ with $|I_k|$ being the cardinality of I_k. The K-fold cross-validation (K-CV) error is thus computed as

$$\widehat{\mathcal{D}}_L = K^{-1} \sum_{k=1}^{K} \widehat{D}(\widehat{\boldsymbol{\theta}}_{L(-k)}). \tag{2.3}$$

We set $K = 5$ for our numerical experiments.

We establish the asymptotic normality of the K-CV error $\widehat{\mathcal{D}}_L$ under the following regularity conditions:

(L1) The probability distribution function of $\boldsymbol{X}$ and the conditional probability $p(\boldsymbol{x})$ are both continuously differentiable.

(L2) The parameter $\boldsymbol{\theta}_{0L}$ is bounded and unique.

(L3) The map $\boldsymbol{\theta} \mapsto L(yf(\boldsymbol{x}; \boldsymbol{\theta}))$ is convex.

(L4) The map $\boldsymbol{\theta} \mapsto L(yf(\boldsymbol{x}; \boldsymbol{\theta}))$ is differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_{0L}$ a.s.. Furthermore, $G(\boldsymbol{\theta}_{0L})$ is element-wisely bounded, where

$$G(\boldsymbol{\theta}_{0L}) = E\Big[\nabla_{\boldsymbol{\theta}} L(Yf(\boldsymbol{X}; \boldsymbol{\theta})) \nabla_{\boldsymbol{\theta}} L(Yf(\boldsymbol{X}; \boldsymbol{\theta}))^T\Big]\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_{0L}}.$$

(L5) The surrogate risk $\mathcal{R}_L(\boldsymbol{\theta})$ is bounded and twice differentiable at $\boldsymbol{\theta} = \boldsymbol{\theta}_{0L}$ with the positive definite Hessian matrix $H(\boldsymbol{\theta}_{0L}) = \nabla^2_{\boldsymbol{\theta}} \mathcal{R}_L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}$.

Assumption (L1) ensures that the GE is continuously differentiable with respect to $\boldsymbol{\theta}$ so that the uniform law of large numbers can be applied. Assumption (L3) ensures that the uniform convergence theorem for convex functions [37] can be applied, and it is satisfied by all the large-margin loss functions considered in this paper. Assumptions (L4) and (L5) are required to obtain the local quadratic approximation to the surrogate risk function around $\boldsymbol{\theta}_{0L}$. Assumptions (L2)–(L5) were previously used by [38] to prove the asymptotic normality of $\widehat{\boldsymbol{\theta}}_L$.

Theorem 2.2.1 below establishes the asymptotic normality of the K-CV error $\widehat{\mathcal{D}}_L$ for any large-margin classifier, which generalizes the result for the SVM in [36].

Theorem 2.2.1 *Suppose Assumptions (L1)–(L5) hold and $\lambda_n = o(n^{-1/2})$. Then for any fixed K,*

$$\mathcal{W}_L = \sqrt{n}\Big(\widehat{\mathcal{D}}_L - D_{0L}\Big) \xrightarrow{d} N\Big(0, E(\psi_1^2)\Big) \quad \text{as } n \to \infty, \tag{2.4}$$

where $\psi_1 = \frac{1}{2}|Y_1 - \text{sign}\{f(\boldsymbol{X}_1;\boldsymbol{\theta}_{0L})\}| - D_{0L} - \dot{d}(\boldsymbol{\theta}_{0L})^T H(\boldsymbol{\theta}_{0L})^{-1} M_1(\boldsymbol{\theta}_{0L})$ with $\dot{d}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} E(\widehat{D}(\boldsymbol{\theta}))$, and $M_1(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} L(Y_1 f(\boldsymbol{X}_1;\boldsymbol{\theta}))$.

An immediate application of Theorem 2.2.1 is to compare two competing classifiers L_1 and L_2. Define their GE difference Δ_{12} and its consistent estimate $\widehat{\Delta}_{12}$ to be $D_{02} - D_{01}$ and $\widehat{\mathcal{D}}_2 - \widehat{\mathcal{D}}_1$, respectively. To test whether the GEs induced by L_1 and L_2 are significantly different, we need to establish an approximate confidence interval for Δ_{12} based on the distribution of $\mathcal{W}_{\Delta_{12}} \equiv \mathcal{W}_2 - \mathcal{W}_1 = n^{1/2}(\widehat{\Delta}_{12} - \Delta_{12})$. In practice, we apply the perturbation-based resampling procedure [39] to approximate the distribution of $\mathcal{W}_{\Delta_{12}}$. This procedure was also employed by [36] to construct the confidence interval of SVM's GE. Specifically, let $\{G_i\}_{i=1}^n$ be i.i.d. random variables drawn from the exponential distribution with unit mean and unit variance. Denote

$$\widehat{\boldsymbol{\theta}}_j^* = \arg\min_{b,\boldsymbol{w}} \left\{ \frac{1}{n} \sum_{i=1}^n G_i L_j\Big(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)\Big) + \frac{\lambda_n}{2} \boldsymbol{w}^T \boldsymbol{w} \right\}. \tag{2.5}$$

Conditionally on $\mathcal{D}_n$, the randomness of $\widehat{\boldsymbol{\theta}}_j^*$ merely comes from that of $G_1, \ldots, G_n$. Denote $W^*_{\Delta_{12}} = W_2^* - W_1^*$ with

$$W_j^* = n^{-1/2} \sum_{i=1}^{n} \Big\{ \frac{1}{2} \Big| y_i - \text{sign}\{f(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_j^*)\} \Big| - \widehat{D}_j \Big\} G_i. \tag{2.6}$$

By repeatedly generating a set of random variables $\{G_i, i = 1, \ldots, n\}$, we can obtain a large number of realizations of $W^*_{\Delta_{12}}$ to approximate the distribution of $\mathcal{W}_{\Delta_{12}}$. In Theorem 2.2.2 below, we prove that this approximation is valid.

Theorem 2.2.2 *Suppose the assumptions in Theorem 2.2.1 hold, we have*

$$\mathcal{W}_{\Delta_{12}} \xrightarrow{d} N\Big(0, Var(\psi_{12} - \psi_{11})\Big),$$

as $n \to \infty$, where ψ_{11} and ψ_{12} are defined in Appendix A.3, and

$$W^*_{\Delta_{12}} \overset{d}{\Longrightarrow} N\Big(0, Var(\psi_{12} - \psi_{11})\Big) \quad \textit{conditional on } \mathcal{D}_n,$$

where "$\Longrightarrow$" means conditional weak convergence in the sense of [40].

Algorithm 1 below summarizes the resampling procedure for establishing the confidence interval of the GE difference Δ_{12}.

Algorithm 1 (Generalization Error Comparison Algorithm)

Input: Training sample $\mathcal{D}_n$ and two candidate classifiers L_1 and L_2.

- Step 1. Calculate K-CV errors $\widehat{\mathcal{D}}_1$ and $\widehat{\mathcal{D}}_2$ of classifiers L_1 and L_2, respectively.
- Step 2. For $r = 1, \ldots, N$, repeat the following steps:
  (a) Generate i.i.d. samples $\{G_i^{(r)}\}_{i=1}^n$ from Exp(1);
  (b) Find $\widehat{\boldsymbol{\theta}}_j^{*(r)}$ via (2.5) and $W_j^{*(r)}$ via (2.6), and calculate $W^{*(r)}_{\Delta_{12}} = W_2^{*(r)} - W_1^{*(r)}$.
- Step 3. Construct the $100(1-\alpha)\%$ confidence interval for Δ_{12} as

$$\Big[\widehat{\Delta}_{12} - n^{-1/2}\phi_{1,2;\alpha/2}, \widehat{\Delta}_{12} - n^{-1/2}\phi_{1,2;1-\alpha/2}\Big],$$

  where $\widehat{\Delta}_{12} = \widehat{\mathcal{D}}_2 - \widehat{\mathcal{D}}_1$ and $\phi_{1,2;\alpha}$ is the αth upper percentile of $\{W^{*(1)}_{\Delta_{12}}, \ldots, W^{*(N)}_{\Delta_{12}}\}$.

In our experiments, we repeated the resampling procedure 100 times, i.e., $N = 100$ in Step 2, and fix $\alpha = 0.1$. The effect of the choice of α will be discussed at the end of Section 2.2.4.

The GEs of two classifiers L_1 and L_2 are significantly different if the confidence interval established in Step 3 does not contain 0. Hence, we can apply *Algorithm 1* to eliminate the classifiers whose GEs are significantly different from the minimal GE of a set of candidate classifiers.

It is worth noting that employing statistical testing for classifier comparison has been successfully applied in practice [41, 42]. In particular, [42] reviewed several statistical tests in comparing two classifiers on multiple data sets and recommended the Wilcoxon sign rank test, which examined whether two classifiers are significantly different by calculating the relative rank of their corresponding performance scores on multiple data sets. Their result relies on an ideal assumption that there is no sampling variability of the measured performance score in each individual data set. Compared to the Wilcoxon sign rank test, our perturbed cross validation estimator has the advantages that it is theoretically justified and it does not rely on the ideal assumption of each performance score.

The remaining classifiers from *Algorithm 1* are potentially good. As will be seen in the next section, the decision boundaries of potentially good classifiers may change dramatically following a small perturbation of the training sample. This indicates that the prediction stability of the classifiers can be different although their GEs are fairly close. Motivated by this observation, in the next section we introduce the DBI to capture the prediction instability and embed it into our classifier selection algorithm.

2.2.2 Stage 2: Final Selection via DBI

In this section, we define the DBI and then provide an efficient way to estimate it in practice.

Toy Example: To motivate the DBI, we start with a simulated example using two classifiers: the squared loss L_1 and the hinge loss L_2. Specifically, we generate 100 observations from a mixture of two Gaussian distributions with equal probability: $N((-0.5,-0.5)^T, I_2)$ and $N((0.5,0.5)^T, I_2)$ with I_2 an identity matrix of dimension two. In Figure 2.2.2, we plot the decision boundary $S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_j)$ (in black) based on $\mathcal{D}_n$, and 100 perturbed decision boundaries $\{S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_j^{*(1)}),\ldots,S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_j^{*(100)})\}$ (in gray) for $j=1,2$; see Step 2 of *Algorithm 1.* Figure 2.2.2 reveals that the perturbed decision boundaries of the squared loss are more stable than those of the SVM given a small perturbation of the training sample. Hence, it is desirable to quantify the variability of the perturbed decision boundaries with respect to the original unperturbed decision boundary $S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_j)$. This is a nontrivial task since the boundaries spread over a d-dimensional space, e.g., $d=2$ in Figure 2.2.2. Therefore, we transform the data in such a way that the above variability can be fully measured in a single dimension. Specifically, we find a $d \times d$ transformation matrix R_L, which is orthogonal with determinant 1, such that the decision boundary based on the transformed data $\mathcal{D}_n^\dagger = \{(\boldsymbol{x}_i^\dagger, y_i), i=1,\ldots,n\}$ with $\boldsymbol{x}_i^\dagger = R_L\boldsymbol{x}_i$ is parallel to the $\mathcal{X}_1,\ldots,\mathcal{X}_{d-1}$ axes; see Section 2.7.3 for the calculation of R_L. The variability of the perturbed decision boundaries with respect to the original unperturbed decision boundary then reduces to the variability along the last axis $\mathcal{X}_d$. For illustration purposes, we next apply the above data-transformation idea to the SVM plotted in the top right plot of Figure 2.2.2. From the bottom plot in Figure 2.2.2, we observe that the variability of the transformed perturbed decision boundaries (in gray) with respect to the transformed unperturbed decision boundary (in black) now reduces to the variability along the $\mathcal{X}_2$ axis only. This is because the transformed unperturbed decision boundary is parallel to the $\mathcal{X}_1$ axis. Note that the choice of data transformation is not unique. For example, we could also transform the data such that the transformed unperturbed decision boundary is parallel to the $\mathcal{X}_2$ axis and then measure the variability along the $\mathcal{X}_1$ axis. Fortunately, the DBI measure we will introduce yields exactly the same value under any transformation, i.e., it is transformation invariant.

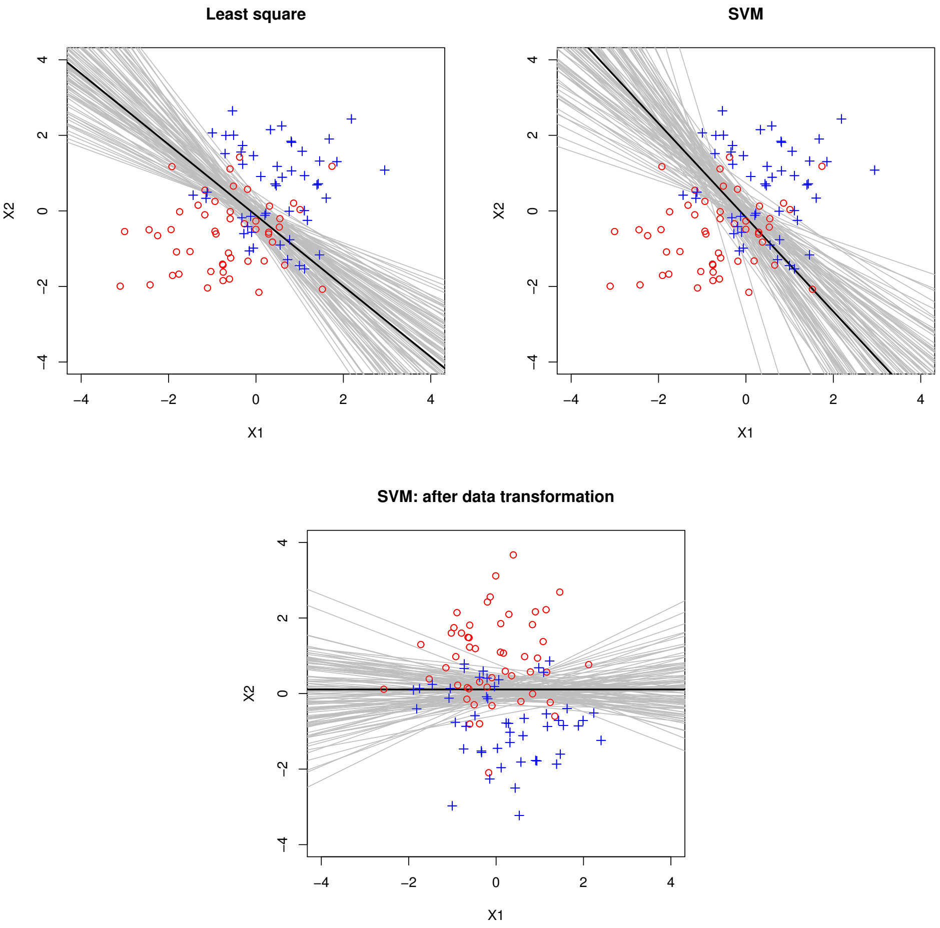

Figure 2.1. Two classes are shown in red circles and blue crosses. The black line is the decision boundary based on the original training sample, and the gray lines are 100 decision boundaries based on perturbed samples. The top left (right) panel corresponds to the least square loss (SVM). The perturbed decision boundaries of SVM after data transformation are shown in the bottom.

Now we are ready to define DBI. Given the loss function L, we define the coefficient estimator based on transformed data $\mathcal{D}_n^\dagger$ as $\widehat{\boldsymbol{\theta}}_L^\dagger$ and the coefficient estimator based on the perturbed samples of $\mathcal{D}_n^\dagger$ as $\widehat{\boldsymbol{\theta}}_L^{\dagger *}$. In addition, we find the following relationship through the transformation matrix R_L:

$$\widehat{\boldsymbol{\theta}}_L \equiv \begin{pmatrix} \widehat{b}_L \\ \widehat{\boldsymbol{w}}_L \end{pmatrix} \Rightarrow \widehat{\boldsymbol{\theta}}_L^\dagger \equiv \begin{pmatrix} \widehat{b}_L \\ R_L \widehat{\boldsymbol{w}}_L \end{pmatrix} \text{ and } \widehat{\boldsymbol{\theta}}_L^* \equiv \begin{pmatrix} \widehat{b}_L^* \\ \widehat{\boldsymbol{w}}_L^* \end{pmatrix} \Rightarrow \widehat{\boldsymbol{\theta}}_L^{\dagger *} \equiv \begin{pmatrix} \widehat{b}_L^* \\ R_L \widehat{\boldsymbol{w}}_L^* \end{pmatrix},$$

which can be shown by replacing $\boldsymbol{x}_i$ with $R_L \boldsymbol{x}_i$ in (2.1) and (2.5) and using the property of R_L.

DBI is defined as the variability of the transformed perturbed decision boundary $S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L^{\dagger *})$ with respect to the transformed unperturbed decision boundary $S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L^\dagger)$ along the direction $\mathcal{X}_d$.

Definition 1 *The decision boundary instability (DBI) of $S(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}_L)$ is defined to be*

$$DBI\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L)\Big) = E\left[Var\Big(S_d | \boldsymbol{X}_{(-d)}^\dagger\Big)\right], \tag{2.7}$$

where S_d is the dth dimension of $S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L^{\dagger *})$ and $\boldsymbol{X}_{(-d)}^\dagger = (X_1^\dagger, \dots, X_{d-1}^\dagger)^T$.

Remark 1 *The conditional variance $Var(S_d | \boldsymbol{X}_{(-d)}^\dagger)$ in (2.7) captures the variability of the transformed perturbed decision boundary along the dth dimension based on a given sample. Note that, after data transformation, the transformed unperturbed decision boundary is parallel to the $\mathcal{X}_1, \dots, \mathcal{X}_{d-1}$ axes. Therefore, this conditional variance precisely measures the variability of the perturbed decision boundary with respect to the unperturbed decision boundary conditioned on the given sample. The expectation in (2.7) then averages out the randomness in the sample.*

Toy Example Continuation: We next give an illustration of (2.7) via the 2-dimensional toy example shown in the bottom plot of Figure 2.2.2. For each sample, the conditional variance in (2.7) is estimated via the sample variability of the projected X_2 values on the perturbed decision boundary (in gray). Then the final DBI is estimated by averaging over all samples.

In Section 2.7.4, we demonstrate an efficient way to simplify (2.7) by approximating the conditional variance via the weighted variance of $\widehat{\boldsymbol{\theta}}_L^\dagger$. Specifically, we show that

$$DBI\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L)\Big) \approx (w_{L,d}^\dagger)^{-2} E\left[\tilde{\boldsymbol{X}}_{(-d)}^{\dagger T} \Big(n^{-1} \Sigma_{0L,(-d)}^\dagger\Big) \tilde{\boldsymbol{X}}_{(-d)}^\dagger\right], \tag{2.8}$$

where $w_{L,d}^{\dagger}$ is the last entry of the transformed coefficient $\boldsymbol{\theta}_{0L}^{\dagger}$, and $n^{-1}\Sigma_{0L,(-d)}^{\dagger}$ is the asymptotic variance of the first d dimensions of $\widehat{\boldsymbol{\theta}}_{L}^{\dagger}$. Therefore, DBI can be viewed as a proxy measure of the asymptotic variance of the decision function.

We next propose a plug-in estimate for the approximate version of DBI in (2.8). Direct estimation of DBI in (2.7) is possible, but it requires perturbing the transformed data. To reduce the computational cost, we can take advantage of our resampling results in Stage 1 based on the relationship between $\Sigma_{0L}^{\dagger}$ and Σ_{0L}. Specifically, we can estimate $\Sigma_{0L}^{\dagger}$ by

$$\widehat{\Sigma}_{L}^{\dagger} = \begin{pmatrix} \widehat{\Sigma}_{b} & \widehat{\Sigma}_{b,\boldsymbol{w}} R_L^T \\ R_L \widehat{\Sigma}_{\boldsymbol{w},b} & R_L \widehat{\Sigma}_{\boldsymbol{w}} R_L^T \end{pmatrix} \quad \text{given that} \quad \widehat{\Sigma}_{L} = \begin{pmatrix} \widehat{\Sigma}_{b} & \widehat{\Sigma}_{b,\boldsymbol{w}} \\ \widehat{\Sigma}_{\boldsymbol{w},b} & \widehat{\Sigma}_{\boldsymbol{w}} \end{pmatrix}, \tag{2.9}$$

where $\widehat{\Sigma}_{L}$ is the sample variance of $\widehat{\boldsymbol{\theta}}_{L}^{*}$ obtained from Stage 1 as a byproduct. Hence, combining (2.8) and (2.9), we propose the following DBI estimate:

$$\widehat{DBI}\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L)\Big) = \frac{\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T} \widehat{\Sigma}_{L,(-d)}^{\dagger} \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger}}{(n\widehat{w}_{L,d}^{\dagger})^2}, \tag{2.10}$$

where $\widehat{w}_{L,d}^{\dagger}$ is the last entry of $\widehat{\boldsymbol{\theta}}_{L}^{\dagger}$, and $\widehat{\Sigma}_{L,(-d)}^{\dagger}$ is obtained by removing the last row and last column of $\widehat{\Sigma}_{L}^{\dagger}$ defined in (2.9). The DBI estimate in (2.10) is the one we will use in the numerical experiments.

2.2.3 Relationship of DBI with Other Variability Measures

In this subsection, we discuss the relationship of DBI with two alternative variability measures.

DBI may appear to be related to the asymptotic variance of the K-CV error, i.e., $E(\psi_1)^2$ in Theorem 1. However, we want to point out that these two quantities are quite different. For example, when data are nearly separable, reasonable perturbations to the data may only lead to a small variation in the K-CV error. On the other hand, small changes in the data (especially those support points near the decision boundary) may lead to a large variation in the decision boundary which implies a large DBI. This is mainly because DBI is conceptually different from the K-CV error. In Section 2.5, we provide concrete examples to show that these two variation measures generally lead to different choices of loss functions, and the loss function with the smallest DBI often corresponds to the classifier that is more accurate and stable.

Moreover, DBI shares similar spirit of the stability-oriented measure introduced in [17]. They defined theoretical stability measures for the purpose of deriving the generalization error bound. Their stability of a classification algorithm is defined as the maximal difference of the decision functions trained from the original dataset and the leave-one-out dataset. Their stability measure mainly focuses on the variability of the decision function and hence suffers from the transformation variant issue since a scale transformation of the decision function coefficients will greatly affect the value of a decision function. On the other hand, our DBI focuses on the variability of the decision boundary and is transformation invariant.

In the experiments, we will compare our classifier selection algorithm with approaches using these two alternative variability measures. Our method achieves superior performance in both classification accuracy and stability.

2.2.4 Summary of Classifier Selection Algorithm

In this section, we summarize our two-stage classifier selection algorithm.

Algorithm 2 (Two-Stage Classifier Selection Procedure):

Input: Training sample $\mathcal{D}_n$ and a collection of candidate classifiers $\{L_j : j \in J\}$.

- Step 1. Obtain the K-CV errors $\widehat{\mathcal{D}}_j$ for each $j \in J$, and let the minimal value be $\widehat{\mathcal{D}}_t$.
- Step 2. Apply *Algorithm 1* to establish the pairwise confidence interval for each GE difference Δ_{tj}. Eliminate the classifier L_j if the corresponding confidence interval does not cover zero. Specifically, the set of potentially good classifiers is defined to be
$$\Lambda = \Big\{ j \in J : \widehat{\Delta}_{tj} - n^{-1/2}\phi_{t,j;\alpha/2} \leq 0 \Big\},$$

where $\widehat{\Delta}_{tj}$ and $\phi_{t,j;\alpha/2}$ are defined in Step 3 of *Algorithm 1*.

- Step 3. Estimate DBI for each L_j with $j \in \Lambda$ using (2.10). The optimal classifier is L_{j^*} with

$$j^* = \arg\min_{j \in \Lambda} \widehat{DBI}\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_j)\Big). \tag{2.11}$$

In Step 2, we fix the confidence level $\alpha = 0.1$ since it provides a sufficient but not too stringent confidence level. Our experiment in Section 6.1 further shows that the set Λ is quite stable against α within a reasonable range around 0.1. The optimal classifier L_{j^*} selected in (2.11) is not necessarily unique. However, according to our experiments, multiple optimal classifiers are quite uncommon. Although in principle we can also perform an additional significance test for DBI in Step 3, the related computational cost is high given that DBI is already a second-moment measure. Hence, we choose not to include this test in our algorithm.

2.3 Large-Margin Unified Machines

This section illustrates our classifier selection algorithm using the LUM [43] as an example. The LUM offers a platform unifying various large margin classifiers ranging from soft ones to hard ones. A soft classifier estimates the class conditional probabilities explicitly and makes the class prediction via the largest estimated probability, while a hard classifier directly estimates the classification boundary without a class-probability estimation [44]. For simplicity of presentation, we rewrite the class of LUM loss functions as

$$L_\gamma(u) = \begin{cases} 1-u & \text{if } u < \gamma \\ (1-\gamma)^2(\frac{1}{u-2\gamma+1}) & \text{if } u \geq \gamma, \end{cases} \tag{2.12}$$

where the index parameter $\gamma \in [0,1]$. As shown by [43], when $\gamma = 1$ the LUM loss reduces to the hinge loss of SVM, which is a typical example of hard classification; when $\gamma = 0.5$ the LUM loss is equivalent to the DWD classifier, which can be viewed as a classifier that is between hard and soft; and when $\gamma = 0$ the LUM loss becomes a soft classifier that has an interesting connection with the logistic loss. Therefore, the LUM framework approximates many of the soft and hard classifiers in the literature. Figure 2.3 displays LUM loss functions for various values of γ and compares them with some commonly used loss functions.

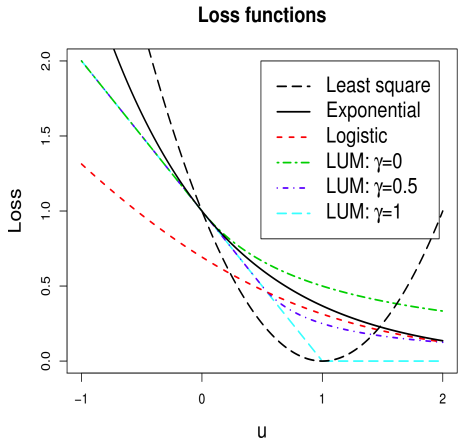

Figure 2.2. Plots of least square, exponential, logistic, and LUM loss functions with $\gamma = 0, 0.5, 1$.

In the LUM framework, we denote the true risk as $\mathcal{R}_\gamma(\boldsymbol{\theta}) = E[L_\gamma(yf(\boldsymbol{x};\boldsymbol{\theta}))]$, the true parameter as $\boldsymbol{\theta}_{0\gamma} = \arg\min_{\boldsymbol{\theta}} \mathcal{R}_\gamma(\boldsymbol{\theta})$, the GE as $D_{0\gamma}$, the empirical generalization error as $\widehat{D}_\gamma$, and the K-CV error as $\widehat{\mathcal{D}}_\gamma$. In practice, given data $\mathcal{D}_n$, LUM solves

$$\widehat{\boldsymbol{\theta}}_\gamma = \arg\min_{b,\boldsymbol{w}} \left\{ \frac{1}{n}\sum_{i=1}^{n} L_\gamma\Big(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\Big) + \frac{\lambda_n \boldsymbol{w}^T\boldsymbol{w}}{2} \right\}. \tag{2.13}$$

We next establish the asymptotic normality of $\widehat{\mathcal{D}}_\gamma$ and $\widehat{\boldsymbol{\theta}}_\gamma$ (with more explicit forms of the asymptotic variances) by verifying the conditions in Theorem 2.2.1, i.e., (L1)–(L5). In particular, we provide a set of sufficient conditions for the LUM, i.e., (L1) and (A1) below.

(A1) $\mathrm{Var}(\boldsymbol{X}|Y) \in \mathbb{R}^{d\times d}$ is a positive definite matrix for $Y \in \{1, -1\}$.

Assumption (A1) is needed to guarantee the uniqueness of the true minimizer $\boldsymbol{\theta}_{0\gamma}$. It is worth pointing out that the asymptotic normality of the estimated coefficients for SVM has also been established by Koo et al. (2008) under another set of assumptions.

Corollary 1 *Suppose that Assumptions (L1) and (A1) hold and $\lambda_n = o(n^{-1/2})$. We have, for each fixed $\gamma \in [0,1]$,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) \xrightarrow{d} N(0, \Sigma_{0\gamma}) \quad \text{as } n \to \infty, \tag{2.14}$$

where $\Sigma_{0\gamma} = H(\boldsymbol{\theta}_{0\gamma})^{-1}G(\boldsymbol{\theta}_{0\gamma})H(\boldsymbol{\theta}_{0\gamma})^{-1}$ with $G(\boldsymbol{\theta}_{0\gamma})$ and $H(\boldsymbol{\theta}_{0\gamma})$ defined in (2.31) and (2.33) in Section 2.7.5.

In practice, direct estimation of $\Sigma_{0\gamma}$ in (2.14) is difficult because of the involvement of the Dirac delta function; see Section 2.7.5 for details. Instead, we find that the perturbation-based resampling procedure proposed in Stage 1 works well.

Next we establish the asymptotic normality of $\widehat{\mathcal{D}}_\gamma$.

Corollary 2 *Suppose that the assumptions in Corollary 1 hold. We have, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\mathcal{D}}_\gamma - D_{0\gamma}) \xrightarrow{d} N\Big(0, E(\psi_{1\gamma}^2)\Big), \tag{2.15}$$

where $\psi_{1\gamma} = \frac{1}{2}|Y_1 - sign\{f(\boldsymbol{X}_1;\boldsymbol{\theta}_{0\gamma})\}| - D_{0\gamma} - \dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1} M_1(\boldsymbol{\theta}_{0\gamma})$, $\dot{d}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} E(\widehat{D}_\gamma(\boldsymbol{\theta}))$, and

$$M_1(\boldsymbol{\theta}_{0\gamma}) = -Y_1\tilde{\boldsymbol{X}}_1 I_{\{Y_1 f(\boldsymbol{X}_1;\boldsymbol{\theta}_{0\gamma})<\gamma\}} - \frac{(1-\gamma)^2 Y_1 \tilde{\boldsymbol{X}}_1 I_{\{Y_1 f(\boldsymbol{X}_1;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}}{\Big(Y_1 f(\boldsymbol{X}_1;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1\Big)^2}.$$

Corollary 2 demonstrates that the K-CV error induced from each LUM loss function yields the desirable asymptotic property under Assumptions (L1) and (A1). It can be applied to justify the perturbation-based resampling procedure for LUM as shown in Theorem 2.2.2.

2.4 Selection Consistency

This section investigates the selection consistency of our two-stage classifier selection algorithm. Selection consistency means that the selected classifier achieves the minimal GE and minimal DBI asymptotically. We focus on the selection consistency of the LUM loss functions in this section; the extension to other large-margin classifiers is straightforward.

For the LUM class, we define the set of potentially good classifiers as

$$\widehat{\Lambda}_0 = \Big\{\gamma \in [0,1] : \widehat{\mathcal{D}}_\gamma \leq \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} + n^{-1/2}\phi_{\gamma,\widehat{\gamma}_0^*;\alpha/2}\Big\}, \tag{2.16}$$

where $\widehat{\gamma}_0^* = \arg\min_{\gamma\in[0,1]} \widehat{\mathcal{D}}_\gamma$, based on $\mathcal{D}_n$. Its population version is thus defined as those classifiers achieving the minimal GE, denoted

$$\Lambda_0 = \Big\{\gamma \in [0,1] : D_{0\gamma} = D_{0\gamma_0^*}\Big\}, \tag{2.17}$$

where $\gamma_0^* = \arg\min_{\gamma\in[0,1]} D_{0\gamma}$. Literally, $\widehat{\gamma}_0^*$ and γ_0^* may not be unique. To show the selection consistency, we require an additional assumption on the Hessian matrix defined in Corollary 1:

(B1) The smallest eigenvalue of the true Hessian matrix $\lambda_{\min}(H(\boldsymbol{\theta}_{0\gamma})) \geq c_1$, and the largest eigenvalue of the true Hessian matrix $\lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma})) \leq c_2$, where the positive constants c_1, c_2 do not depend on γ.

As seen in the proof of Corollary 1, the true Hessian matrix $H(\boldsymbol{\theta}_{0\gamma})$ is positive definite for any fixed $\gamma \in [0,1]$ under Assumptions (L1) and (A1). Therefore, Assumption (B1) is slightly stronger in the uniform sense. It is required to guarantee the uniform convergence results, i.e., (2.38) and (2.40), in Appendix A.1.

Our Lemma 1 first ensures that the minimum K-CV error converges to the minimum GE at a root-n rate.

Lemma 1 *Suppose that Assumptions (L1),(A1), and (B1) hold. We have, if* $\lambda_n = o(n^{-1/2})$,

$$\Big|\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma_0^*}\Big| = O_P(n^{-1/2}). \tag{2.18}$$

In the second stage, we denote the index of the selected optimal classifier as

$$\widehat{\gamma}_0 = \arg\min_{\gamma\in\widehat{\Lambda}_0} \widehat{DBI}\Big(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)\Big), \tag{2.19}$$

and its population version as

$$\gamma_0 = \arg\min_{\gamma \in \Lambda_0} DBI\Big(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)\Big). \tag{2.20}$$

Again, $\widehat{\gamma}_0$ and γ_0 are not necessarily unique.

Theorem 2.4.1 *Suppose the assumptions in Lemma 1 hold, we have, as $N \to \infty$,*

$$\Big|\widehat{DBI}\Big(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})\Big) - DBI\Big(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})\Big)\Big| = o_P(n^{-1}). \tag{2.21}$$

Recall that N is the number of resamplings defined in Step 2 of Algorithm 1.

Theorem 2.4.1 implies that the estimated DBI of the selected classifier converges to the DBI of the true optimal classifier, which has the smallest DBI. Therefore, the proposed two-stage algorithm is able to select the classifier with the minimal DBI among those classifiers having the minimal GE. It is not uncommon to have several classifiers obtain the same minimal GE, especially when the two classes are well separable. In summary, we have shown that the selected optimal classifier has achieved the minimal GE and the minimal DBI asymptotically.

2.5 Experiments

In this section, we first demonstrate the DBI estimation procedure introduced in Section 2.2.2, and then illustrate the applicability of our classifier selection method in various simulated and real examples. In all experiments, we compare our selection procedure, denoted as "cv+dbi", with two alternative methods: 1) "cv+varcv" which is the two-stage approach selecting the loss with the minimal variance of the K-CV error in Stage 2, and 2) "cv+be" which is the two-stage approach selecting the loss with the minimal classification stability defined in Bousquet and Elisseeff (2002) in Stage 2. Stage 1 of each alternative approach is the same as ours. We consider six large-margin classifier candidates: least squares loss, exponential loss, logistic loss, and LUM with $\gamma = 0, 0.5, 1$. Recall that LUM with $\gamma = 0.5$ ($\gamma = 1$) is equivalent to DWD (SVM). In all the large-margin classifiers, the tuning parameter λ_n is selected via cross-validation.

2.5.1 Illustration

This subsection demonstrates the DBI estimation procedure and checks the sensitivity of the confidence level α in Algorithm 2.

We generated labels $y \in \{-1, 1\}$ with equal probability. Given $Y = y$, the predictor vector (x_1, x_2) was generated from a bivariate normal $N((\mu y, \mu y)^T, I_2)$ with the signal level $\mu = 0.8$.

We first illustrate the DBI estimation procedure in Section 2.2.2 by comparing the estimated DBIs with the true DBIs for various sample sizes. We varied the sample size n among 50, 100, 200, 500, and 1000. The classifier with the least squares loss was investigated due to its simplicity. Simple algebra implied that the true parameter $\boldsymbol{\theta}_{0L} = (0, 0.351, 0.351)$ and the transformed parameter $\boldsymbol{\theta}_{0L}^{\dagger} = (0, 0, 0.429)$. Furthermore, the covariance matrix Σ_{0L} and the transformed covariance matrix $\Sigma_{0L}^{\dagger}$ were computed as

$$\Sigma_{0L} = \begin{pmatrix} 0.439 & 0 & 0 \\ 0 & 0.268 & -0.170 \\ 0 & -0.170 & 0.268 \end{pmatrix} \quad \text{and} \quad \Sigma_{0L}^{\dagger} = \begin{pmatrix} 0.439 & 0 & 0 \\ 0 & 0.439 & 0 \\ 0 & 0 & 0.098 \end{pmatrix},$$

given the transformation matrix

$$R_L = \begin{pmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

Finally, plugging all these terms into (2.8) led to

$$DBI\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L)\Big) \approx \frac{3.563}{n}. \tag{2.22}$$

The left plot of Figure 2.5.1 compares the estimated DBIs in (2.10) with the true DBIs in (2.22). Clearly, they match very well for various sample sizes and their difference vanishes as the sample size increases. This experiment empirically validates the estimation procedure in Section 2.2.2.

In order to show the sensitivity of the confidence level α to the set Λ in Algorithm 2, we randomly selected one replication and display the proportion of potentially good

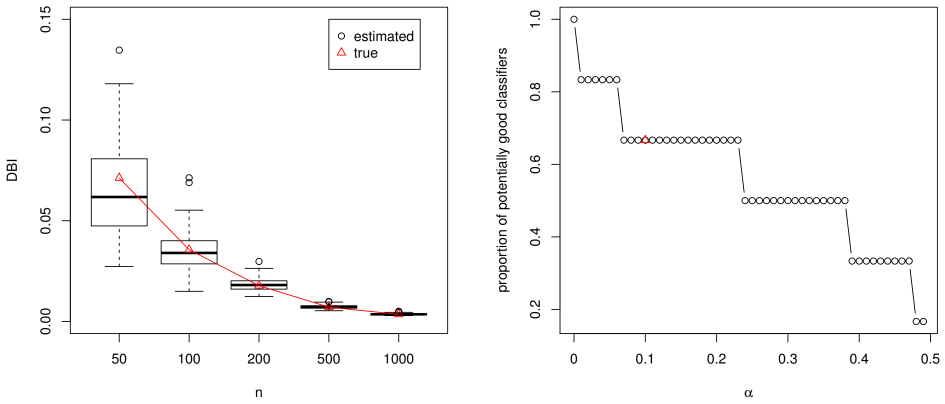

Figure 2.3. Comparison of true and estimated DBIs in Example 6.1 is shown in the left plot. The true DBIs are denoted as red triangles and the estimated DBIs from replicated experiments are illustrated by box plots. The sensitivity of confidence level α to the proportion of potentially good classifiers in Stage 1 is shown on the right.

classifiers over all six classifiers. Note that as α increases, the confidence interval for the difference of GEs will be narrower, and hence the size of Λ will be smaller. Therefore, the change of the proportion reflects exactly the change of Λ since Λ is monotone with respect to α. For each $\alpha \in \{\frac{l}{100};\ l = 0, \ldots, 50\}$, we computed the proportion of potentially good classifiers. As shown in the right plot of Figure 2.5.1, the proportion is stable in a reasonable range around 0.1.

2.5.2 Simulations

In this section, we illustrate the superior performance of our method using four simulated examples. These simulations were previously studied by [45] and [43]. In all of the simulations, the size of training data sets was 100 and that of testing data sets was 1000. All the procedures were repeated 100 times and the averaged test errors and averaged test DBIs of the selected classifier were reported.

Simulation 1: Two predictors were uniformly generated over $\{(x_1, x_2) : x_1^2+x_2^2 \leq 1\}$. The class label y was 1 when $x_2 \geq 0$ and -1 otherwise. We generated 100 samples and then contaminated the data by randomly flipping the labels of 15% of the instances.

Simulation 2: The setting was the same as Simulation 1 except that we contaminated the data by randomly flipping the labels of 25% of the instances.

Simulation 3: The setting was the same as Simulation 1 except that we contaminated the data by randomly flipping the labels of 80% of the instances whose $|x_2| \geq 0.7$.

Simulation 4: Two predictors were uniformly generated over $\{(x_1, x_2) : |x_1| + |x_2| \leq 2\}$. Conditionally on $X_1 = x_1$ and $X_2 = x_2$, the class label y took 1 with probability $e^{3(x_1+x_2)}/(1 + e^{3(x_1+x_2)})$ and -1 otherwise.

We first demonstrate the mechanism of our proposed method for one repetition of Simulation 1. As shown in the upper left plot of Figure 2.5.2, exponential loss and LUMs with $\gamma = 0.5$ or 1 are potentially good classifiers in Stage 1; they happen to have the same K-CV error. Their corresponding DBIs are compared in the second stage. As shown in the upper right plot of Figure 2.5.2, LUM with $\gamma = 0.5$ gives the minimal DBI and is selected as the final classifier. In this example, although exponential loss also gives the minimal K-CV error, its decision boundary is unstable compared to that of LUM with $\gamma = 0.5$. This shows that the K-CV estimate itself is not sufficient for classifier comparison, since it ignores the variation in the classifier. To show that our DBI estimation is reasonable, we display the perturbed decision boundaries for these three potentially good classifiers on the bottom of Figure 2.5.2. The relationship among their instabilities is precisely captured by our DBI estimate: compared with the exponential loss and LUM with $\gamma = 1$, LUM with $\gamma = 0.5$ is more stable.

We report the averaged test errors and averaged test DBIs of the classifier selected from our method as well as two alternative approaches, see Table 2.1. In all four simulated examples, our "cv+dbi" achieves the smallest test errors, while the difference

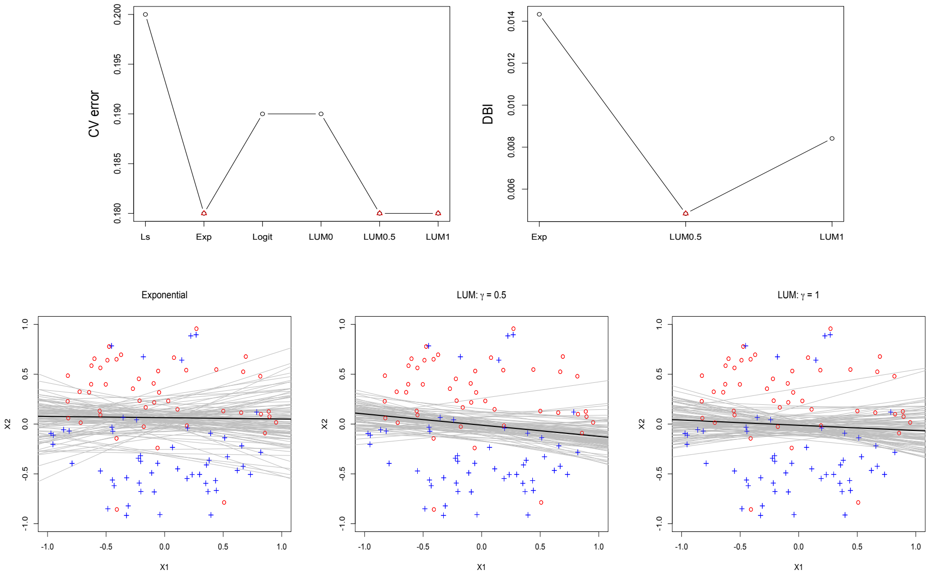

Figure 2.4. The K-CV error, the DBI estimate, and the perturbed decision boundaries in Simulation 1 with flipping rate 15%. The minimal K-CV error and minimal DBI estimate are indicated with red triangles. The labels Ls, Exp, Logit, LUM0, LUM0.5, and LUM1 refer to least squares loss, exponential loss, logistic loss, and LUM loss with index $\gamma = 0, 0.5, 1$, respectively.

among test errors of all algorithms is generally not significant. This phenomenon of indistinguishable test errors agrees with the fact that all methods are the same during the first stage and those left from Stage 1 are all potentially good in terms of classification accuracy. However, our "cv+dbi" is able to choose the classifiers with minimal test DBIs in all simulations and the improvements over other algorithms are significant. Overall, our method is able to choose the classifier with outstanding performance in both classification accuracy and stability.

2.5.3 Real Examples

In this subsection, we compare our method with the alternatives on two real datasets in the UCI Machine Learning Repository [46].

Table 2.1.
The averaged test errors and averaged test DBIs (multiplied by 100) of all methods: "cv+varcv" is the two-stage approach which selects the loss with the minimal variance of the K-CV error in Stage 2; "cv+be" is the two-stage approach which in Stage 2 selects the loss with the minimal classification stability defined in Bousquet and Elisseeff (2002); "cv+dbi" is our method. The smallest value in each case is given in bold. Standard errors are given in subscript.

Simulations		cv+varcv	cv+be	cv+dbi
Sim 1	Error	$0.191_{0.002}$	$0.194_{0.002}$	$\mathbf{0.190}_{0.002}$
	DBI	$0.139_{0.043}$	$0.135_{0.019}$	$\mathbf{0.081}_{0.002}$
Sim 2	Error	$0.296_{0.002}$	$0.303_{0.003}$	$\mathbf{0.295}_{0.002}$
	DBI	$0.291_{0.044}$	$0.318_{0.036}$	$\mathbf{0.229}_{0.012}$
Sim 3	Error	$0.218_{0.006}$	$0.234_{0.006}$	$\mathbf{0.209}_{0.004}$
	DBI	$0.124_{0.008}$	$0.291_{0.037}$	$\mathbf{0.107}_{0.003}$
Sim 4	Error	$0.120_{0.001}$	$0.121_{0.001}$	$\mathbf{0.119}_{0.001}$
	DBI	$0.884_{0.207}$	$0.414_{0.106}$	$\mathbf{0.235}_{0.038}$

The first data set is the liver disorders data set (*liver*) which consists of 345 samples with 6 variables of blood test measurements. The class label splits the data into 2 classes with sizes 145 and 200. The second data set is the breast cancer data set (*breast*) which consists of 683 samples after removing missing values [47]. Each sample has 10 experimental measurement variables and one binary class label indicating whether the sample is benign or malignant. These 683 samples arrived periodically as Dr. Wolberg reported his clinical cases. In total, there are 8 groups of samples which reflect the chronological order of the data. It is expected that a good classification procedure should generate a classifier that is stable across these groups of samples.

For each dataset, we randomly split the data into 2/3 training samples and 1/3 testing samples, and reported the averaged test errors and averaged test DBIs based on all classifier selection algorithms over 50 replications, see Table 2.2. Compared with the alternatives, our "cv+dbi" method obtains significant improvements in DBIs and simultaneously attains minimal test errors in both real data sets. This indicates that the proposed method could serve as a practical tool for selecting a most accurate and stable classifier.

Table 2.2.
The averaged test errors and averaged test DBIs of all methods in real example. The smallest value in each case is given in bold. Standard errors are given in subscript.

Data		cv+varcv	cv+be	cv+dbi
Liver	Error	$0.331_{0.006}$	$0.335_{0.006}$	$\mathbf{0.327}_{0.006}$
	DBI	$0.140_{0.013}$	$0.157_{0.024}$	$\mathbf{0.113}_{0.012}$
Breast	Error	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$	$\mathbf{0.038}_{0.002}$
	DBI	$0.388_{0.066}$	$0.152_{0.028}$	$\mathbf{0.124}_{0.023}$

2.6 Nonlinear Extension

The extension of our two-stage algorithm to nonlinear classifiers contains two aspects: (1) asymptotic normality of the K-CV error in Stage 1; (2) the application of DBI in Stage 2. The former is still valid due to [48], and the latter is feasible by mapping the nonlinear decision boundaries to a higher dimensional space where the projected decision boundaries are linear. Details of these two extensions are as follows.

Extension of Stage 1: We first modify several key concepts. The loss $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ is convex if it is convex in its third argument for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$. A reproducing kernel Hilbert space (RKHS) H is a space of functions $f : \mathcal{X} \to \mathbb{R}$ which is generated by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Here the kernel k could be a linear kernel, a Gaussian RBF kernel, or a polynomial kernel.

Given i.i.d training samples $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}$ drawn from $P = (X, Y)$, the empirical function $f_{L,\mathcal{D}_n,\lambda_n}$ solves

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i)) + \lambda_n \|f\|_{\mathcal{H}}^2.$$

In the nonparametric case, the optimization problem of minimizing population risk is ill-posed because a solution is not necessarily unique, and small changes in P may have large effects on the solution. Therefore it is common to impose a bound on the complexity of the predictor and estimate a smoother approximation to the population version [48]. For a fixed $\lambda_0 \in (0, \infty)$, we denote f_{L,P,λ_0} as the population function which solves

$$\min_{f \in \mathcal{H}} \int L(x, y, f(x)) P(d(x, y)) + \lambda_0 \|f\|_{\mathcal{H}}^2.$$

The following conditions are assumed in [48] to prove the asymptotic normality of the estimated kernel decision function.

(N1) The loss L is a convex, P-square-integrable Nemitski loss function of order $p \in [1, \infty)$. That is, there is a P-square-integrable function $b : \mathcal{X} \times \mathcal{Y} \to R$ such that $|L(x, y, t)| \leq b(x, y) + |t|^p$ for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$.

(N2) The partial derivatives $L^{'}(x, y, t) := \frac{\partial L}{\partial t}(x, y, t)$ and $L^{''}(x, y, t) := \frac{\partial^2 L}{\partial^2 t}(x, y, t)$ exist for every $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$ and are continuous.

(N3) For every $a \in (0, \infty)$, there is $b_a^{'} \in L_2(P)$ and $b_a^{''} \in [0, \infty)$ such that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\sup_{t \in [-a,a]} |L^{'}(x, y, t)| \leq b_a^{'}(x, y)$ and $\sup_{t \in [-a,a]} |L^{''}(x, y, t)| \leq b_a^{''}$.

Proposition 1 *(Theorem 3.1, [48]) Under Assumptions (N1)-(N3) and* $\lambda_n = \lambda_0 + o(n^{-1/2})$*, for every* $\lambda_0 \in (0, \infty)$*, there is a tight, Borel-measurable Gaussian process* $\mathbb{H} : \Omega \to H$ *such that* $\sqrt{n}\Big(f_{L,\mathcal{D}_n,\lambda_n} - f_{L,P,\lambda_0}\Big) \to \mathbb{H}$.

Remark 2 *Among the loss functions considered in this paper, the least squares, exponential, and logistic losses all satisfy the assumptions (N1)-(N3), while the LUM loss is not differentiable and does not satisfy Assumption (N2). However, [48] showed that any Lipschitz-continuous loss function (e.g. LUM loss) can always be modified as* *a differentiable $\epsilon-$version of the loss function such that the assumptions (N1)-(N3) are satisfied; see Remark 3.5 in [48].*

In the nonlinear case, the GE D_{0L} and the K-CV error $\widehat{\mathcal{D}}_L$ are modified accordingly. The asymptotic normality of $\mathcal{W}_L = \sqrt{n}(\widehat{\mathcal{D}}_L - D_{0L})$ follows from Proposition 1, Corollary 3.3 in [49], and a slight modification of the proof of our Theorem 1. Then a perturbation-based resampling approach can be constructed analogously to Algorithm 1.

Extension of Stage 2: The concept of DBI is defined for linear decision boundaries. In order to measure the instability of nonlinear decision boundaries, we can map the nonlinear decision boundaries to a higher dimensional space where the projected decision boundaries are linear.

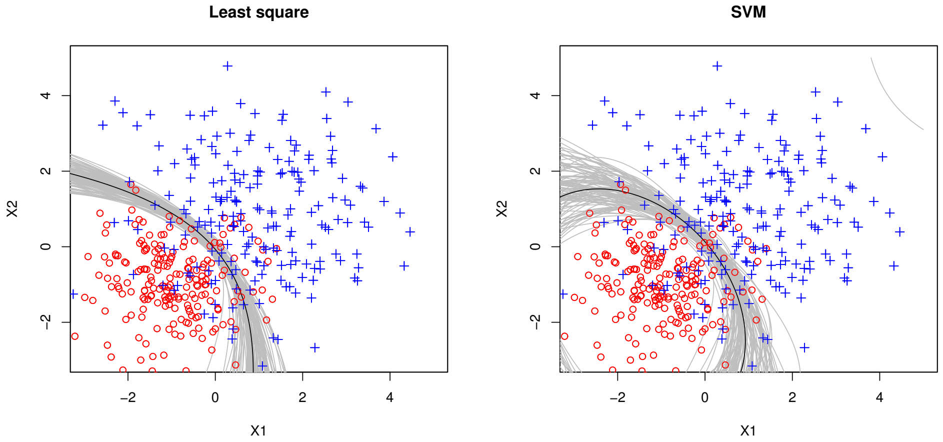

Figure 2.5. The nonlinear perturbed decision boundaries for the least squares loss (left) and SVM (right) in the bivariate normal example with unequal variances.

Here we illustrate the estimation procedure via a bivariate normal example with sample size $n = 400$. Assume the underlying distributions of the two classes are $f_1 = N((-1,-1)^T, I_2)$ and $f_2 = N((1,1)^T, 2I_2)$ with equal prior probability. We map the input $\{x_1, x_2\}$ to the polynomial basis $\{x_1, x_2, x_1x_2, x_1^2, x_2^2\}$ and fit the linear large-margin classifiers using the expanded inputs. The instability of the original nonlinear decision boundary boils down to the instability of the linear boundaries in the expanded space. Figure 2.6 demonstrates 100 nonlinear perturbed decision boundaries for the least squares and SVM losses, where the former is visually more stable than the latter. Indeed, their corresponding DBI estimations in the expanded space capture this relationship in that the estimated DBI of the former is 0.017 and that of the latter is 0.354. ■

2.7 Technical Proofs

In the section, we provide proofs of all theorems, calculation of transformation matrix, and detailed estimation of DBI.

2.7.1 Proof of Theorem 2.2.1:

Before we prove Theorem 2.2.1, we show an intermediate result in Lemma 2.

Lemma 2 *Suppose Assumptions (L1)–(L3) hold and* $\lambda_n = o(n^{-1/2})$*. Then we have* $\widehat{\boldsymbol{\theta}}_L \xrightarrow{P} \boldsymbol{\theta}_{0L}$ *and* $\widehat{D}_L \xrightarrow{P} D_{0L}$ *as* $n \to \infty$*.*

To show $\widehat{\boldsymbol{\theta}}_L \to \boldsymbol{\theta}_{0L}$, we apply Theorem 5.7 of van der Vaart (1998). Firstly, we show that, uniformly in $\boldsymbol{\theta}$, the empirical risk $O_{nL}(\boldsymbol{\theta})$ converges to the true risk $\mathcal{R}_L(\boldsymbol{\theta})$ in probability. Assumption (L3) guarantees that the loss function $L(yf(\boldsymbol{x};\boldsymbol{\theta}))$ is convex in $\boldsymbol{\theta}$, and it is easy to see that $O_{nL}(\boldsymbol{\theta})$ converges to $\mathcal{R}_L(\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$. Then we have $\sup_{\boldsymbol{\theta}} |O_{nL}(\boldsymbol{\theta}) - \mathcal{R}_L(\boldsymbol{\theta})| \to 0$ in probability by uniform convergence Theorem for convex functions in Pollard (1991). Secondly, according to assumption (L2), we have that $\mathcal{R}_L(\boldsymbol{\theta})$ has a unique minimizer $\boldsymbol{\theta}_{0L}$. Therefore, we know that $\widehat{\boldsymbol{\theta}}_L$ converges to $\boldsymbol{\theta}_{0L}$ in probability. The consistency of $\widehat{D}(\widehat{\boldsymbol{\theta}}_L)$ can be obtained by the uniform law of large numbers. According to Assumption (L1), $p(\boldsymbol{x})$ is continuously differentiable, and hence $|y - \text{sign}\{f(\boldsymbol{x};\boldsymbol{\theta})\}| = |y - \text{sign}\{\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}\}|$ is continuous in each $\boldsymbol{\theta}$ for almost all $\boldsymbol{x}$. This together with $|y - \text{sign}\{f(\boldsymbol{x};\boldsymbol{\theta})\}| \leq 2$ leads to uniform convergence

$\sup_{\boldsymbol{\theta}} |\widehat{D}(\boldsymbol{\theta}) - \frac{1}{2}E|y_0 - \text{sign}\{f(\boldsymbol{x}_0;\boldsymbol{\theta})\}|| \to 0$. Therefore, we have $\widehat{D}(\widehat{\boldsymbol{\theta}}_L) \to D_{0L}$ in probability. This concludes the proof of Lemma 2. ■

Proof of Theorem 2.2.1:

We next prove (2.4) in three steps. Let $M_i(\boldsymbol{\theta}_{0L}) = \nabla_{\boldsymbol{\theta}} L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}$. In step 1, we show that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L}) = -n^{-1/2} H(\boldsymbol{\theta}_{0L})^{-1} \sum_{i=1}^{n} M_i(\boldsymbol{\theta}_{0L}) + o_P(1) \tag{2.23}$$

by applying Theorem 2.1 in Hjort and Pollard (1993). Denote $Z = (\boldsymbol{X}^T, Y)$ and $\Delta\boldsymbol{\theta} = (\Delta b, \Delta \boldsymbol{w}^T)^T$. Taylor expansion leads to

$$L(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0L} + \Delta\boldsymbol{\theta})) - L(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0L})) = M(\boldsymbol{\theta}_{0L})^T \Delta\boldsymbol{\theta} + R(Z, \Delta\boldsymbol{\theta}), \tag{2.24}$$

where

$$\begin{aligned} M(\boldsymbol{\theta}_{0L}) &= \nabla_{\boldsymbol{\theta}} L(Yf(\boldsymbol{X};\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}} \\ R(Z, \Delta\boldsymbol{\theta}) &= \frac{(\Delta\boldsymbol{\theta})^T \Big(\nabla^2_{\boldsymbol{\theta}} L(Yf(\boldsymbol{X};\boldsymbol{\theta}))\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}}\Big) \Delta\boldsymbol{\theta}}{2} + o(\|\Delta\boldsymbol{\theta}\|^2). \end{aligned}$$

According to Assumption (L1), it is easy to check that $E(M(\boldsymbol{\theta}_{0L})) = \nabla_{\boldsymbol{\theta}} \mathcal{R}_L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0L}} = 0$, and

$$E[R(Z, \Delta\boldsymbol{\theta})] = \frac{1}{2}(\Delta\boldsymbol{\theta})^T H(\boldsymbol{\theta}_{0L})(\Delta\boldsymbol{\theta}) + o(\|\Delta\boldsymbol{\theta}\|^2); \quad E[R^2(Z, \Delta\boldsymbol{\theta})] = o(\|\Delta\boldsymbol{\theta}\|^3).$$

Denote $s = (b_s, \boldsymbol{w}_s^T)^T$, $Z_i = (\boldsymbol{X}_i^T, Y_i)$, and

$$\begin{aligned} A_n(s) &= \sum_{i=1}^{n} \Big\{ L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0L} + s/\sqrt{n})) - L(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0L})) \Big\} \\ &\quad + \lambda_n (\boldsymbol{w}_{0L} + \boldsymbol{w}_s/\sqrt{n})^T (\boldsymbol{w}_{0L} + \boldsymbol{w}_s/\sqrt{n}) - \lambda_n \boldsymbol{w}_{0L}^T \boldsymbol{w}_{0L}. \end{aligned}$$

Note that $A_n(s)$ is minimized when $s = \sqrt{n}(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L})$ and $nE[R(Z, s/\sqrt{n})] = \frac{1}{2}s^T H(\boldsymbol{\theta}_{0L})s + o(\|s\|^2)$. Based on the above Taylor expansion (2.24), we have

$$\begin{aligned} A_n(s) &= \sum_{i=1}^{n} \Big\{ M_i(\boldsymbol{\theta}_{0L})^T s/\sqrt{n} + R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n}) \Big\} + nE[R(Z, s/\sqrt{n})] \\ &\quad + \lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s \\ &= U_n^T s + \frac{1}{2}s^T H(\boldsymbol{\theta}_{0L})s + o(\|s\|^2) + \sum_{i=1}^{n} \Big\{ R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n}) \Big\} \\ &\quad + \lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s, \end{aligned}$$

where $U_n = n^{-1/2}\sum_{i=1}^{n} M_i(\boldsymbol{\theta}_{0L})$. Note that $\sum_{i=1}^{n}\{R(Z_i, s/\sqrt{n}) - ER(Z_i, s/\sqrt{n})\} \to 0$, and $\lambda_n \boldsymbol{w}_s^T \boldsymbol{w}_s \to 0$ since $\lambda_n \to 0$ and $\boldsymbol{w}_s$ is bounded. In addition, Hessian matrix $H(\boldsymbol{\theta}_{0L})$ is positive definite due to Assumption (L5). Therefore, we can conclude that (2.23) holds by Theorem 2.1 in Hjort and Pollard (1993).

In step 2, we show that $W_L = \sqrt{n}\{\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - D_{0L}\} \to N(0, E(\psi_1^2))$. As shown in Jiang et al. (2008), the class of functions $\mathcal{G}_{\boldsymbol{\theta}}(\delta) = \Big\{|Y - \text{sign}\{f(\boldsymbol{X}; \boldsymbol{\theta})\}| : \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0L}\| \leq \delta\Big\}$ is a P-Donsker class for any fixed $0 < \delta < \infty$. This together with (2.23) and consistency of $\widehat{\boldsymbol{\theta}}_L$ implies that

$$\begin{aligned}
& \sqrt{n}\Big(\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - D_{0L}\Big) \\
= \;& \sqrt{n}\Big(\widehat{D}(\widehat{\boldsymbol{\theta}}_L) - \widehat{D}(\boldsymbol{\theta}_{0L})\Big) + \sqrt{n}\Big(\widehat{D}(\boldsymbol{\theta}_{0L}) - D_{0L}\Big) \\
\overset{d}{=} \;& \sqrt{n}\dot{d}(\boldsymbol{\theta}_{0L})^T(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L}) + \sqrt{n}\Big(\widehat{D}(\boldsymbol{\theta}_{0L}) - D_{0L}\Big) \\
\overset{d}{=} \;& n^{-1/2}\sum_{i=1}^{n}\Big\{\frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i; \boldsymbol{\theta}_{0L})\}| - D_{0L} - \dot{d}(\boldsymbol{\theta}_{0L})^T H(\boldsymbol{\theta}_{0L})^{-1} M_1(\boldsymbol{\theta}_{0L})\Big\} \\
= \;& n^{-1/2}\sum_{i=1}^{n}\psi_i \xrightarrow{d} N(0, E(\psi_1^2)),
\end{aligned}$$

where " $\overset{d}{=}$ " means asymptotical equivalence in the distributional sense.

In step 3, the distribution of $\mathcal{W}_L = n^{1/2}\{\widehat{\mathcal{D}}_L - D_{0L}\}$ is asymptotically equivalent to that of W_L as shown in Theorem 3 in Jiang et al. (2008). This concludes the proof of Theorem 2.2.1. ■

2.7.2 Proof of Theorem 2.2.2

According to Appendix D in Jiang et al. (2008), we have

$$W_1^* \overset{d}{=} n^{-1/2}\sum_{i=1}^{n}\psi_{i1}(G_i - 1) \;\text{ and }\; W_2^* \overset{d}{=} n^{-1/2}\sum_{i=1}^{n}\psi_{i2}(G_i - 1),$$

where $\psi_{ij} = \frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i; \boldsymbol{\theta}_{0j})\}| - D_{0j} - \dot{d}(\boldsymbol{\theta}_{0j})^T H(\boldsymbol{\theta}_{0j})^{-1} M_i(\boldsymbol{\theta}_{0j})$, for $j = 1, 2$. Recall that " $\overset{d}{=}$ " means the distributional equivalence. As shown in Jiang et al.

(2008), conditional on the data, W_j^* converges to a normal with mean 0 and variance $n^{-1}\sum_{i=1}^n \psi_{ij}^2$ for $j = 1, 2$. Note that

$$W_2^* - W_1^* \overset{d}{=} n^{-1/2}\sum_{i=1}^n (\psi_{i2} - \psi_{i1})(G_i - 1).$$

Here, $(\psi_{i2} - \psi_{i1})$'s, $i = 1, \ldots, n$, are i.i.d random vectors with $E(\psi_{i2} - \psi_{i1}) = 0$ and $E|\psi_{i2} - \psi_{i1}|^2 < \infty$. Independent of $(\psi_{i2} - \psi_{i1})$, $(G_i - 1)$'s are i.i.d random variables with mean 0 and variance 1. Since $(\psi_{i2} - \psi_{i1})$ depends on the sample $(\boldsymbol{x}_i, y_i)$, Lemma 2.9.5 in van der Vaart and Wellner (1996) implies that, conditional on the data,

$$n^{-1/2}\sum_{i=1}^n (\psi_{i2} - \psi_{i1})(G_i - 1) \overset{d}{\Longrightarrow} N(0, Var(\psi_{12} - \psi_{11})). \tag{2.25}$$

Next, as shown in Theorem 2.2.1, $W_1 \overset{d}{=} n^{-1/2}\sum_{i=1}^n \psi_{i1}$ and $W_2 \overset{d}{=} n^{-1/2}\sum_{i=1}^n \psi_{i2}$, therefore,

$$W_2 - W_1 \overset{d}{=} n^{-1/2}\sum_{i=1}^n (\psi_{i2} - \psi_{i1}) \xrightarrow{d} N(0, Var(\psi_{12} - \psi_{11})).$$

This together with (2.25) and the asymptotic equivalence of W_L and $\mathcal{W}_L$ (Jiang et al. 2008) lead to the asymptotic equivalence between $\mathcal{W}_{\Delta 12}$ and $W^*_{\Delta_{12}}$, which concludes the proof. ■

2.7.3 Calculation of the Transformation Matrix in Section 2.2.2

Given a d dimensional hyperplane $f(\boldsymbol{x};\boldsymbol{\theta}) = b + w_1x_1 + \cdots + w_dx_d = 0$, we aim to find a transformation matrix $R \in \mathbb{R}^{d\times d}$ such that the transformed hyperplane $f(\boldsymbol{x};\boldsymbol{\theta}^\dagger) = b^\dagger + w_1^\dagger x_1 + \cdots + w_d^\dagger x_d = 0$ is parallel to $\mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$, where $(w_1^\dagger, \cdots, w_d^\dagger)^T = R(w_1, \cdots, w_d)^T$ and $b^\dagger = b$. Here, we implicitly assume that $w_d \neq 0$.

We construct a class of linearly independent vectors spanning the hyperplane:

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ -\frac{w_1}{w_d} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ -\frac{w_2}{w_d} \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ -\frac{w_{d-1}}{w_d} \end{bmatrix}.$$

Denote these vectors as v_1, v_2,...,v_{d-1}. Then, by Gram-Schmidt process, we can produce the following orthogonal vectors $\bar{v}_1$, $\bar{v}_2$,..., $\bar{v}_{d-1}$:

$$\begin{aligned}\bar{v}_1 &= v_1,\\ \bar{v}_2 &= v_2 - \tfrac{<v_2,\bar{v}_1>}{<\bar{v}_1,\bar{v}_1>}\bar{v}_1,\\ \bar{v}_{d-1} &= v_{d-1} - \tfrac{<v_{d-1},\bar{v}_1>}{<\bar{v}_1,\bar{v}_1>}\bar{v}_1 - \cdots - \tfrac{<v_{d-1},\bar{v}_{d-2}>}{<\bar{v}_{d-2},\bar{v}_{d-2}>}\bar{v}_{d-2},\end{aligned}$$

where the inner product $< u, v >= \sum_{i=1}^d u_i v_i$ for $u = (u_1, \ldots, u_d)$ and $v = (v_1, \ldots, v_d)$. Denote $\bar{v}_d = [w_1, \cdots, w_d]^T$, which is orthogonal to every $\bar{v}_i$, $i = 1, \cdots, d-1$ by the above construction. In the end, we normalize $u_i = \bar{v}_i \|\bar{v}_i\|^{-1}$ for $i = 1, \cdots, d$, and define the orthogonal transformation matrix R as $[u_1, \ldots, u_d]^T$. By some elementary calculation, we can verify that that $w_i^\dagger = 0$ for $i = 1, \cdots, d-1$ but $w_d^\dagger \neq 0$ under the above construction. Therefore, the transformed hyperplane $f(\boldsymbol{x}; \boldsymbol{\theta}^\dagger)$ is parallel to $\mathcal{X}_1, \ldots, \mathcal{X}_{d-1}$. ■

2.7.4 Approximation of DBI

We propose an approximate version of DBI, i.e., (2.8), which can be easily estimated in practice.

According to (2.7), we can calculate $DBI(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L))$ as

$$E\left[\tilde{\boldsymbol{X}}_{(-d)}^{\dagger T} Var\Big(\widehat{\boldsymbol{\eta}}_L^{\dagger *} | \boldsymbol{X}_{(-d)}^\dagger\Big) \tilde{\boldsymbol{X}}_{(-d)}^\dagger\right], \tag{2.26}$$

where $\tilde{\boldsymbol{X}}_{(-d)}^\dagger = (1, \boldsymbol{X}_{(-d)}^{\dagger T})^T$ and $\widehat{\boldsymbol{\eta}}_L^{\dagger *} = \Big(-\widehat{b}_L^{\dagger *}/\widehat{w}_{L,d}^{\dagger *}, -\widehat{w}_{L,1}^{\dagger *}/\widehat{w}_{L,d}^{\dagger *} \ldots, -\widehat{w}_{L,d-1}^{\dagger *}/\widehat{w}_{L,d}^{\dagger *}\Big)$. To further simplify (2.26), we need the following theorem as an intermediate step.

Theorem 2.7.1 *Suppose that Assumptions (L1)–(L5) hold and* $\lambda_n = o(n^{-1/2})$. *We have, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L - \boldsymbol{\theta}_{0L}) \xrightarrow{d} N(0, \Sigma_{0L}), \tag{2.27}$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L^* - \widehat{\boldsymbol{\theta}}_L) \overset{d}{\Longrightarrow} N(0, \Sigma_{0L}) \quad \textit{conditional on } \mathcal{D}_n, \tag{2.28}$$

where $\Sigma_{0L} = H(\boldsymbol{\theta}_{0L})^{-1}G(\boldsymbol{\theta}_{0L})H(\boldsymbol{\theta}_{0L})^{-1}$. *After data transformation, we have, as* $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L^{\dagger} - \boldsymbol{\theta}_{0L}^{\dagger}) \quad \xrightarrow{d} \quad N(0, \Sigma_{0L}^{\dagger}), \tag{2.29}$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_L^{\dagger *} - \widehat{\boldsymbol{\theta}}_L^{\dagger}) \quad \overset{d}{\Longrightarrow} \quad N(0, \Sigma_{0L}^{\dagger}) \quad \textit{conditional on } \mathcal{D}_n^{\dagger}, \tag{2.30}$$

where $\boldsymbol{\theta}_{0L}^{\dagger} = (b_{0L}, \boldsymbol{w}_{0L}^T R_L^T)^T$ *and*

$$\Sigma_{0L}^{\dagger} = \begin{pmatrix} \Sigma_b & \Sigma_{b,\boldsymbol{w}} R_L^T \\ R_L \Sigma_{\boldsymbol{w},b} & R_L \Sigma_{\boldsymbol{w}} R_L^T \end{pmatrix} \quad \textit{if we partition } \Sigma_{0L} \textit{ as } \begin{pmatrix} \Sigma_b & \Sigma_{b,\boldsymbol{w}} \\ \Sigma_{\boldsymbol{w},b} & \Sigma_{\boldsymbol{w}} \end{pmatrix}.$$

We omit the proof of Theorem 2.7.1 since (2.27) and (2.28) directly follow from (2.23) and Appendix D in Jiang et al. (2008), and (2.29) and (2.30) follow from the Delta method.

Let $\widehat{\boldsymbol{\eta}}_L^{\dagger} = \Big(- \widehat{b}_L^{\dagger}/\widehat{w}_{L,d}^{\dagger}, -\widehat{w}_{L,1}^{\dagger}/\widehat{w}_{L,d}^{\dagger} \ldots, -\widehat{w}_{L,d-1}^{\dagger}/\widehat{w}_{L,d}^{\dagger}\Big)$. According to (2.29) and (2.30), we know that $Var(\widehat{\boldsymbol{\eta}}_L^{\dagger *} | \boldsymbol{X}_{(-d)}^{\dagger})$ is a consistent estimate of $Var(\widehat{\boldsymbol{\eta}}_L^{\dagger})$ because $\widehat{\boldsymbol{\eta}}_L^{\dagger *}$ and $\widehat{\boldsymbol{\eta}}_L^{\dagger}$ can be written as the same function of $\widehat{\boldsymbol{\theta}}_L^{\dagger *}$ and $\widehat{\boldsymbol{\theta}}_L^{\dagger}$, respectively. Hence, we claim that

$$DBI\Big(S(\boldsymbol{X}; \widehat{\boldsymbol{\theta}}_L)\Big) \approx E\left(\tilde{\boldsymbol{X}}_{(-d)}^{\dagger T} Var(\widehat{\boldsymbol{\eta}}_L^{\dagger}) \tilde{\boldsymbol{X}}_{(-d)}^{\dagger}\right).$$

Furthermore, we can approximate $Var(\widehat{\boldsymbol{\eta}}_L^{\dagger})$ by $(w_{L,d}^{\dagger})^{-2}[n^{-1}\Sigma_{0L,(-d)}^{\dagger}]$, where $n^{-1}\Sigma_{0L,(-d)}^{\dagger}$ is the asymptotic variance of the first d dimensions of $\widehat{\boldsymbol{\theta}}_L^{\dagger}$, since $\widehat{w}_{L,d}^{\dagger}$ asymptotically follows a normal distribution with mean $w_{L,d}^{\dagger}$ and variance converging to 0 as n grows (Hinkley, 1969). Finally, we get the desirable approximation (2.8) for DBI. ■

2.7.5 Proof of Corollary 1

It suffices to show that (A1) and (L1) imply Assumptions (L2)-(L5).

(L2). We first show that the minimizer $\boldsymbol{\theta}_{0\gamma}$ exists for each fixed γ. It is easy to see that $\mathcal{R}_{\gamma}(\boldsymbol{\theta})$ is continuous w.r.t. $\boldsymbol{\theta}$. We next show that, for any large enough M, the closed set $S(M) = \Big\{\boldsymbol{\theta} \in R^d : \mathcal{R}_{\gamma}(\boldsymbol{\theta}) \leq M\Big\}$ is bounded. When $yf(\boldsymbol{x}, \boldsymbol{\theta}) < \gamma$, we need to show $S(M) = \Big\{\boldsymbol{\theta} \in R^d : E[1 - Yf(\boldsymbol{X}; \boldsymbol{\theta})] \leq M\Big\}$ is contained in a box around the origin. Denote e_j as the vector with one in the j-th component and zero otherwise. Motivated by Rocha et al. (2009), we can show that, for any M, there exists a $\alpha_{j,M}$ such that any $\boldsymbol{\theta}$ satisfying $| < \boldsymbol{\theta}, e_j > | > \alpha_{j,M}$ leads to $E[(1-Yf(\boldsymbol{X};\boldsymbol{\theta})I_{(Yf(\boldsymbol{X};\boldsymbol{\theta})<\gamma)})] > M$. Similarly, when $yf(\boldsymbol{x},\boldsymbol{\theta}) \geq \gamma$, $S(M)$ is contained in a sphere around the origin, that is, for any M, there exists a σ such that any $\boldsymbol{\theta}$ satisfying $| < \boldsymbol{\theta}, \boldsymbol{\theta} > | > \sigma$ leads to $E[\frac{(1-\gamma)^2}{Yf(\boldsymbol{X};\boldsymbol{\theta})-2\gamma+1}I_{(Yf(\boldsymbol{X};\boldsymbol{\theta})\geq\gamma)})] > M$. These imply the existence of $\boldsymbol{\theta}_{0\gamma}$. The uniqueness of $\boldsymbol{\theta}_{0\gamma}$ is implied by the positive definiteness of Hessian matrix as verified in (L5) below.

(L3). The loss function $L_\gamma(yf(\boldsymbol{x};\boldsymbol{\theta}))$ is convex by noting that two segments of $L_\gamma(yf(\boldsymbol{x};\boldsymbol{\theta}))$ are convex, and the sum of convex functions is convex.

(L4). The loss function $L_\gamma(yf(\boldsymbol{x};\boldsymbol{\theta}))$ is not differentiable only on the set $\{\boldsymbol{x} : \tilde{\boldsymbol{x}}^T\boldsymbol{\theta} = \gamma \text{ or } \tilde{\boldsymbol{x}}^T\boldsymbol{\theta} = -\gamma\}$, which is assumed to be a zero probability event. Therefore, with probability one, it is differentiable with

$$\nabla_{\boldsymbol{\theta}} L_\gamma(yf(\boldsymbol{x};\boldsymbol{\theta})) = -\tilde{\boldsymbol{x}}yI_{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}<\gamma)} - \frac{(1-\gamma)^2\tilde{\boldsymbol{x}}y}{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}-2\gamma+1)^2}I_{(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}\geq\gamma)},$$

and hence

$$\begin{aligned} G(\boldsymbol{\theta}_{0\gamma}) &= E\Big[\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}} L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))^T|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}}\Big] \\ &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T Y^2 I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma)} + \frac{(1-\gamma)^4\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T Y^2}{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^4}I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}\Big\} \\ &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})A(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) + (1-p(\boldsymbol{X}))B(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big]\Big\}, \qquad (2.31) \end{aligned}$$

where $A(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})$ and $B(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})$ are defined as

$$\begin{aligned} A(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) &= I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma)} + \frac{(1-\gamma)^4}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^4}I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}; \\ B(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) &= I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma)} + \frac{(1-\gamma)^4}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^4}I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}. \end{aligned}$$

Obviously, $|A(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})|$ and $|B(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})|$ are both bounded by one. Therefore, $G(\boldsymbol{\theta}_{0\gamma}) < \infty$ based on the moment condition of $\boldsymbol{X}$.

(L5). We prove it in three steps. First, we show the risk $R_\gamma(\boldsymbol{\theta})$ is bounded. For each fixed $\gamma \in [0,1]$,

$$\begin{aligned}
\mathcal{R}_\gamma(\boldsymbol{\theta}) &\leq E\Big|L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))\Big| \\
&= E\Big|(1-Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta})I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma)} + \frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1}I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma)}\Big| \\
&\leq E\Big|(1-Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta})I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma)}\Big| + E\Big|\frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1}I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma)}\Big| \\
&\leq E\Big|(1-Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta})I_{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<1)}\Big| + |1-\gamma| < \infty, \qquad (2.32)
\end{aligned}$$

where the first term in (2.32) was shown to be bounded in Rocha et al. (2009).

Next, we derive the form of Hessian matrix. The moment assumption of $\boldsymbol{x}$ and the inequality $(y\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}-2\gamma+1)^2 \leq (1-\gamma)^2$ lead to $E|\nabla_{\boldsymbol{\theta}}L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))| \leq E|-\tilde{\boldsymbol{X}}Y|+E|-\tilde{\boldsymbol{X}}Y| \leq 2E|\tilde{\boldsymbol{X}}| < \infty$. Then, dominated convergence theorem implies that $\nabla_{\boldsymbol{\theta}}\mathcal{R}_\gamma(\boldsymbol{\theta}) = E[\nabla_{\boldsymbol{\theta}}L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))]$. Hence, the Hessian matrix equals $\nabla_{\boldsymbol{\theta}}E[\nabla_{\boldsymbol{\theta}}L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))]$. We next derive the form of $E[\nabla_{\boldsymbol{\theta}}L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))]$. Note that

$$\begin{aligned}
&E[\nabla_{\boldsymbol{\theta}}L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}))] \\
&= E\Big[-\tilde{\boldsymbol{X}}YI_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}Y}{(Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2}I_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big] \\
&= E\Big\{I_{\{Y=1\}}\Big[-\tilde{\boldsymbol{X}}I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2}I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big] \\
&\qquad + I_{\{Y=-1\}}\Big[\tilde{\boldsymbol{X}}I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} + \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}+2\gamma-1)^2}I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&= E\Big\{p(\boldsymbol{X})\Big[-\tilde{\boldsymbol{X}}I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} - \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}-2\gamma+1)^2}I_{\{\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&\qquad + E\Big\{(1-p(\boldsymbol{X}))\Big[\tilde{\boldsymbol{X}}I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}<\gamma\}} + \frac{(1-\gamma)^2\tilde{\boldsymbol{X}}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}+2\gamma-1)^2}I_{\{-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}\geq\gamma\}}\Big]\Big\} \\
&= E_1(\boldsymbol{\theta}) + E_2(\boldsymbol{\theta}).
\end{aligned}$$

After tedious algebra, we can show

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}E_1(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}} &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T p(\boldsymbol{X})C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big\}, \\
\nabla_{\boldsymbol{\theta}}E_2(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}} &= E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T (1-p(\boldsymbol{X}))D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big\},
\end{aligned}$$

where

$$\begin{aligned} C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) &= \delta(\gamma-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}) - \frac{(1-\gamma)^2\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-\gamma)}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^2} + \frac{2(1-\gamma)^2 I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^3}, \\ D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) &= \delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}) - \frac{(1-\gamma)^2\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+\gamma)}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^2} - \frac{2(1-\gamma)^2 I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^3}, \end{aligned}$$

and $\delta(\cdot)$ is the Dirac delta function. Hence, we can write the Hessian matrix as

$$H(\boldsymbol{\theta}_{0\gamma}) = E\Big\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})C(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma}) + (1-p(\boldsymbol{X}))D(\boldsymbol{X},\boldsymbol{\theta}_{0\gamma})\Big]\Big\}. \tag{2.33}$$

Finally, we establish the positive definiteness of $H(\boldsymbol{\theta}_{0\gamma})$. We write $H(\boldsymbol{\theta}_{0\gamma}) = R_1(\boldsymbol{\theta}_{0\gamma}) + R_2(\boldsymbol{\theta}_{0\gamma})$ with

$$\begin{aligned} R_1(\boldsymbol{\theta}_{0\gamma}) &= E\Bigg\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[p(\boldsymbol{X})\delta(\gamma-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}) + (1-p(\boldsymbol{X}))\delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})\Big]\Bigg\}, \\ R_2(\boldsymbol{\theta}_{0\gamma}) &= E\Bigg\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\Big[(1-p(\boldsymbol{X}))\Big(\frac{\delta(\gamma+\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^2} - \frac{2I_{(-\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}+2\gamma-1)^3}\Big) \\ &\qquad - p(\boldsymbol{X})\Big(\frac{\delta(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-\gamma)}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^2} - \frac{2I_{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\leq\gamma)}}{(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}-2\gamma+1)^3}\Big)\Big]\Bigg\}(1-\gamma)^2. \end{aligned}$$

Next we show the positive definiteness of $R_1(\boldsymbol{\theta}_{0\gamma})$. Let $f_{\boldsymbol{x}}$ be the density of $\tilde{\boldsymbol{x}}^T\boldsymbol{\theta}_{0\gamma}$. According to Lemma 9 in Rocha et al. (2009), Assumption $(L1)$ implies that $f_{\boldsymbol{x}}(\gamma) > 0$, $f_{\boldsymbol{x}}(-\gamma) > 0$, $P(Y=1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma) > 0$, and $P(Y=-1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma) > 0$. Note that $R_1(\boldsymbol{\theta}_{0\gamma})$ can be rewritten as

$$\begin{aligned} R_1(\boldsymbol{\theta}_{0\gamma}) &= E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma\Big]P(Y=1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma)f_{\boldsymbol{X}}(\gamma) \\ &\quad + E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=-1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma\Big]P(Y=-1|\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma)f_{\boldsymbol{X}}(-\gamma). \end{aligned}$$

In order to show $R_1(\boldsymbol{\theta}_{0\gamma})$ is positive definite, it remains to show that $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=\gamma\Big]$ or $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y=-1,\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}=-\gamma\Big]$ is strictly positive definite. Rocha et al. (2009) showed that

$$\begin{aligned} & E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y,\tilde{\boldsymbol{X}}^T\theta_{0\gamma}=\gamma\Big] = E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y,\boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}} = \frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big] \\ \succeq\ & \Big(\frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big)^2(v_{\boldsymbol{w}_{0\gamma}}v_{\boldsymbol{w}_{0\gamma}}^T) + Var\Big(\boldsymbol{X}|Y,\boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}} = \frac{\gamma-b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big), \end{aligned} \tag{2.34}$$

where $S_1 \succeq S_2$ means $S_1 - S_2$ is positive semi-definite, and $v_{\boldsymbol{w}_{0\gamma}} = \frac{\boldsymbol{w}_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}$. By assumption $(A1)$, $Var(\boldsymbol{X}|Y)$ is non-singular, and hence $Var\Big(\boldsymbol{X}|Y, \boldsymbol{X}^T v_{\boldsymbol{w}_{0\gamma}} = \frac{\gamma - b_{0\gamma}}{\|\boldsymbol{w}_{0\gamma}\|}\Big)$ has rank $(d-1)$. Therefore, the right hand side of (2.34) is strictly positive definite when $\gamma \neq b_{0\gamma}$. Similarly, $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y, \tilde{\boldsymbol{X}}^T\theta_{0\gamma} = -\gamma\Big]$ is strictly positive definite when $\gamma \neq -b_{0\gamma}$. Therefore, either $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y = 1, \boldsymbol{X}^T\boldsymbol{w}_{0\gamma} + b_{0\gamma} = \gamma\Big]$ or $E\Big[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T|Y = -1, \boldsymbol{X}^T\boldsymbol{w}_{0\gamma} + b_{0\gamma} = -\gamma\Big]$ will be strictly positive definite at $\boldsymbol{\theta}_{0\gamma}$. This leads to the positive definiteness of $R_1(\boldsymbol{\theta}_{0\gamma})$.

In addition, similar argument implies that $R_2(\boldsymbol{\theta}_{0\gamma})$ is positive definite at $\boldsymbol{\theta}_{0\gamma}$. This is due to the fact that $(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} + 2\gamma - 1)^3 < 0$ when $\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} + \gamma \leq 0$, and $(\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} - 2\gamma + 1)^3 > 0$ when $\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} - \gamma \geq 0$. Therefore, the Hessian matrix $H(\boldsymbol{\theta}_{0\gamma})$ is strictly positive definite for any $\gamma \in [0, 1]$. This concludes Corollary 1. ■

2.7.6 Proof of Corollary 2

Following the proof of Theorem 2.2.1, we only need to show that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) = -n^{-1/2}H(\boldsymbol{\theta}_{0\gamma})^{-1}\sum_{i=1}^{n} M_i(\boldsymbol{\theta}_{0\gamma}) + o_P(1),$$

where

$$M_i(\boldsymbol{\theta}_{0\gamma}) = -Y_i\tilde{\boldsymbol{X}}_i I_{\{Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<\gamma\}} - \frac{(1-\gamma)^2 Y_i \tilde{\boldsymbol{X}}_i I_{\{Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}}{\Big(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1\Big)^2}.$$

Similarly, we denote $Z = (\boldsymbol{X}^T, Y)$ and $t = (b_t, \boldsymbol{w}_t^T)^T$, and write the difference of the loss function according to their definitions,

$$\begin{aligned}
&L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma} + t)) - L_\gamma(Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma}))\\
&= (1 - Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma} + t))I_{\{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t)<\gamma\}} + \frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma} + t) - 2\gamma + 1} I_{\{Y\tilde{\boldsymbol{X}}^T(\boldsymbol{\theta}_{0\gamma}+t)\geq\gamma\}}\\
&\qquad - (1 - Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma})I_{\{Y\widetilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}<\gamma\}} - \frac{(1-\gamma)^2}{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma} - 2\gamma + 1} I_{\{Y\tilde{\boldsymbol{X}}^T\boldsymbol{\theta}_{0\gamma}\geq\gamma\}}\\
&= M(\boldsymbol{\theta}_{0\gamma})^T t + R(Z, t).
\end{aligned}$$

Here we reorganize the complicated terms into two parts $M(\boldsymbol{\theta}_{0\gamma})^T t$ which is linear in t and $R(Z, t)$ which contains higher-order functions of t. In particular, we denote

$$\begin{aligned}
M(\boldsymbol{\theta}_{0\gamma}) &= -Y\tilde{\boldsymbol{X}}^T I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})<\gamma\}} - \frac{(1-\gamma)^2 Y\tilde{\boldsymbol{X}}^T}{(Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1)^2} I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}; \\
R(Z, t) &= \Big(1 - Yf(\boldsymbol{X};\boldsymbol{\theta}_{0\gamma} + t)\Big)\Big[I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}+t)<\gamma\}} - I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})<\gamma\}}\Big] \\
&\quad + \frac{(1-\gamma)^2 I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}+t)\geq\gamma\}}}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma} + t) - 2\gamma + 1} \\
&\quad - \left[\frac{(1-\gamma)^2}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1} - \frac{(1-\gamma)^2 Yf(\boldsymbol{X}, t)}{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1}\right] I_{\{Yf(\tilde{\boldsymbol{X}}^T;\boldsymbol{\theta}_{0\gamma})\geq\gamma\}}.
\end{aligned}$$

It is easy to check that $E(M(\boldsymbol{\theta}_{0\gamma})) = \nabla_{\boldsymbol{\theta}}\mathcal{R}_\gamma(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{0\gamma}}$,

$$E[R(Z, t)] = \frac{1}{2} t^T H(\boldsymbol{\theta}_{0\gamma}) t + o(\|t\|^2) \text{ and } E[R^2(Z, t)] = O(\|t\|^3).$$

The remaining arguments follow exactly from the proof of Theorem 2.2.1. ■

2.7.7 Proof of Lemma 1

In the proof of Corollary 2, we showed that for any $\gamma \in [0, 1]$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma}) = -n^{-1/2} H(\boldsymbol{\theta}_{0\gamma})^{-1} \sum_{i=1}^n M_i(\boldsymbol{\theta}_{0\gamma}) + o_P(1); \tag{2.35}$$

$$\sqrt{n}(\widehat{\mathcal{D}}_\gamma - D_{0\gamma}) = n^{-1/2} \sum_{i=1}^n \psi_{i\gamma} + o_P(1), \tag{2.36}$$

where $\psi_{i\gamma} = \frac{1}{2}|Y_i - \text{sign}\{f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\}| - D_{0\gamma} - \dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1} M_i(\boldsymbol{\theta}_{0\gamma})$. In addition, (2.35) and (2.36) converge to normal distributions.

Next, we show that the right hand sides of (2.35) and (2.36) are uniformly bounded over $\gamma \in [0, 1]$. Denoting the L_1 norm as $\|\cdot\|_1$, we have

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} \Big\| M_i(\boldsymbol{\theta}_{0\gamma}) \Big\|_1 \\
\leq\ & \sup_{\gamma\in[0,1]} \Big\| - Y_i \tilde{\boldsymbol{X}}_i I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<\gamma)} \Big\|_1 + \sup_{\gamma\in[0,1]} \left\| \frac{(1-\gamma)^2 Y_i \tilde{\boldsymbol{X}}_i I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})\geq\gamma)}}{\Big(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma}) - 2\gamma + 1\Big)^2} \right\|_1 \\
\leq\ & 2 \Big\| \tilde{\boldsymbol{X}}_i \Big\|_1.
\end{aligned} \tag{2.37}$$

In addition, $\lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma})) \leq c_2$ in Assumption (B1) implies that each component of the Hessian matrix is uniformly bounded since $\|H(\boldsymbol{\theta}_{0\gamma})\|_{\max} \leq \|H(\boldsymbol{\theta}_{0\gamma})\|_2 = \lambda_{\max}(H(\boldsymbol{\theta}_{0\gamma}))$. This combining with (2.37) and Central Limit Theorem leads to

$$\sup_{\gamma\in[0,1]} \left\|\sqrt{n}(\widehat{\boldsymbol{\theta}}_\gamma - \boldsymbol{\theta}_{0\gamma})\right\|_1 = O_P(1). \tag{2.38}$$

Similarly,

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} \left|\psi_{i\gamma}\right| \\
\leq\ & \sup_{\gamma\in[0,1]} \frac{1}{2}|Y_i - \text{sign}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{\theta}_{0\gamma})| + \sup_{\gamma\in[0,1]} |D_{0\gamma}| + \sup_{\gamma\in[0,1]} \left|\dot{d}(\boldsymbol{\theta}_{0\gamma})^T H(\boldsymbol{\theta}_{0\gamma})^{-1} M_i(\boldsymbol{\theta}_{0\gamma})\right| \\
\leq\ & 1 + 1 + \sup_{\gamma\in[0,1]} \left\|\dot{d}(\boldsymbol{\theta}_{0\gamma})\right\|_1 \sup_{\gamma\in[0,1]} \left\|H(\boldsymbol{\theta}_{0\gamma})^{-1}\right\|_{\max} \sup_{\gamma\in[0,1]} \left\|M_i(\boldsymbol{\theta}_{0\gamma})\right\|_1 \\
\leq\ & 2 + c_3\|\tilde{\boldsymbol{X}}_i\|_1, \qquad (2.39)
\end{aligned}$$

where c_3 in (2.39) is a constant according to $\|H(\boldsymbol{\theta}_{0\gamma})^{-1}\|_{\max} \leq \|H(\boldsymbol{\theta}_{0\gamma})^{-1}\|_2 = 1/\lambda_{\min}(H(\boldsymbol{\theta}_{0\gamma})) \leq 1/c_1$ from Assumption (B1), and

$$\|\dot{d}(\boldsymbol{\theta}_{0\gamma})\|_1 \leq 4\left\|\nabla E\Big(I_{(Y_i f(\boldsymbol{X}_i;\boldsymbol{\theta}_{0\gamma})<0)}\Big)\right\|_1 \leq 4\delta(-Y_i \boldsymbol{\theta}_{0\gamma}^T \tilde{\boldsymbol{X}}_i)\|\tilde{\boldsymbol{X}}_i\|_1 = 0 \ \text{ a.s.}$$

with $\delta(z) = 0$ for $z \neq 0$ and ∞ at $z = 0$. So (2.39) leads to

$$\sup_{\gamma\in[0,1]} \sqrt{n}\left|\widehat{\mathcal{D}}_\gamma - D_{0\gamma}\right| = O_P(1). \tag{2.40}$$

In the end, the definitions of γ_0^* and $\widehat{\gamma}_0^*$ imply that

$$D_{0\gamma_0^*} - D_{0\widehat{\gamma}_0^*} \leq 0 \ \ \textit{and} \ \ \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\gamma_0^*} \leq 0. \tag{2.41}$$

Therefore, we have $D_{0\gamma_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} = D_{0\gamma_0^*} - D_{0\widehat{\gamma}_0^*} + D_{0\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} \leq D_{0\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} = O_P(n^{-1/2})$ based on (2.40) and (2.41). Using similar arguments, we have $\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma_0^*} \leq O_P(n^{-1/2})$. The above discussions imply that $\left|\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma_0^*}\right| = O_P(n^{-1/2})$. This concludes the proof of Lemma 1. ■

2.7.8 Proof of Theorem 2.4.1

Before we prove Theorem 2.4.1, we introduce two useful lemmas.

Lemma 3 *The generalization error* $D_{0\gamma} = \frac{1}{2}E|Y_0 - sign\{\tilde{\boldsymbol{X}}_0^T \widehat{\boldsymbol{\theta}}_\gamma\}|$ *is continuous w.r.t.* γ *a.s.*

Proof of Lemma 3: The discontinuity of sign function happens only at $\tilde{\boldsymbol{X}}_0^T \widehat{\boldsymbol{\theta}}_\gamma = 0$, which is assumed to have probability zero. Hence, it is sufficient to show $\widehat{\boldsymbol{\theta}}_\gamma$ is continuous in γ by dominated convergence theorem. Recall that $\widehat{\boldsymbol{\theta}}_\gamma = \arg\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma}(\boldsymbol{\theta})$ with

$$O_{n\gamma}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L_\gamma\Big(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\Big) + \frac{\lambda_n \boldsymbol{w}^T\boldsymbol{w}}{2}.$$

Note that $O_{n\gamma}(\boldsymbol{\theta})$ is continuous w.r.t. γ due to the continuity of $L_\gamma(u)$ w.r.t. γ. Then, for any sequence $\gamma_n \to \gamma_{00}$ with $\gamma_{00} \in [0,1]$, continuous mapping theorem implies that $|O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta})| < \delta$ for any $\delta > 0$ when n is sufficiently large. Denote $\widehat{\boldsymbol{\theta}}_{\gamma_{00}} = \arg\min_{\boldsymbol{\theta}} O_{n\gamma_{00}}(\boldsymbol{\theta})$ and $\mathcal{G} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\gamma_{00}}\| \le \epsilon\}$. For each fixed ϵ, we construct

$$\delta = \frac{\min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}})}{2}.$$

Then we have

$$\begin{aligned} O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) &= \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) - 2\delta \\ &< \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_{00}}(\boldsymbol{\theta}) + O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta}) - \delta \\ &\le O_{n\gamma_n}(\boldsymbol{\theta}) - \delta, \end{aligned}$$

which is true for any $\boldsymbol{\theta} \in R^{d+1}$. Therefore,

$$O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) < \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_n}(\boldsymbol{\theta}) - \delta. \tag{2.42}$$

On the other hand, $|O_{n\gamma_n}(\boldsymbol{\theta}) - O_{n\gamma_{00}}(\boldsymbol{\theta})| < \delta$ implies that $O_{n\gamma_n}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) - O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) < \delta$ and hence $\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) < O_{n\gamma_{00}}(\widehat{\boldsymbol{\theta}}_{\gamma_{00}}) + \delta$. This combining with (2.42) leads to

$$\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) < \min_{\boldsymbol{\theta}\in R^{d+1}\setminus\mathcal{G}} O_{n\gamma_n}(\boldsymbol{\theta}).$$

Therefore, $\arg\min_{\boldsymbol{\theta}\in R^{d+1}} O_{n\gamma_n}(\boldsymbol{\theta}) \in \mathcal{G}$, and hence $\widehat{\boldsymbol{\theta}}_\gamma$ is continuous at γ_{00}. Note that ϵ can be made arbitrarily small and γ_{00} is an arbitrary element within $[0,1]$. This concludes Lemma 3. ■

The following Lemma 4 shows the (element-wise) asymptotic equivalence between Λ_0 and $\widehat{\Lambda}_0$.

Lemma 4 *Suppose that the assumptions in Lemma 1 hold. We have, as $n \to \infty$, (i) for any $\widehat{\gamma} \in \widehat{\Lambda}_0$, there exists a $\gamma \in \Lambda_0$ such that $\widehat{\gamma} \xrightarrow{P} \gamma$; (ii) for any $\gamma \in \Lambda_0$, there exists a $\widehat{\gamma} \in \widehat{\Lambda}_0$ satisfying $\widehat{\gamma} \xrightarrow{P} \gamma$.*

Proof of Lemma 4: Our proof consists of two steps. In the first step, for any $\widehat{\gamma} \in \widehat{\Lambda}_0$ with $\widehat{\gamma} \xrightarrow{P} \gamma$, we have

$$\begin{aligned} D_{0\gamma} - D_{0\gamma_0^*} &= (D_{0\gamma} - D_{0\widehat{\gamma}}) + (D_{0\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\gamma_0^*}) + (\widehat{\mathcal{D}}_{\gamma_0^*} - D_{0\gamma_0^*}) \\ &= I + II + III + IV. \end{aligned}$$

Obviously, we have $I = o_P(1)$ according to continuous mapping theorem and Lemma 3, and $II, IV = o_P(1)$ due to (2.40). As for III, we have $III \leq \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - \widehat{\mathcal{D}}_{\gamma_0^*} + n^{-1/2}\phi_{\widehat{\gamma},\widehat{\gamma}_0^*;\alpha/2} \leq o_P(1)$ since $\widehat{\gamma} \in \widehat{\Lambda}_0$ defined in (2.16). The above discussions lead to the conclusion that $D_{0\gamma} - D_{0\gamma_0^*} \leq o_P(1)$. Therefore, we have $P(\gamma \in \Lambda_0) \geq P(D_{0\gamma} - D_{0\gamma_0^*} \leq 0) \to 1$.

In the second step, we apply the contradiction argument. Assume there exists some $\gamma \in \Lambda_0$ such that $\widehat{\gamma} \notin \widehat{\Lambda}_0$ for any $\widehat{\gamma} \xrightarrow{P} \gamma$. The above assumption directly implies that $\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} > o_P(1)$. The analysis in the first step further implies that there exists some $\gamma^* \in \Lambda_0$, i.e., $D_{0\gamma^*} = D_{0\gamma_0^*}$, with probability tending to one such that $\widehat{\gamma}_0^* \xrightarrow{P} \gamma^*$. Then, we have

$$\begin{aligned} D_{0\gamma} - D_{0\gamma^*} &= (D_{0\gamma} - D_{0\widehat{\gamma}}) + (D_{0\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}} - \widehat{\mathcal{D}}_{\widehat{\gamma}_0^*}) + (\widehat{\mathcal{D}}_{\widehat{\gamma}_0^*} - D_{0\gamma^*}) \\ &= I + II + III' + IV'. \end{aligned}$$

Recall that $I, II = o_P(1)$ and $III' > o_P(1)$ as shown in the above. We also have $IV' = o_P(1)$ due to (2.40) and the fact that $\widehat{\gamma}_0^* \xrightarrow{P} \gamma^*$. In summary, we have $D_{0\gamma} - D_{0\gamma^*} > o_P(1)$, which contradicts the definition of γ. This concludes the proof of Lemma 4. ■

Proof of Theorem 2.4.1: The proof consists of two major steps. In the first step, we show that

$$\sup_{\gamma\in[0,1]} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| \to 0. \tag{2.43}$$

Denote $\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) = \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T} Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger$, where $\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger = (1,(R_\gamma\boldsymbol{x}_i)_{(-d)}^T)^T$ and R_γ is the transformation matrix associated with the loss function L_γ. Then we have

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| \\
\le &\sup_{\gamma\in[0,1]} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - \overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| \\
&+ \sup_{\gamma\in[0,1]} n\Big|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big|.
\end{aligned} \tag{2.44}$$

Next we show each summand in (2.44) converges to 0. For the first one, we have

$$\begin{aligned}
&\sup_{\gamma\in[0,1]} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - \overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| \\
= &\sup_{\gamma\in[0,1]} n\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T}\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger - \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T} Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger\Big| \\
= &\sup_{\gamma\in[0,1]} \Big|\sum_{i=1}^n \widetilde{\boldsymbol{x}}_{i(-d)}^{\dagger T}[(\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) - Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger))]\widetilde{\boldsymbol{x}}_{i(-d)}^\dagger\Big|,
\end{aligned} \tag{2.45}$$

where

$$Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) = \frac{\Sigma_{0\gamma,(-d)}^\dagger}{n(w_{\gamma,d}^\dagger)^2} \text{ and } \widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) = \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger}{n(\widehat{w}_{\gamma,d}^\dagger)^2}.$$

Here, $\widehat{w}_{\gamma,d}^\dagger$ is the last dimension of $\widehat{\boldsymbol{\theta}}_\gamma^*$. Since $\widehat{w}_{\gamma,d}^\dagger$ follows the normal distribution with mean $w_{\gamma,d}^\dagger$ and variance converging to 0, we have $\widehat{w}_{\gamma,d}^\dagger = w_{\gamma,d}^\dagger + o_P(1)$, and hence $(\widehat{w}_{\gamma,d}^\dagger)^2 = (w_{\gamma,d}^\dagger)^2 + o_P(1)$ due to the boundedness of $w_{\gamma,d}^\dagger$. In addition, uniform law of large numbers implies that each component of $\widehat{\Sigma}_\gamma^\dagger - \Sigma_{0\gamma}^\dagger$ uniformly converges to 0 w.r.t. γ, because each element of $\widehat{\Sigma}_\gamma^\dagger$ is continuous w.r.t. γ (by similar arguments as in Lemma 3). Therefore, we have

$$\begin{aligned}
n\Big[\widehat{Var}(\widehat{\boldsymbol{\eta}}_\gamma^\dagger) - Var(\widehat{\boldsymbol{\eta}}_\gamma^\dagger)\Big] &= \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger}{(\widehat{w}_{\gamma,d}^\dagger)^2} - \frac{\Sigma_{0\gamma,(-d)}^\dagger}{(w_{\gamma,d}^\dagger)^2} \\
&= \frac{\widehat{\Sigma}_{\gamma,(-d)}^\dagger - \Sigma_{0\gamma,(-d)}^\dagger}{(w_{\gamma,d}^\dagger)^2 + o_P(1)} - \frac{\Sigma_{0\gamma,(-d)}^\dagger o_P(1)}{(w_{\gamma,d}^\dagger)^2[(w_{\gamma,d}^\dagger)^2 + o_P(1)]},
\end{aligned} \tag{2.46}$$

where the second term in (2.46) uniformly converges to 0 due to Assumption (B1) and the boundedness of $w^{\dagger}_{\gamma,d}$. Therefore, each element of (2.46) uniformly converges to 0, which implies that (2.45) converges to 0.

As for the second summand of (2.44), we again apply uniform law of large numbers to show

$$\sup_{\gamma\in[0,1]} n\Big|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma}))\Big| \to 0.$$

Note that $\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)} Var(\widehat{\boldsymbol{\eta}}_{\gamma})^{\dagger}\tilde{\boldsymbol{X}}^{\dagger}_{(-d)}$ is continuous w.r.t. γ by similar arguments as in Lemma 3, and

$$\begin{aligned} n\Big|\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)} Var(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})\tilde{\boldsymbol{X}}^{\dagger}_{(-d)}\Big| &= \Big|(1,(R_{\gamma}\boldsymbol{x})^{T}_{(-d)})^{T} nVar(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})(1,(R_{\gamma}\boldsymbol{x})^{T}_{(-d)})\Big| \\ &\le c_4\Big|1+\boldsymbol{x}^{T}_{(-d)}\boldsymbol{x}_{(-d)}\Big| \le c_5, \end{aligned}$$

where the first inequality holds because each component of $nVar(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})$ is uniformly bounded due to the boundedness of $w^{\dagger}_{\gamma,d}$ and Assumption (B1). Then the uniform law of large number implies

$$\begin{aligned} &\sup_{\gamma\in[0,1]} n\Big|\overline{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma}))\Big| \\ = &\sup_{\gamma\in[0,1]}\Big|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}^{\dagger T}_{i(-d)}(w^{\dagger}_{\gamma,d})^{-2}\Sigma^{\dagger}_{0\gamma,(-d)}(\widehat{\boldsymbol{\eta}}^{\dagger}_{\gamma})\widetilde{\boldsymbol{x}}^{\dagger}_{i(-d)} - E\Big(\tilde{\boldsymbol{X}}^{\dagger T}_{(-d)}(w^{\dagger}_{\gamma,d})^{-2}\Sigma^{\dagger}_{0\gamma,(-d)}\tilde{\boldsymbol{X}}^{\dagger}_{(-d)}\Big)\Big| \\ \to &\ 0. \end{aligned} \tag{2.47}$$

Combining (2.45) and (2.47) leads to (2.43).

In the second step of the proof, we show $n(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))-DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))) \le o_P(1)$ and $n(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))) \le o_P(1)$, from which the desirable result (2.21) follows.

Firstly, we prove

$$n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big) \le o_P(1).$$

Denote $\widehat{\gamma}^{\sharp}_0 = \arg\min_{\gamma\in\widehat{\Lambda}_0} DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma}))$. For γ_0 defined in (2.20), Theorem 4 implies that there exists a $\widehat{\gamma}^{\triangle}_0 \in \widehat{\Lambda}_0$ such that $\widehat{\gamma}^{\triangle}_0 \xrightarrow{P} \gamma_0$, then we have

$$n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big) = I + II + III,$$

where

$$\begin{aligned} I &= n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))\Big), \\ II &= n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle}))\Big), \\ III &= n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0}))\Big). \end{aligned}$$

Note that $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) \leq \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))$ according to (2.19) and hence

$$I \leq n\Big(\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp}))\Big).$$

According to $DBI(S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\sharp})) \leq DBI(S(\boldsymbol{x};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle}))$ due to $\widehat{\gamma}_0^\sharp \in \widehat{\Lambda}_0$, we have $II \leq 0$. Moreover, $DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0^\triangle})) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) = o_P(n^{-1})$ according to $\widehat{\gamma}_0^\triangle \xrightarrow{P} \gamma_0$ and continuous mapping theorem. All these together with (2.43) lead to

$$I + II + III \leq \sup_{\gamma\in\widehat{\Lambda}_0} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| + o_P(1) \leq o_P(1). \quad (2.48)$$

Secondly, we prove

$$n(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))) \leq o_P(1).$$

Denote $\widetilde{\gamma}_0 = \arg\min_{\gamma\in\Lambda_0} \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))$. For $\widehat{\gamma}_0$ defined in (2.19), Lemma 4 implies that there exists $\widetilde{\gamma}_0^\sharp \in \Lambda_0$ such that $\widehat{\gamma}_0 \xrightarrow{P} \widetilde{\gamma}_0^\sharp$, then we have

$$\begin{aligned} & DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) \leq DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) \\ + \; & \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp})) + \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})). \end{aligned}$$

Here $DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) \leq DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0}))$ by the definition of γ_0, $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0})) \leq \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp}))$ due to the definition of $\widetilde{\gamma}_0$, and $\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widetilde{\gamma}_0^\sharp})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0})) = o_P(n^{-1})$ according to $\widehat{\gamma}_0 \xrightarrow{P} \widetilde{\gamma}_0^\sharp$ and continuous mapping theorem. Therefore,

$$\begin{aligned} & n\Big(DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\gamma_0})) - \widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_{\widehat{\gamma}_0}))\Big) \\ \leq \; & \sup_{\gamma\in\Lambda_0} n\Big|\widehat{DBI}(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma)) - DBI(S(\boldsymbol{X};\widehat{\boldsymbol{\theta}}_\gamma))\Big| + o_P(1) \leq o_P(1), \end{aligned} \quad (2.49)$$

Consequently, combining (2.48) and (2.49) leads to the desirable conclusion in Theorem 2.4.1. ■

3. STABILIZED NEAREST NEIGHBOR CLASSIFIER AND ITS THEORETICAL PROPERTIES

The k-nearest neighbor (kNN) classifier [50, 51] is one of the most popular nonparametric classification methods, due to its conceptual simplicity and powerful prediction capability. In the literature, extensive research have been done to justify various nearest neighbor classifiers based on the risk, which calibrates the inaccuracy of the classifier [52–57]. We refer the readers to [58] for a comprehensive study. Recently, [27] has proposed an optimal weighted nearest neighbor (OWNN) classifier. Like most other existing nearest neighbor classifiers, OWNN focuses on the risk without paying much attention to the classification stability.

In this chapter, we define a general measure of stability for a classification method. It characterizes the sampling variability of the prediction, and is named the classification instability (CIS). CIS applies to both linear and non-linear classifiers. An important result we obtain is that the CIS of a weighted nearest neighbor (WNN) classifier is asymptotically proportional to the ℓ_2 norm of the weight vector. This rather concise form is crucial in our methodological development and theoretical analysis. To illustrate the interplay between risk and CIS, we apply the kNN classifier to a bivariate toy example (see details in Section 3.6.1) and plot in Figure 3.1 the regret (that is the risk minus a constant known as the Bayes risk) versus CIS, calculated according to Proposition 2 and Theorem 3.2.1 in Section 3.2, for different k. As k increases, the classifier becomes more and more stable, while the regret first decreases and then increases. In view of the kNN classifier with the minimal regret, marked as the red square in Figure 3.1, one may have the impression that there are other k values with similar regret but much smaller CIS, such as the one marked as the blue triangle shown in the plot.

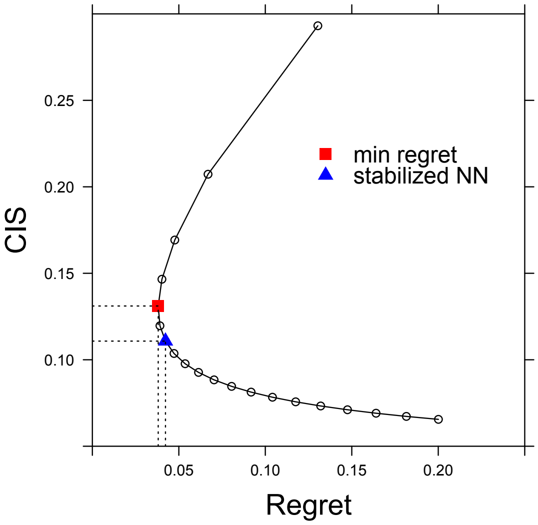

Figure 3.1. Regret and CIS of the kNN classifier. From top to bottom, each circle represents the kNN classifier with $k \in \{1, 2, \ldots, 20\}$. The red square corresponds to the classifier with the minimal regret and the classifier depicted by the blue triangle improves it to have a lower CIS.

Inspired by Figure 3.1, we propose a novel method called stabilized nearest neighbor (SNN) classifier, which takes the stability into consideration. We construct the SNN procedure by minimizing the CIS of WNN over an acceptable region where the regret is small, indexed by a tuning parameter. SNN encompasses the OWNN classifier as a special case.

To understand the theoretical property of SNN, we establish a sharp convergence rate of CIS for general plug-in classifiers. This sharp rate is slower than but approaching n^{-1}, shown by adapting the framework of [21]. Furthermore, the proposed SNN method is shown to achieve both the minimax optimal rate in the regret established in the literature, and this sharp rate in CIS established in this article.

In order to illustrate the advantage of the SNN classifier, we offer a comprehensive asymptotic comparison among various classifiers, through which new insights are obtained. It is theoretically verified that the CIS of our SNN procedure is much smaller than those of others. Figure 3.2 shows the regret and CIS of kNN, OWNN,

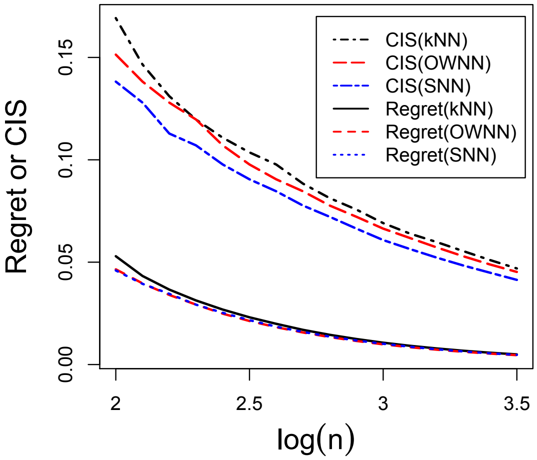

Figure 3.2. Regret and CIS of kNN, OWNN, and SNN procedures for a bivariate normal example. The top three lines represent CIS's of kNN, OWNN, and SNN. The bottom three lines represent regrets of kNN, SNN, and OWNN. The sample size shown on the x-axis is in the $\log_{10}$ scale.

and SNN for a bivariate example (see details in Section 3.6.1). Although OWNN is *theoretically* the best in regret, its regret curve appear to overlap with that of SNN. On the other hand, the SNN procedure has a visibly smaller CIS than OWNN. A compelling message is that with almost the same accuracy, our SNN could greatly improve stability. Extensive experiments further illustrate that SNN has a significant improvement in CIS, and sometimes even improves the accuracy. Such appealing results are supported by the theoretical finding (in Corollary 3) that the regret of SNN approaches that of OWNN at a faster rate than the rate at which the CIS of OWNN approaches that of SNN, where both rates are shown to be sharp. As a by-product, we also show that OWNN is more stable than the kNN and the bagged nearest neighbor (BNN) classifiers.

3.1 Classification Instability

Let $(X, Y) \in \mathbb{R}^d \times \{1, -1\}$ be a random couple with joint distribution P. We regard X as a d-dimensional vector of features for an object and Y as the label indicating that the object belongs to one of two classes. Denote the prior class probability as $\pi_1 = \mathbb{P}(Y = 1)$, where $\mathbb{P}$ is the probability with respect to P, and the distribution of X given $Y = 1$ as P_1. Similarly, we denote the distribution of X given $Y = -1$ as P_2. The marginal distribution of X can be written as $\bar{P} = \pi_1 P_1 + (1 - \pi_1)P_2$. For a classifier $\phi : \mathbb{R}^d \mapsto \{1, -1\}$, the risk of ϕ is defined as $R(\phi) = \mathbb{P}(\phi(X) \neq Y)$. It is well known that the Bayes rule, denoted as ϕ^{Bayes}, minimizes the above risk. Specifically, $\phi^{\text{Bayes}}(x) = 1 - 2\mathbb{1}\{\eta(x) < 1/2\}$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$ and $\mathbb{1}\{\cdot\}$ is the indicator function. In practice, a classification procedure Ψ is applied to a training data set $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ to produce a classifier $\widehat{\phi}_n = \Psi(\mathcal{D})$. We define the risk of the procedure Ψ as $\mathbb{E}_{\mathcal{D}}[R(\widehat{\phi}_n)]$, and the regret of Ψ as $\mathbb{E}_{\mathcal{D}}[R(\widehat{\phi}_n)] - R(\phi^{\text{Bayes}})$, where $\mathbb{E}_{\mathcal{D}}$ denotes the expectation with respect to the distribution of $\mathcal{D}$, and $R(\phi^{\text{Bayes}})$ is the risk of the Bayes rule, called the Bayes risk. The risk describes the inaccuracy of a classification method. In practice, for a classifier ϕ, the classification error for a test data can be calculated as an emprical version of the risk $R(\phi)$.

For a classification procedure, it is desired that, with high probability, classifiers trained from different samples yield the same prediction for the same object. Our first step in formalizing the classification instability is to define the distance between two generic classifiers ϕ_1 and ϕ_2, which measures the level of disagreement between them.

Definition 2 *(Distance of Classifiers) Define the distance between two classifiers ϕ_1 and ϕ_2 as $d(\phi_1, \phi_2) = \mathbb{P}(\phi_1(X) \neq \phi_2(X))$.*

We next define the classification instability (CIS). Throughout the article, we denote $\mathcal{D}_1$ and $\mathcal{D}_2$ as two i.i.d. copies of the training sample $\mathcal{D}$. For ease of notation, we have suppressed the dependence of $\text{CIS}(\Psi)$ on the sample size n of $\mathcal{D}$.

Definition 3 *(Classification Instability) Define the classification instability of a classification procedure* Ψ *as*

$$CIS(\Psi) = \mathbb{E}_{\mathcal{D}_1,\mathcal{D}_2}\Big[d(\widehat{\phi}_{n1}, \widehat{\phi}_{n2})\Big] \tag{3.1}$$

where $\widehat{\phi}_{n1} = \Psi(\mathcal{D}_1)$ *and* $\widehat{\phi}_{n2} = \Psi(\mathcal{D}_2)$ *are the classifiers obtained by applying the classification procedure* Ψ *to samples* $\mathcal{D}_1$ *and* $\mathcal{D}_2$*.*

Intuitively, CIS is the average probability that the same object is classified to two different classes in two separate runs of the learning algorithm. By definition, $0 \leq \text{CIS}(\Psi) \leq 1$, and a small $\text{CIS}(\Psi)$ represents a stable classification procedure Ψ.

3.2 Stabilized Nearest Neighbor Classifier

In this section, we introduce a novel classification method called the *stabilized* nearest neighbor (SNN) procedure which well balances the trade-off between classification accuracy and classification stability. Section 3.2.1 reviews the general weighted nearest neighbor (WNN) framework. Section 3.2.2 derives an asymptotic equivalent form of the CIS measure for the WNN procedure, which turns out to be proportional to the Euclidean norm of the weight vector. Based on WNN's explicit expressions of the regret and CIS, we propose the SNN classification procedure in Section 3.2.3.

3.2.1 Review of WNN

For any fixed x, let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ be a sequence of observations with ascending distance to x. For a nonnegative weight vector $\boldsymbol{w}_n = (w_{ni})_{i=1}^n$ satisfying $\sum_{i=1}^n w_{ni} = 1$, the WNN classifier $\widehat{\phi}_n^{\boldsymbol{w}_n}$ predicts the label of x as $\widehat{\phi}_n^{\boldsymbol{w}_n}(x) = 1 - 2\mathbb{1}\{\sum_{i=1}^n w_{ni}\mathbb{1}\{Y_{(i)} = 1\} < 1/2\}$.

[27] revealed a nice asymptotic expansion formula of the regret of WNN under the following Assumptions. For a smooth function g, we denote $\dot{g}(x)$ as its gradient vector at x. We assume the following conditions through all the article.

(A1) The set $\mathcal{R} \subset \mathbb{R}^d$ is a compact d-dimensional manifold with boundary $\partial\mathcal{R}$.

(A2) The set $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$ is nonempty. There exists an open subset U_0 of $\mathbb{R}^d$ which contains $\mathcal{S}$ such that: (i) η is continuous on $U \backslash U_0$ with U an open set containing $\mathcal{R}$; (ii) the restriction of the conditional distributions of X, P_1 and P_2, to U_0 are absolutely continuous with respect to Lebesgue measure, with twice continuously differentiable Randon-Nikodym derivatives f_1 and f_2.

(A3) There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^\rho d\bar{P}(x) < \infty$. Moreover, for sufficiently small $\delta > 0$, $\inf_{x\in\mathcal{R}} \bar{P}(B_\delta(x))/(a_d \delta^d) \geq C_3 > 0$, where $a_d = \pi^{d/2}/\Gamma(1 + d/2)$, $\Gamma(\cdot)$ is gamma function, and C_3 is a constant independent of δ.

(A4) For all $x \in \mathcal{S}$, we have $\dot{\eta}(x) \neq 0$, and for all $x \in \mathcal{S} \cap \partial\mathcal{R}$, we have $\dot{\partial\eta}(x) \neq 0$, where $\partial\eta$ is the restriction of η to $\partial\mathcal{R}$. ■

Remark 3 *Assumptions (A1)–(A4) have also been employed to show the asymptotic expansion of the regret of the kNN classifier [59]. The condition $\dot{\eta}(x) \neq 0$ in (A4) is equivalent to the margin condition with $\alpha = 1$; see (2.1) in [27]. These assumptions ensure that $\bar{f}(x_0)$ and $\dot{\eta}(x_0)$ are bounded away from zero and infinity on $\mathcal{S}$.*

Moreover, for a smooth function g: $\mathbb{R}^d \to \mathbb{R}$, let $g_j(x)$ its jth partial derivative at x, $\ddot{g}(x)$ the Hessian matrix at x, and $g_{jk}(x)$ the (j, k)th element of $\ddot{g}(x)$. Let $c_{j,d} = \int_{v:\|v\|\leq 1} v_j^2 dv$. Define

$$a(x) = \sum_{j=1}^{d} \frac{c_{j,d}\{\eta_j(x)\bar{f}_j(x) + 1/2\eta_{jj}(x)\bar{f}(x)\}}{a_d^{1+2/d}\bar{f}(x)^{1+2/d}}.$$

We further define two distribution-related constants

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x)}{4\|\dot{\eta}(x)\|} d\mathrm{Vol}^{d-1}(x), \quad B_2 = \int_{\mathcal{S}} \frac{\bar{f}(x)}{\|\dot{\eta}(x)\|} a(x)^2 d\mathrm{Vol}^{d-1}(x),$$

where Vol^{d-1} is the natural $(d-1)$-dimensional volume measure that $\mathcal{S}$ inherits. Based on Assumptions (A1)-(A4), B_1 and B_2 are finite with $B_1 > 0$ and $B_2 \geq 0$, where $B_2 = 0$ only when $a(x)$ equals zero on $\mathcal{S}$. In addition, for $\beta > 0$, we denote $W_{n,\beta}$ as the set of $\boldsymbol{w}_n$ satisfying (w.1)–(w.5).

(w.1) $\sum_{i=1}^n w_{ni}^2 \leq n^{-\beta}$,

(w.2) $n^{-4/d}(\sum_{i=1}^n \alpha_i w_{ni})^2 \leq n^{-\beta}$, where $\alpha_i = i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}}$, (w.3) $n^{2/d}\sum_{i=k_2+1}^n w_{ni}/\sum_{i=1}^n \alpha_i w_{ni} \leq 1/\log n$ with $k_2 = \lfloor n^{1-\beta} \rfloor$,

(w.4) $\sum_{i=k_2+1}^n w_{ni}^2/\sum_{i=1}^n w_{ni}^2 \leq 1/\log n$,

(w.5) $\sum_{i=1}^n w_{ni}^3/(\sum_{i=1}^n w_{ni}^2)^{3/2} \leq 1/\log n$.

For the kNN classifier with $w_{ni} = k^{-1}\mathbb{1}\{1 \leq i \leq k\}$, [27] showed that (w.1)–(w.5) reduce to $\max(n^{\beta}, (\log n)^2) \leq k \leq \min(n^{(1-\beta d/4)}, n^{1-\beta})$. More discussions on $W_{n,\beta}$ can be found in [27].

Proposition 2 *[27] Under Assumptions (A1)–(A4), for each $\beta \in (0, 1/2)$, we have, as $n \to \infty$,*

$$\textit{Regret}(\textit{WNN}) = \left\{ B_1 \sum_{i=1}^n w_{ni}^2 + B_2 \Big(\sum_{i=1}^n \frac{\alpha_i w_{ni}}{n^{2/d}} \Big)^2 \right\} \{1 + o(1)\}, \tag{3.2}$$

uniformly for $\boldsymbol{w}_n \in W_{n,\beta}$ with $W_{n,\beta}$, where $\alpha_i = i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}}$.

[27] further derived a weight vector that minimizes the asymptotic regret (3.2) which led to the optimal weighted nearest neighbor (OWNN) classifier. As will be shown in Section 3.2.3, the OWNN classifier is a special case of our SNN classifier.

3.2.2 Asymptotically Equivalent Formulation of CIS

Denote the two resulting WNN classifiers trained on $\mathcal{D}_1$ and $\mathcal{D}_2$ as $\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(x)$ and $\widehat{\phi}_{n2}^{\boldsymbol{w}_n}(x)$ respectively. With a slight abuse of notation, we denote the CIS of a WNN classification procedure by CIS(WNN). According to the definition in (3.1), classification instability of a WNN procedure is $\text{CIS(WNN)} = \mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2,X}\Big(\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(X) \neq \widehat{\phi}_{n2}^{\boldsymbol{w}_n}(X)\Big)$.

We first do some initial analysis of the above CIS measure, which helps derive its asymptotic equivalent formulation. Based on the fact $\mathcal{D}_1, \mathcal{D}_2 \overset{i.i.d.}{\sim} \mathcal{D}$, the CIS of WNN can be reformulated as

$$\begin{aligned} & \text{CIS(WNN)} \\ = \; & \mathbb{E}_X \Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2} \Big(\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(X) \neq \widehat{\phi}_{n2}^{\boldsymbol{w}_n}(X) \Big| X \Big) \Big] \\ = \; & \mathbb{E}_X \Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2} \Big(\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(X) = 1, \widehat{\phi}_{n2}^{\boldsymbol{w}_n}(X) = -1 \Big| X \Big) \Big] \\ & + \mathbb{E}_X \Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2} \Big(\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(X) = -1, \widehat{\phi}_{n2}^{\boldsymbol{w}_n}(X) = 1 \Big| X \Big) \Big] \\ = \; & 2\mathbb{E}_X \Big[\mathbb{P}_{\mathcal{D}} \Big(\widehat{\phi}_{n}^{\boldsymbol{w}_n}(X) = 1 | X \Big) \Big] - 2\mathbb{E}_X \Big[\mathbb{P}_{\mathcal{D}}^2 \Big(\widehat{\phi}_{n}^{\boldsymbol{w}_n}(X) = 1 | X \Big) \Big] \end{aligned} \tag{3.3}$$

The above derivations indicate that WNN's CIS can be fully captured by its performance on a single data set, and we can develop the formulation by studying each term in (3.3) separately.

Theorem 3.2.1 provides an asymptotic expansion formula for the CIS of WNN in terms of the weight vector $\boldsymbol{w}_n$.

Theorem 3.2.1 *(Asymptotic Equivalent Form of CIS) Under Assumptions (A1)–(A4), for each* $\beta \in (0, 1/2)$*, we have, as* $n \to \infty$*,*

$$\mathit{CIS}(\mathit{WNN}) = B_3 \Big(\sum_{i=1}^{n} w_{ni}^2 \Big)^{1/2} \{1 + o(1)\}, \tag{3.4}$$

uniformly for all $\boldsymbol{w}_n \in W_{n,\beta}$ *with* $W_{n,\beta}$*, where the constant* $B_3 = 4B_1/\sqrt{\pi} > 0$ *with* B_1 *defined in Proposition 2.*

Theorem 3.2.1 demonstrates that CIS of the WNN procedure is asymptotically proportional to $(\sum_{i=1}^{n} w_{ni}^2)^{1/2}$. For example, for the kNN procedure (that is the WNN procedure with $w_{ni} = k^{-1}\mathbb{1}\{1 \le i \le k\}$), its CIS is asymptotically $B_3\sqrt{1/k}$. Therefore, a larger value of k leads to a more stable kNN procedure, which was seen in Figure 3.1. Furthermore, we note that the CIS expansion in (3.4) is related to the first term in (3.2). The expansions in (3.2) and (3.4) allow precise calibration of the regret and CIS. This interesting connection is important in the development of our SNN procedure.

3.2.3 Stabilized Nearest Neighbor Classifier

To stabilize WNN, we consider a weight vector which minimizes the CIS over an acceptable region where the classification regret is less than some constant $c_1 > 0$, that is,

$$\begin{aligned} \min_{\boldsymbol{w}_n} \quad & \mathrm{CIS}(\mathrm{WNN}) \\ \text{subject to} \quad & \mathrm{Regret}(\mathrm{WNN}) \leq c_1, \ \sum_{i=1}^{n} w_{ni} = 1, \ \boldsymbol{w}_n \geq 0. \end{aligned} \tag{3.5}$$

By considering $\mathrm{CIS}(\mathrm{WNN})^2$ in the objective function and the Lagrangian formulation, we can see that (3.5) is equivalent to minimizing $\mathrm{Regret}(\mathrm{WNN}) + \lambda_0 \mathrm{CIS}^2(\mathrm{WNN})$ subject to the constraints that $\sum_{i=1}^{n} w_{ni} = 1$ and $\boldsymbol{w}_n \geq 0$, where $\lambda_0 > 0$. The expansions (3.2) and (3.4) in Proposition 2 and Theorem 3.2.1 imply the following asymptotically equivalent formulation:

$$\begin{aligned} \min_{\boldsymbol{w}_n} \quad & \left(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2 \\ \text{subject to} \quad & \sum_{i=1}^{n} w_{ni} = 1, \ \boldsymbol{w}_n \geq 0, \end{aligned} \tag{3.6}$$

where $\lambda = (B_1 + \lambda_0 B_3^2)/B_2$ depends on constants B_1 and B_2 and λ_0. When $\lambda \to \infty$, (3.6) leads to the most stable but trivial kNN classifier with $k = n$. The classifier in (3.6) with $\lambda \downarrow B_1/B_2$ (*i.e.*, $\lambda_0 \downarrow 0$) approaches the OWNN classifier considered in [27]. Note that the two terms $(n^{-2/d} \sum_{i=1}^{n} \alpha_i w_{ni})^2$ and $\sum_{i=1}^{n} w_{ni}^2$ in (3.6) represent the bias and variance terms of the regret expansion given in Proposition 2 [27]. By varying the weights of these two terms through λ, we are able to stablize a nearest neighbor classifier. Moreover, the stabilized classifier achieves desirable convergence rates in both regret and CIS.

Theorem 3.2.2 gives the optimal weight w_{ni}^* with respect to the optimization problem (3.6). We formally define the stabilized nearest neighbor (SNN) classifier as the WNN classifier with the optimal weight w_{ni}^*.

Theorem 3.2.2 *(Optimal Weight) For any fixed $\lambda > 0$, the minimizer of* (3.6) *is*

$$w_{ni}^* = \begin{cases} \frac{1}{k^*}\left[1 + \frac{d}{2} - \frac{d}{2(k^*)^{2/d}}\alpha_i\right], & \text{for } i = 1, \ldots, k^*, \\ 0, & \text{for } i = k^* + 1, \ldots, n, \end{cases}$$

where $\alpha_i = i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}}$ and $k^* = \lfloor \{\frac{d(d+4)}{2(d+2)}\}^{\frac{d}{d+4}} \lambda^{\frac{d}{d+4}} n^{\frac{4}{d+4}} \rfloor$.

The SNN classifier encompasses the OWNN classifier as a special case when $\lambda = B_1/B_2$. Note that Theorem 3.2.2 is valid for any $\lambda > 0$, which includes the case $\lambda > B_1/B_2$.

The computational complexity of our SNN classifier is comparable to that of existing nearest neighbor classifiers. If we preselect a value for λ, SNN requires no training at all. The testing time consists of two parts: the $O(n)$ complexity for the computation of n distances, where n is the size of training data; and the $O(n \log n)$ complexity for sorting n distances. The kNN classifier, for example, shares the same computational complexity. In practice, λ is not predetermined and we may treat it as a tuning parameter, whose optimal value is selected via cross validation. See Algorithm 1 in Section 3.5 for details. We will show in Section 3.5 that the complexity of tuning in SNN is also comparable to existing methods.

3.3 Theoretical Properties

In this section, we establish a sharp convergence rate of CIS for a general plug-in classifier, which includes SNN as a special case. We then demonstrate that SNN attains this established sharp convergence rate in CIS, as well as the minimax optimal convergence rate in regret.

3.3.1 A Sharp Rate of CIS

Motivated by [21], we establish a sharp convergence rate of CIS for a general plug-in classifier. A plug-in classification procedure Ψ first estimates the regression function $\eta(x)$ by $\widehat{\eta}_n(x)$, and then plugs it into the Bayes rule, that is, $\widehat{\phi}_n(x) = 1 - 2\mathbb{1}\{\widehat{\eta}_n(x) < 1/2\}$.

The following *margin condition* [21] is assumed for deriving the upper bound of the convergence rate, while two additional conditions are required for showing the lower bound. A distribution function P satisfies the *margin condition* if there exist constants $C_0 > 0$ and $\alpha \geq 0$ such that for any $\epsilon > 0$,

$$\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \epsilon) \leq C_0\epsilon^{\alpha}. \tag{3.7}$$

The parameter α characterizes the behavior of the regression function η near 1/2, and a larger α implies a lower noise level and hence an easier classification scenario.

The second condition is on the smoothness of $\eta(x)$. Specifically, we assume that η belongs to a *Hölder class of functions* $\Sigma(\gamma, L, \mathbb{R}^d)$ (for some fixed $L, \gamma > 0$) containing the functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor\gamma\rfloor$ times continuously differentiable and satisfy, for any $x, x' \in \mathbb{R}^d$, $|g(x') - g_x(x')| \leq L\|x - x'\|^{\gamma}$, where $\lfloor\gamma\rfloor$ is the largest integer not greater than γ, g_x is the Taylor polynomial series of degree $\lfloor\gamma\rfloor$ at x, and $\|\cdot\|$ is the Euclidean norm.

Our last condition assumes that the marginal distribution $\bar{P}$ satisfies the *strong density assumption* which satisfies that for a compact set $\mathcal{R} \subset \mathbb{R}^d$ and constants $c_0, r_0 > 0$, $\bar{P}$ is supported on a compact (c_0, r_0)-regular set $A \subset \mathcal{R}$ satisfying $\nu_d(A \cap B_r(x)) \geq c_0\nu_d(B_r(x))$ for all $r \in [0, r_0]$ and all $x \in A$, where ν_d denotes the d-dimensional Lebesgue measure and $B_r(x)$ is a closed Euclidean ball in $\mathbb{R}^d$ centered at x and of radius $r > 0$. Moreover, for all $x \in A$, the Lebesgue density $\bar{f}$ of $\bar{P}$ satisfies $\bar{f}_{\min} \leq \bar{f}(x) \leq \bar{f}_{\max}$ for some $0 < \bar{f}_{\min} < \bar{f}_{\max}$, and $\bar{f}(x) = 0$ otherwise. In addition, $\bar{f} \in \Sigma(\gamma - 1, L, A)$.

We first derive the rate of convergence of CIS by assuming the exponential convergence rate of the corresponding regression function estimator.

Theorem 3.3.1 *(Upper Bound) Let $\widehat{\eta}_n$ be an estimator of the regression function η and let $\mathcal{R} \subset \mathbb{R}^d$ be a compact set. Let $\mathcal{P}$ be a set of probability distributions supported* *on $\mathcal{R} \times \{1, -1\}$ such that for some constants $C_1, C_2 > 0$, some positive sequence $a_n \to \infty$, and almost all x with respect to $\bar{P}$,*

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{\mathcal{D}}\Big(|\widehat{\eta}_n(x) - \eta(x)| \geq \delta\Big) \leq C_1 \exp(-C_2 a_n \delta^2) \tag{3.8}$$

holds for any $n > 1$ and $\delta > 0$, where $\mathbb{P}_{\mathcal{D}}$ is the probability with respect to $P^{\otimes n}$. Furthermore, if all the distributions $P \in \mathcal{P}$ satisfy the margin condition for a constant C_0, then the plug-in classification procedure Ψ corresponding to $\widehat{\eta}_n$ satisfies

$$\sup_{P \in \mathcal{P}} \mathit{CIS}(\Psi) \leq C a_n^{-\alpha/2},$$

for any $n > 1$ and some constant $C > 0$ depending only on α, C_0, C_1, and C_2.

It is worth noting that condition (3.8) holds for various types of estimators. For example, Theorem 3.2 in [21] showed that the local polynomial estimator satisfies (3.8) with $a_n = n^{2\gamma/(2\gamma+d)}$ when the bandwidth is of the order $n^{-1/(2\gamma+d)}$. In addition, Theorem 3.3.3 in Section 3.3.2 implies that (3.8) holds for the newly proposed SNN classifier with the same a_n. Hence, in both cases, the upper bound is of the order $n^{-\alpha\gamma/(2\gamma+d)}$.

We next derive the lower bound of CIS in Theorem 3.3.2. As will be seen, this lower bound implies that the obtained rate of CIS, that is, $n^{-\alpha\gamma/(2\gamma+d)}$, cannot be further improved for the plug-in classification procedure.

Theorem 3.3.2 *(Lower Bound) Let $\mathcal{P}_{\alpha,\gamma}$ be a set of probability distributions supported on $\mathcal{R} \times \{1, -1\}$ such that for any $P \in \mathcal{P}_{\alpha,\gamma}$, P satisfies the margin condition* (3.7)*, the regression function $\eta(x)$ belongs to the Hölder class $\Sigma(\gamma, L, \mathbb{R}^d)$, and the marginal distribution $\bar{P}$ satisfies the strong density assumption. Suppose further that $\mathcal{P}_{\alpha,\gamma}$ satisfies* (3.8) *with $a_n = n^{2\gamma/(2\gamma+d)}$ and $\alpha\gamma \leq d$. We have*

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathit{CIS}(\Psi) \geq C' n^{-\alpha\gamma/(2\gamma+d)},$$

for any $n > 1$ and some constant $C' > 0$ independent of n.

Theorems 3.3.1 and 3.3.2 together establish a sharp convergence rate of the CIS for the general plug-in classification procedure on the set $\mathcal{P}_{\alpha,\gamma}$. The requirement $\alpha\gamma \leq d$ in Theorem 3.3.2 implies that α and γ cannot be large simultaneously. As pointed out in [21], this is intuitively true because a very large γ implies a very smooth regression function η, while a large α implies that η cannot stay very long near $1/2$, and hence when η hits $1/2$, it should take off quickly. Lastly, we note that this rate is slower than n^{-1}, but approaches n^{-1} as the dimension d increases when $\alpha\gamma = d$.

3.3.2 Optimal Convergence Rates of SNN

This subsection illustrates that SNN's convergence rate of regret is minimax optimal and its convergence rate of CIS achieves the sharp rate established in Section 3.3.1. We further show the asymptotic difference between the SNN procedure and the OWNN procedure.

In Theorem 3.3.3 and Corollary 3 below, we consider SNN with $k^* \asymp n^{2\gamma/(2\gamma+d)}$ in Theorem 3.2.2, where $a_n \asymp b_n$ means the ratio sequence a_n/b_n stays away from zero and infinity as $n \to \infty$. Note that under Assumptions (A1)–(A4), we have $\gamma = 2$ and hence $k^* \asymp n^{4/(4+d)}$, which agrees with the formulation in Theorem 3.2.2.

Theorem 3.3.3 *For any $\alpha \geq 0$ and $\gamma \in (0, 2]$, the SNN procedure with any fixed $\lambda > 0$ satisfies*

$$\begin{aligned} \sup_{P\in\mathcal{P}_{\alpha,\gamma}} \mathit{Regret}(\mathit{SNN}) &\leq \tilde{C} n^{-(\alpha+1)\gamma/(2\gamma+d)}, \\ \sup_{P\in\mathcal{P}_{\alpha,\gamma}} \mathit{CIS}(\mathit{SNN}) &\leq C n^{-\alpha\gamma/(2\gamma+d)}, \end{aligned}$$

for any $n > 1$ and some constants $\tilde{C}, C > 0$, where $\mathcal{P}_{\alpha,\gamma}$ is defined in Theorem 3.3.2.

Corollary 3 below further investigates the difference between the SNN procedure (with $\lambda \neq B_1/B_2$) and the OWNN procedure in terms of both regret and CIS.

Corollary 3 *For any $\alpha \geq 0$, $\gamma \in (0, 2]$, we have, when $\lambda \neq B_1/B_2$,*

$$\begin{aligned} \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \Big\{ \mathit{Regret}(\mathit{SNN}) - \mathit{Regret}(\mathit{OWNN}) \Big\} &\asymp n^{-(1+\alpha)\gamma/(2\gamma+d)}, \\ \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \Big\{ \mathit{CIS}(\mathit{OWNN}) - \mathit{CIS}(\mathit{SNN}) \Big\} &\asymp n^{-\alpha\gamma/(2\gamma+d)}, \end{aligned} \tag{3.9}$$

where $\mathcal{P}_{\alpha,\gamma}$ is defined in Theorem 3.3.2.

Corollary 3 illustrates that the regret of SNN approaches that of the OWNN (from above) at a faster rate than the rate at which the CIS of OWNN approaches that of the SNN procedure (from above). Intuitively, this may imply that there is some room for improvement in CIS, while the difference in regret between the two methods is very small.

Remark 4 *Under Assumptions (A1)–(A4), which are the assumptions in Theorem 3.2.1, and the assumption that $\gamma = 2$, the conclusion in* (3.9) *can be strengthened to that for any $P \in \mathcal{P}_{1,2}$, $\mathit{CIS}(\mathit{OWNN}) - \mathit{CIS}(\mathit{SNN}) \asymp n^{-2/(d+4)}$, that is, the room for improvement in CIS is relatively large for all distributions in $\mathcal{P}_{1,2}$.*

3.4 Asymptotic Comparisons

This section starts with an asymptotic comparison of the CIS among several existing nearest neighbor classifiers. We then demonstrate that SNN significantly improves OWNN in CIS.

3.4.1 CIS Comparison of Existing Methods

We compare the CIS for three existing methods, kNN, OWNN and the bagged nearest neighbor (BNN) classifier. The kNN classifier is a special case of the WNN classifier with weight $w_{ni} = 1/k$ for $i = 1, \ldots, k$ and $w_{ni} = 0$ otherwise. Another special case of the WNN classifier is the BNN classifier. After generating subsamples from the original data set, the BNN classifier applies 1-nearest neighbor classifier to each bootstrapped subsample and returns the final predictor by majority vote. If the resample size m is sufficiently smaller than n, *i.e.*, $m \to \infty$ and $m/n \to 0$, the BNN classifier is shown to be a consistent classifier [60]. In particular, [60] showed that, for large n, the BNN classifier (with or without replacement) is approximately equivalent to a WNN classifier with the weight $w_{ni} = q(1-q)^{i-1}/[1-(1-q)^n]$ for $i = 1, \dots, n$, where q is the resampling ratio m/n.

We denote the CIS of the above classification procedures as CIS(kNN), CIS(BNN) and CIS(OWNN), respectively. Here k in the kNN classifier is selected as the one minimizing the regret [59]. The optimal q in the BNN classifier and the optimal weight in the OWNN classifier are both calculated based on their asymptotic relations with the optimal k in kNN, which were defined in (2.9) and (3.5) of [27]. Corollary 4 gives the pairwise CIS ratios of these classifiers. Note that these ratios depend on the feature dimension d only.

Corollary 4 *Under Assumptions (A1)-(A4) and the assumption that B_2 is positive, we have, as $n \to \infty$,*

$$\begin{aligned} \frac{CIS(OWNN)}{CIS(kNN)} &\longrightarrow 2^{2/(d+4)}\Big(\frac{d+2}{d+4}\Big)^{(d+2)/(d+4)}, \\ \frac{CIS(BNN)}{CIS(kNN)} &\longrightarrow 2^{-2/(d+4)}\Gamma(2+2/d)^{d/(d+4)}, \\ \frac{CIS(BNN)}{CIS(OWNN)} &\longrightarrow 2^{-4/(d+4)}\Gamma(2+2/d)^{d/(d+4)}\Big(\frac{d+4}{d+2}\Big)^{(d+2)/(d+4)}. \end{aligned}$$

The limiting CIS ratios in Corollary 4 are plotted in Figure 3.3 which delivers several messages. A major one is that the OWNN procedure is more stable than the kNN and BNN procedures for any d. The largest improvement of the OWNN procedure over kNN is achieved when $d = 4$ and the improvement diminishes as $d \to \infty$. The CIS ratio of BNN over kNN equals 1 when $d = 2$ and is less than 1 when $d > 2$, which is consistent with the common perception that bagging can generally reduce the variability of the nearest neighbor classifiers. Similar phenomenon has been shown in the ratio of their regrets [27]. Therefore, bagging can be used to improve the kNN procedure in terms of both accuracy and stability when $d > 2$. Furthermore, the CIS ratio of OWNN over BNN is less than 1 for all d and it quickly converges to 1

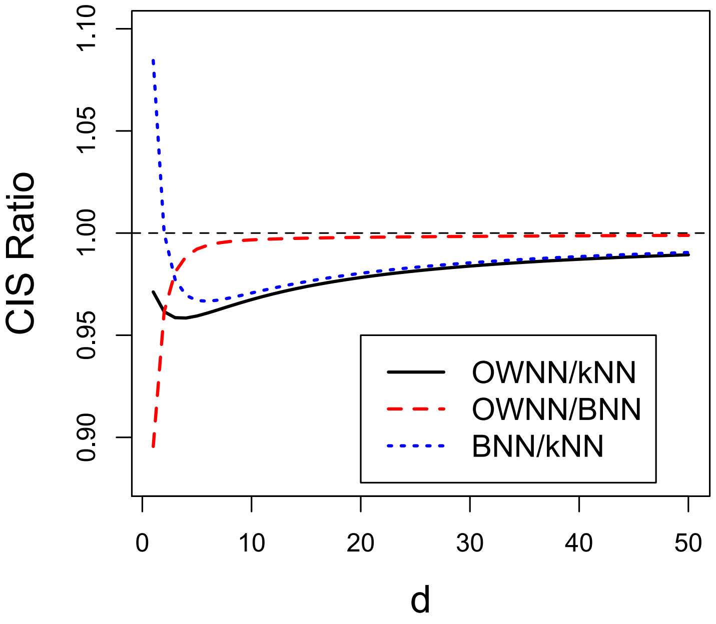

Figure 3.3. Pairwise CIS ratios between kNN, BNN and OWNN for different feature dimension d.

as d increases. This implies that although the BNN procedure is asymptotically less stable than the OWNN procedure, their difference sharply vanishes as d increases.

3.4.2 Comparisons between SNN and OWNN

Corollary 3 in Section 3.3.2 implies that OWNN and SNN share the same convergence rates of regret and CIS (note that OWNN is a special case of SNN). Hence, it is of great interest to go one step further and compare their relative magnitude. The asymptotic comparisons between SNN and OWNN are characterized in Corollary 5.

Corollary 5 *Under Assumptions (A1)-(A4) and the assumption that B_2 is positive, we have, as $n \to \infty$,*

$$\begin{aligned}\frac{Regret(SNN)}{Regret(OWNN)} &\longrightarrow \Big\{\frac{B_1}{\lambda B_2}\Big\}^{d/(d+4)}\Big\{\frac{4+d\lambda B_2/B_1}{4+d}\Big\},\\ \frac{CIS(SNN)}{CIS(OWNN)} &\longrightarrow \Big\{\frac{B_1}{\lambda B_2}\Big\}^{d/(2(d+4))},\end{aligned}$$

where constants B_1 and B_2 are defined in Proposition 2.

As can be seen from Corollary 5, both ratios of the SNN procedure over the OWNN procedure depend on λ, and unknown constants B_1 and B_2. Since $\lambda = (B_1+\lambda_0 B_3^2)/B_2$ in (3.6) and $B_3 = 4B_1/\sqrt{\pi}$ in (3.4), we further have

$$\frac{\text{Regret(SNN)}}{\text{Regret(OWNN)}} \longrightarrow \Big\{\frac{1}{1+16B_1\lambda_0/\pi}\Big\}^{d/(d+4)}\Big\{\frac{4+d(1+16B_1\lambda_0/\pi)}{4+d}\Big\}, \quad (3.10)$$

$$\frac{\text{CIS(SNN)}}{\text{CIS(OWNN)}} \longrightarrow \Big\{\frac{1}{1+16B_1\lambda_0/\pi}\Big\}^{d/(2(d+4))}. \quad (3.11)$$

For any $\lambda_0 > 0$, SNN has an improvement in CIS over the OWNN. As a mere illustration, we consider a case that the regret and the squared CIS are given equal weight, that is, $\lambda_0 = 1$. In this case, the ratios in (3.10) and (3.11) only depend on B_1 and d.

Figure 3.4 shows 3D plots of these ratios as functions of B_1 and d. As expected, the CIS of the SNN procedure is universally smaller than OWNN (ratios less than 1 on the right panel), while the OWNN procedure has a smaller regret (ratios greater than 1 on the left panel). For a fixed B_1, as the dimension d increases, the regret of SNN approaches that of OWNN, while the advantage of SNN in terms of CIS grows. For a fixed dimension d, as B_1 increases, the regret ratio between SNN and OWNN gets larger, but the CIS advantage of SNN also grows. According to the definition of B_1, a great value of B_1 indicates a harder problem for classification; see the discussion after Theorem 1 of [27].

Since SNN improves OWNN in CIS, but has a greater regret, it is of interest to know when the improvement of SNN in CIS is greater than its loss in regret. We thus consider the relative gain, defined as the absolute ratio of the percentages of CIS reduction and regret increment, that is, $|\Delta\text{CIS}/\Delta\text{Regret}|$, where

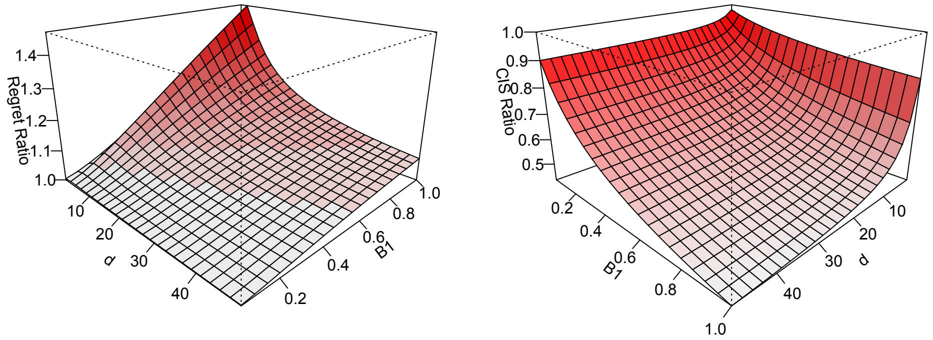

Figure 3.4. Regret ratio and CIS ratio of SNN over OWNN as functions of B_1 and d. The darker the color, the larger the value.

$\Delta\text{CIS} = [\text{CIS(SNN)} - \text{CIS(OWNN)}]/\text{CIS(OWNN)}$ and $\Delta\text{Regret} = [\text{Regret(SNN)} - \text{Regret(OWNN)}]/\text{Regret(OWNN)}$. As an illustration, when $\lambda_0 = 1$, we have the relative gain converges to $\left[1 - (1 + 16B_1/\pi)^{-d/(2d+8)}\right] \left[(1 + 16B_1/\pi)^{4/(d+4)} - 1\right]^{-1}$. Figure 3.5 shows the log(relative gain) as a function of B_1 and d. For most combinations of B_1 and d, the logarithm is greater than 0 (shown in grey in Figure 3.5), where SNN's improvement in CIS is greater than its loss in regret. In particular, when $B_1 \leq 0.2$, the logarithm of relative gain is positive for all d.

3.5 Tuning Parameter Selection

To select the parameter λ for the SNN classifier, we first identify a set of values for λ whose corresponding (estimated) risks are among the smallest, and then choose from them an optimal one which has the minimal estimated CIS. Let $\widehat{\phi}_{\mathcal{D}}^{\lambda}$ denote an SNN classifier with parameter λ trained from sample $\mathcal{D}$. Given a predetermined set of tuning parameter values $\Lambda = \{\lambda_1, \dots, \lambda_K\}$, the tuning parameter $\widehat{\lambda}$ is selected using Algorithm 1 below, which involves estimating the CIS and risk in Steps 1–3 and a two-stage selection in Steps 4 and 5.

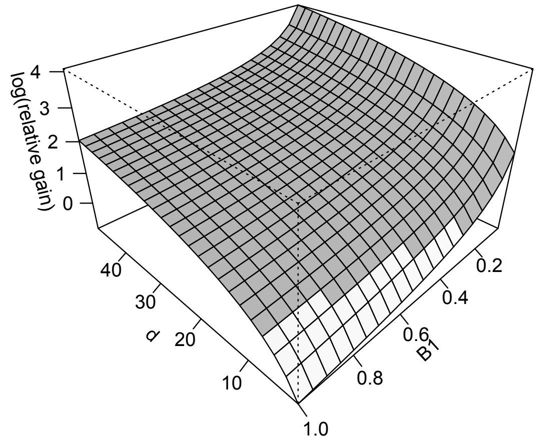

Figure 3.5. Logarithm of relative gain of SNN over OWNN as a function of B_1 and d when $\lambda_0 = 1$. The grey (white) color represents the case where the logarithm of relative gain is greater (less) than 0.

Algorithm 1:

Step 1. Randomly partition $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ into five subsets I_i, $i = 1, \cdots, 5$.

Step 2. For $i = 1$, let I_1 be the test set and I_2, I_3, I_4 and I_5 be training sets. Obtain predicted labels from $\widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j)$ and $\widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j)$ respectively for each $X_j \in I_1$. Estimate the CIS and risk of the classifier with parameter λ by

$$\begin{aligned}
\widehat{\text{CIS}}_i(\lambda) &= \frac{1}{|I_1|} \sum_{(X_j, Y_j) \in I_1} \mathbb{1}\{\widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j) \neq \widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j)\}, \\
\widehat{\text{Risk}}_i(\lambda) &= \frac{1}{2|I_1|} \sum_{(X_j, Y_j) \in I_1} \Big\{ \mathbb{1}\{\widehat{\phi}^{\lambda}_{I_2 \cup I_3}(X_j) \neq Y_j\} + \mathbb{1}\{\widehat{\phi}^{\lambda}_{I_4 \cup I_5}(X_j) \neq Y_j\} \Big\}.
\end{aligned}$$

Step 3. Repeat *Step 2* for $i = 2, \ldots, 5$ and estimate the CIS and risk, with I_i being the test set and the rest being the training sets. Finally, the estimated CIS and risk are,

$$\widehat{\text{CIS}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \widehat{\text{CIS}}_i(\lambda), \quad \widehat{\text{Risk}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \widehat{\text{Risk}}_i(\lambda).$$

Step 4. Perform *Step 2* and *Step 3* for each $\lambda_k \in \Lambda$. Denote the set of tuning parameters with top accuracy as

$$\mathcal{A} := \{\lambda : \widehat{\text{Risk}}(\lambda) \text{ is less than the 10th percentile of } \widehat{\text{Risk}}(\lambda_k),\ k = 1, \dots, K\}.$$

Step 5. Output the optimal tuning parameter $\widehat{\lambda}$ as

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \mathcal{A}} \widehat{\text{CIS}}(\lambda).$$

In our experiments, the predetermined set of tuning parameters Λ are of size 100. In Step 1, the sample sizes of the subsets I_i are chosen to be roughly equal. The estimation scheme based on cross-validation in Steps 1 − 3 can be replaced by other data re-sampling strategies such as bootstrap, while in the latter case each subsets are no longer independent. In Step 4, the threshold 10% reflects how the set of most accurate classifiers is defined. Based on our experiments, the choice of 10% leads to superior performance.

Compared with the tuning method for the kNN classifier, which minimizes the estimated risk, Algorithm 1 requires additional estimation of the CIS. However, as revealed in Step 2, the estimation of the CIS is concurrently conducted with the estimation of the risk. Therefore, the complexity of tuning for our SNN classifier is in the same order as that for the kNN classifier. As will be seen in the numerical experiment, the additional effort on estimating the CIS leads to improvement over existing nearest neighbor methods in both accuracy and stability.

3.6 Numerical Studies

We first validate our theoretical findings using a simple example, and then illustrate the improvements of the SNN classifier over existing nearest neighbor classifiers using simulations and real examples.

3.6.1 Validation of Asymptotically Equivalent Forms

This subsection aims to support the asymptotically equivalent forms of CIS derived in Theorem 3.2.1 and the CIS and regret ratios in Corollary 5. We focus on a multivariate Gaussian example in which regret and CIS have explicit expressions.

Assume that the underlying distributions of both classes are $P_1 \sim N(0_2, \mathbb{I}_2)$ and $P_2 \sim N(1_2, \mathbb{I}_2)$ and the prior class probability $\pi_1 = 1/3$. We choose $\mathcal{R} = [-2, 3]^2$, which covers at least 95% probability of the sampling region, and set $n = 50, 100, 200$ and 500. In addition, a test set with 1000 observations were independently generated. The estimated risk and CIS were calculated based on 100 replications. In this example, some calculus practice leads to $B_1 = 0.1299$, $B_2 = 10.68$ and $B_3 = 0.2931$. According to Proposition 2, Theorems 3.2.1 and 3.2.2, we obtain that

$$\text{Regret(SNN)} = \frac{0.1732}{k^*} + \frac{4.7467(k^*)^2}{n^2} \tag{3.12}$$

$$\text{CIS(SNN)} = \frac{0.3385}{(k^*)^{1/2}}, \tag{3.13}$$

with $k^* = \lfloor 1.5^{1/3} \lambda^{1/3} n^{2/3} \rfloor$. For mere illustration, we choose $\lambda = (B_1 + B_3^2)/B_2$, which corresponds to $\lambda_0 = 1$. So we have $k^* = \lfloor 0.3118 n^{2/3} \rfloor$. Similarly, the asymptotic regret and CIS of OWNN are (3.12) and (3.13) with $k^* = \lfloor 0.2633 n^{2/3} \rfloor$ due to (2.4) in [27].

In Figures 3.6, we plot the asymptotic CIS of the SNN and OWNN classifiers computed using the above formulae, shown in red curve, along with the estimated CIS based on the simulated data, shown as the box plots over 100 replications. As the sample size n increases, the estimated CIS approximates its asymptotic value very well. For example, when $n = 500$, the asymptotic CIS of the SNN (OWNN) classifier is 0.078 (0.085) while the estimated CIS is 0.079 (0.086).

Similarly, in Figure 3.7, we plot the asymptotic risk, that is, the asymptotic regret in (3.12) plus the true Bayes risk (0.215 in this example), for the SNN and OWNN classifiers, along with the estimated risk. Here we compute the Bayes risk by Monte Carlo integration. Again the difference of the estimated risk and asymptotic risk decreases as the sample size grows.

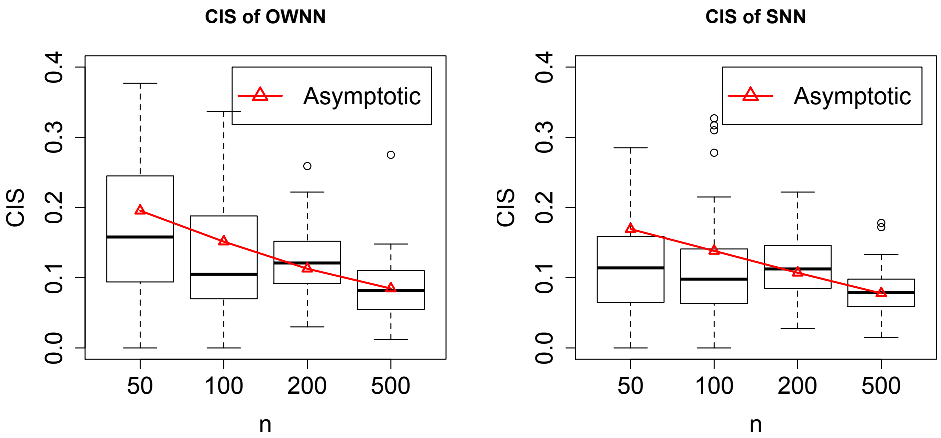

Figure 3.6. Asymptotic CIS (red curve) and estimated CIS (box plots over 100 simulations) for OWNN (left) and SNN (right) procedures. These plots show that the estimated CIS converges to its asymptotic equivalent value as n increases.

Furthermore, according to (3.11), the asymptotic CIS ratio of the SNN classifier over the OWNN classifier is 0.9189 in this example, and the empirically estimated CIS ratios are 0.6646, 0.9114, 0.8940 and 0.9219, for $n = 50, 100, 200, 500$. This indicates that the estimated CIS ratio converges to its asymptotic value as n increases. However, by (3.10), the asymptotic regret ratio of the SNN classifier over the OWNN classifier is 1.0305, while the estimated ones are 1.0224, 1.1493, 0.3097 and 0.1136, for $n = 50, 100, 200, 500$. It appears that the estimated regret ratio matches with its asymptotic value for small sample size, but they differ for large n. This may be caused by the fact that the classification errors are very close to Bayes risk for large n and hence the estimated regret ratio has a numerical issue. For example, when $n = 500$, the average errors of the SNN classifier and the OWNN classifier are 0.2152 and 0.2161, respectively, while the Bayes risk is 0.215 (see Figure 3.7). A similar issue was also reported in [27].

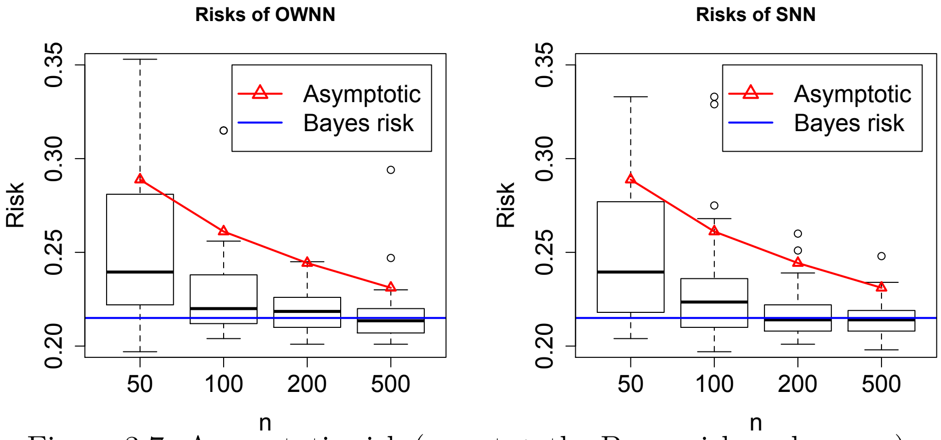

Figure 3.7. Asymptotic risk (regret + the Bayes risk; red curves) and estimated risk (black box plots) for OWNN (left) and SNN procedures (right). The blue horizontal line indicates the Bayes risk, 0.215. These plots show that the estimated risk converges to its asymptotic version (and also the Bayes risk) as n increases.

3.6.2 Simulations

In this section, we compare SNN with the kNN, OWNN and BNN classifiers. The parameter k in kNN was tuned from 100 equally spaced grid points from 5 to $n/2$. For a fair comparison, in the SNN classifier, the parameter λ was tuned so that the corresponding parameter k^* (see Theorem 3.2.2) were equally spaced and fell into the roughly same range.

In Simulation 1, we assumed that the two classes were from $P_1 \sim N(0_d, \mathbb{I}_d)$ and $P_2 \sim N(\mu_d, \mathbb{I}_d)$ with the prior probability $\pi_1 = 1/3$. We set sample size $n = 200$ and chose μ such that the resulting B_1 was fixed as 0.1 for different d. Specifically, in Section 3.7.8 we show that

$$B_1 = \frac{\sqrt{2\pi}}{3\pi\mu d} \exp\left(-\frac{(\mu d/2 - \ln 2/\mu)^2}{2d}\right). \tag{3.14}$$

Hence, we set $\mu = 2.076$, 1.205, 0.659, 0.314, 0.208 for $d = 1, 2, 4, 8$ and 10, respectively.

In Simulation 2, the training data set were generated by setting $n = 200$, $d = 2$ or 5, $P_1 \sim 0.5N(0_d, \mathbb{I}_d) + 0.5N(3_d, 2\mathbb{I}_d)$, $P_2 \sim 0.5N(1.5_d, \mathbb{I}_d) + 0.5N(4.5_d, 2\mathbb{I}_d)$, and $\pi_1 = 1/2$ or $1/3$.

Simulation 3 has the same setting as Simulation 2, except that $P_1 \sim 0.5N(0_d, \Sigma) + 0.5N(3_d, 2\Sigma)$ and $P_2 \sim 0.5N(1.5_d, \Sigma) + 0.5N(4.5_d, 2\Sigma)$, where Σ is the Toeplitz matrix whose jth entry of the first row is 0.6^{j-1}.

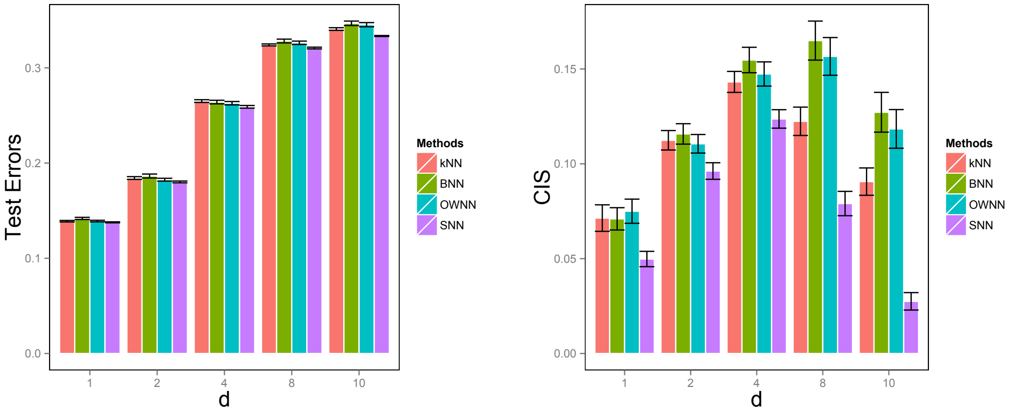

Figure 3.8. Average test errors and CIS's (with standard error bar marked) of the kNN, BNN, OWNN, and SNN methods in Simulation 1. The x-axis indicates different settings with various dimensions. Within each setting, the four methods are horizontally lined up (from the left are kNN, BNN, OWNN, and SNN).

Simulation 1 is a relatively easy classification problem. Simulation 2 examines the bimodal effect and Simulation 3 combines bimodality with dependence between the components. In each simulation setting, a test data set of size 1000 is independently generated and the average classification error and average (estimated) CIS for the test set are reported over 100 replications. To calculate the average CIS, for each replication we build two classifiers based on the randomly divided training data, and then estimate CIS by the average disagreement of these two classifiers on the test data.

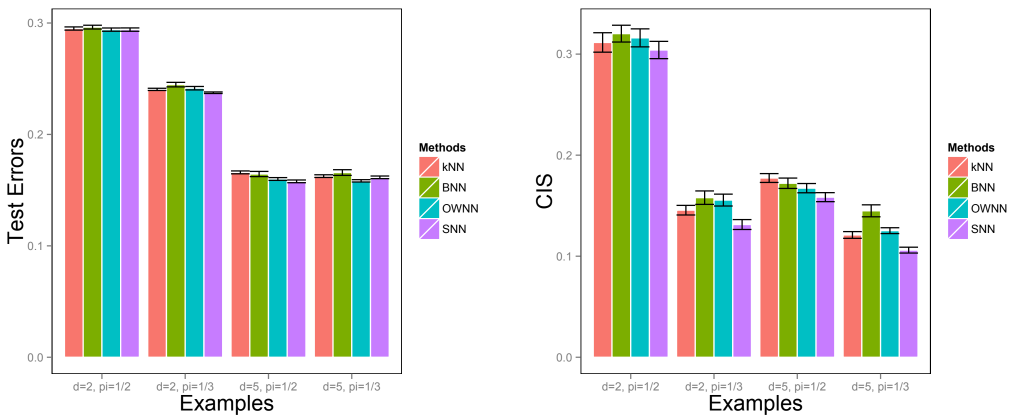

Figure 3.9. Average test errors and CIS's (with standard error bar marked) of the kNN, BNN, OWNN, and SNN methods in Simulation 2. The ticks on the x-axis indicate the dimensions and prior class probability π for different settings. Within each setting, the four methods are horizontally lined up (from the left are kNN, BNN, OWNN, and SNN).

Figure 3.8 shows the average error (on the left) and CIS (on the right) for Simulation 1. As a first impression, the error is similar among different classification methods, while the CIS differs a lot. In terms of the stability, SNN always has the smallest CIS; in particular, as d increases, the improvement of SNN over all other procedures becomes even larger. This agrees with the asymptotic findings in Section 3.4.2. For example, when $d = 10$, all the kNN, BNN, and OWNN procedures are at least five times more unstable than SNN. In terms of accuracy, SNN obtains the minimal test errors in all five scenarios, although the improvements in test errors are not significant when $d = 1,\ 2$ or 4. This result reveals that although SNN would be asymptotically less accurate than OWNN in theory, the actual empirical difference in accuracy between SNN and OWNN is often ignorable. In contrast, SNN additionally relies on the classification stability, which in turn provides a significant improvement in stability over other nearest neighbor procedures.

Figures 3.9 and 3.10 summarize the results for Simulations 2 and 3. Again, in general, the difference in CIS is much obvious than the difference in the error. The

SNN procedure obtains the minimal CIS for all 8 cases. Interestingly, the improvements are significant in all the four cases when $\pi_1 = 1/3$. Moreover, among 3 out of the 8 cases, our SNN achieves the smallest test errors and the improvements are significant. Even in cases where the error is not the smallest, the performance of SNN is close to the best classifier.

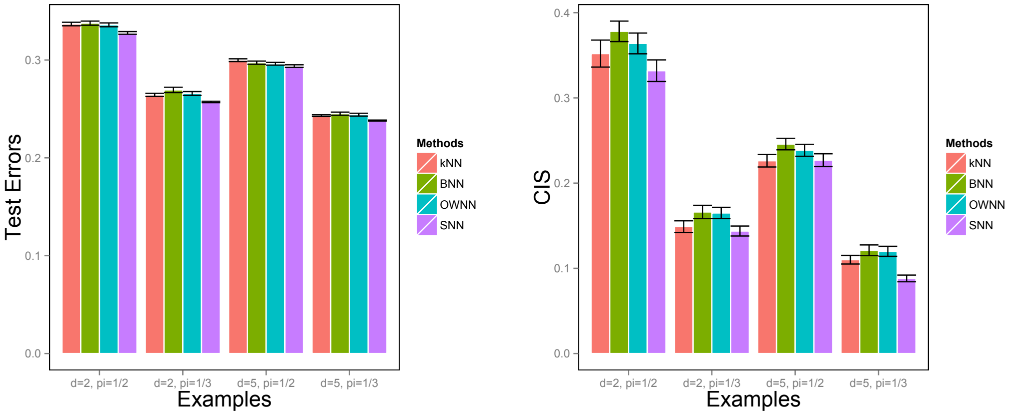

Figure 3.10. Average test errors and CIS's (with standard error bar marked) of the kNN, BNN, OWNN, and SNN methods in Simulation 3. The ticks on the x-axis indicate the dimensions and prior class probability π for different settings. Within each setting, the four methods are horizontally lined up (from the left are kNN, BNN, OWNN, and SNN).

3.6.3 Real Examples

In this subsection, we extend the comparison to four real data sets publicly available in the UCI Machine Learning Repository [46].

The first data set is the breast cancer data set (*breast*) collected by [47]. There are 683 samples and 10 experimental measurement variables. The binary class label indicates whether the sample is benign or malignant. These 683 samples arrived periodically. In total, there are 8 groups of samples which reflect the chronological order of the data. A good classification procedure is expected to produce a stable

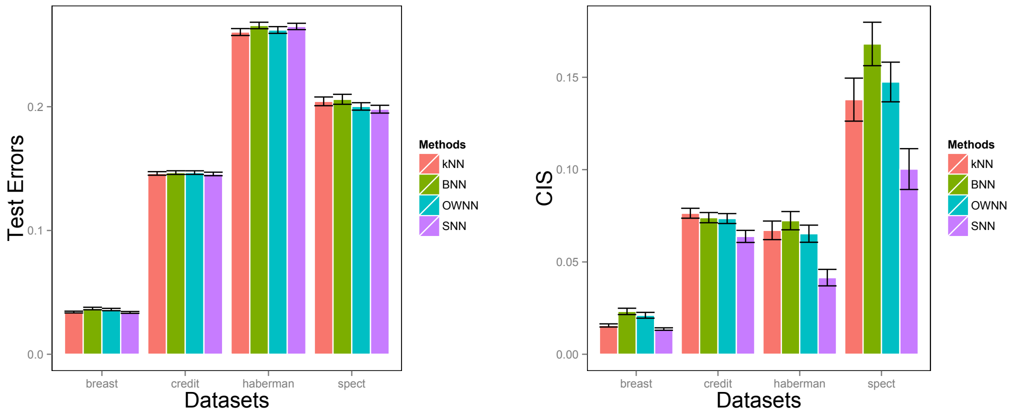

Figure 3.11. Average test errors and CIS's (with standard error bar marked) of the kNN, BNN, OWNN and SNN methods for four data examples. The ticks on the x-axis indicate the names of the examples. Within each example, the four methods are horizontally lined up (from the left are kNN, BNN, OWNN, and SNN).

classifier across these groups of samples. The second data set is the credit approval data set (*credit*). It consists of 690 credit card applications and each application has 14 attributes reflecting the user information. The binary class label refers to whether the application is positive or negative. The third data set is the haberman's survival data set (*haberman*) which contains 306 cases from study conducted on the survival of patients who had undergone surgery for breast cancer. It has three attributes, age, patient's year of operation, and number of positive axillary nodes detected. The response variable indicates the survival status: either the patient survived 5 years or longer or the patient died within 5 years. The last data set is the SPECT heart data set (*spect*) which describes the diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the 267 image sets (patients) had 22 binary feature patterns and was classified into two classes: normal and abnormal.

For each data set, we randomly split it into training and test sets with the equal size. The same tuning procedure as in the simulation is applied here. We compute the classification error and (estimated) CIS on each test set. These procedures are repeated 100 times and the average error and CIS are reported in Figure 3.11.

Similar to the simulation results, the SNN procedure obtains the minimal CIS in all four real data sets and the improvements in CIS are significant. The errors of OWNN and our SNN procedures have no significant difference, although OWNN is theoretically the best in accuracy. These real experiments further illustrate that, with almost the same classification accuracy, our SNN procedure can achieve a significant improvement in stability.

3.7 Technical Proofs

This section contains detailed proofs of all theoretical results and the calculation of B_1 in a particular example.

3.7.1 Proof of Theorem 3.2.1

Before we prove Theorem 3.2.1, we first introduce a useful Lemma.

Lemma 5 *For any distribution function G, constant a, and constant $b > 0$, we have*

$$\begin{aligned}\int_{-\infty}^{\infty} \{G(-bu-a) - \mathbb{1}\{u<0\}\}\, du &= -\frac{1}{b}\left\{a + \int_{-\infty}^{\infty} t dG(t)\right\},\\ \int_{-\infty}^{\infty} u\{G(-bu-a) - \mathbb{1}\{u<0\}\}\, du &= \frac{1}{b^2}\left\{\frac{1}{2}a^2 + \frac{1}{2}\int_{-\infty}^{\infty} t^2 dG(t) + a\int_{-\infty}^{\infty} t dG(t)\right\}.\end{aligned}$$

Proof of Lemma 5: We show the second equality. The proof of the first equality is similar. Note

$$\begin{aligned}&\int_{-\infty}^{\infty} u\{G(-bu-a) - \mathbb{1}\{u<0\}\}\, du\\ =\ &\int_{-\infty}^{0} u\{G(-bu-a) - 1\}\, du + \int_{0}^{\infty} uG(-bu-a)du \qquad (3.15)\end{aligned}$$

After substitute $t = -bu - a$ for each term, we have

$$\begin{aligned}\int_{-\infty}^{0} u\left\{G(-bu-a)-1\right\}du &= \frac{1}{b^2}\int_{-a}^{\infty}(t+a)(1-G(t))dt \\ \int_{0}^{\infty} uG(-bu-a)du &= \frac{1}{b^2}\int_{-\infty}^{-a}(t+a)(-G(t))dt\end{aligned}$$

Plugging these two into (3.15), we have

$$\begin{aligned}&\int_{-\infty}^{\infty} u\left\{G(-bu-a)-\mathbb{1}\{u<0\}\right\}du \\ = \ & \frac{1}{b^2}\left\{-\int_{-\infty}^{-a} tG(t)dt - a\int_{-\infty}^{-a} G(t)dt + \int_{-a}^{\infty} t(1-G(t))dt + a\int_{-a}^{\infty}(1-G(t))dt\right\} \\ = \ & \frac{1}{b^2}\left\{I + II + III + IV\right\}.\end{aligned}$$

Applying integration by part, we can calculate

$$\begin{aligned}I &= -\frac{1}{2}\Big[a^2G(-a) - \int_{-\infty}^{-a} t^2dG(t)\Big] \\ II &= a\Big[aG(-a) + \int_{-\infty}^{-a} tdG(t)\Big] \\ III &= \frac{1}{2}\Big[-a^2(1-G(-a)) + \int_{-a}^{\infty} t^2dG(t)\Big] \\ IV &= a\Big[a(1-G(-a)) + \int_{-a}^{\infty} tdG(t)\Big]\end{aligned}$$

Plugging I-IV into (3.15) leads to desirable equality. This concludes the proof of Lemma 5. ■

Proof of Theorem 3.2.1: Note that CIS(WNN) $= \mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2,X}\Big(\widehat{\phi}_{n1}^{\boldsymbol{w}_n}(X) \neq \widehat{\phi}_{n2}^{\boldsymbol{w}_n}(X)\Big)$ can be expressed in the following way.

$$\begin{aligned}&\text{CIS(WNN)} \\ = \ & \mathbb{E}_X\Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2}\Big(\widehat{\phi}_{\mathcal{D}_1}^{\boldsymbol{w}_n}(X) \neq \widehat{\phi}_{\mathcal{D}_2}^{\boldsymbol{w}_n}(X)\Big|X\Big)\Big] \\ = \ & \mathbb{E}_X\Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2}\Big(\widehat{\phi}_{\mathcal{D}_1}^{\boldsymbol{w}_n}(X) = 1, \widehat{\phi}_{\mathcal{D}_2}^{\boldsymbol{w}_n}(X) = -1\Big|X\Big)\Big] \\ & + \mathbb{E}_X\Big[\mathbb{P}_{\mathcal{D}_1,\mathcal{D}_2}\Big(\widehat{\phi}_{\mathcal{D}_1}^{\boldsymbol{w}_n}(X) = -1, \widehat{\phi}_{\mathcal{D}_2}^{\boldsymbol{w}_n}(X) = 1\Big|X\Big)\Big] \\ = \ & \mathbb{E}_X\Big[2\mathbb{P}_{\mathcal{D}_1}\Big(\widehat{\phi}_{\mathcal{D}_1}^{\boldsymbol{w}_n}(X) = 1|X\Big)\Big(1 - \mathbb{P}_{\mathcal{D}_1}\Big(\widehat{\phi}_{\mathcal{D}_1}^{\boldsymbol{w}_n}(X) = 1|X\Big)\Big)\Big],\end{aligned}$$

where the last equality is valid because $\mathcal{D}_1$ and $\mathcal{D}_2$ are i.i.d. samples. Without loss of generality, we consider a generic sample $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$. Given

$X = x$, we define $(X_{(i)}, Y_{(i)})$ such that $\|X_{(1)} - x\| \le \|X_{(2)} - x\| \le \ldots \le \|X_{(n)} - x\|$ with $\|\cdot\|$ the Euclidean norm. Denote the estimated regression function $S_n(x) = \sum_{i=1}^n w_{ni} \mathbb{1}\{Y_{(i)} = 1\}$. We have

$$\mathbb{E}_X\Big[\mathbb{P}\Big(\widehat{\phi}_{\mathcal{D}}^{\boldsymbol{w}_n}(X) = 1|X\Big)\Big] = \int_{\mathcal{R}} \mathbb{P}\Big(S_n(x) \ge 1/2\Big) d\bar{P}(x),$$
$$\mathbb{E}_X\Big[\mathbb{P}^2\Big(\widehat{\phi}_{\mathcal{D}}^{\boldsymbol{w}_n}(X) = 1|X\Big)\Big] = \int_{\mathcal{R}} \mathbb{P}^2\Big(S_n(x) \ge 1/2\Big) d\bar{P}(x),$$

where $\bar{P}(x)$ is the marginal distribution of X. For the sake of simplicity, $\mathbb{P}$ denotes the probability with respect to $\mathcal{D}$. Hence, CIS satisfies

$$\begin{aligned} \mathrm{CIS(WNN)}/2 &= \int_{\mathcal{R}} \mathbb{P}(S_n(x) \ge 1/2)\Big(1 - \mathbb{P}(S_n(x) \ge 1/2)\Big) d\bar{P}(x) \\ &= \int_{\mathcal{R}} \{\mathbb{P}(S_n(x) < 1/2) - \mathbb{1}\{\eta(x) < 1/2\}\} \, d\bar{P}(x) \\ &\quad - \int_{\mathcal{R}} \big\{\mathbb{P}^2(S_n(x) < 1/2) - \mathbb{1}\{\eta(x) < 1/2\}\big\} \, d\bar{P}(x) \end{aligned}$$

Denote the boundary $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$. For $\epsilon > 0$, let $\mathcal{S}^{\epsilon\epsilon} = \{x \in \mathbb{R}^d : \eta(x) = 1/2 \text{ and } \mathrm{dist}(x, \mathcal{S}) < \epsilon\}$, where $\mathrm{dist}(x, \mathcal{S}) = \inf_{x_0 \in \mathcal{S}} \|x - x_0\|$. We will focus on the set

$$\mathcal{S}^{\epsilon} = \left\{ x_0 + t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|} : x_0 \in \mathcal{S}^{\epsilon\epsilon}, |t| < \epsilon \right\}.$$

Let $\mu_n(x) = \mathbb{E}\{S_n(x)\}$, $\sigma_n^2(x) = \mathrm{Var}\{S_n(x)\}$, and $\epsilon_n = n^{-\beta d/4}$. Denote $s_n^2 = \sum_{i=1}^n w_{ni}^2$ and $t_n = n^{-2/d} \sum_{i=1}^n \alpha_i w_{ni}$. [27] showed that, uniformly for $\boldsymbol{w}_n \in W_{n,\beta}$,

$$\sup_{x \in \mathcal{S}^{\epsilon_n}} |\mu_n(x) - \eta(x) - a(x) t_n| = o(t_n), \tag{3.16}$$
$$\sup_{x \in \mathcal{S}^{\epsilon_n}} \left|\sigma_n^2(x) - \frac{1}{4} s_n^2\right| = o(s_n^2). \tag{3.17}$$

We organize our proof in three steps. In Step 1, we focus on analyzing on the set $\mathcal{R} \cap \mathcal{S}^{\epsilon_n}$; in Step 2, we focus on the complement set $\mathcal{R} \backslash \mathcal{S}^{\epsilon_n}$; Step 3 combines the results and applies a normal approximation to yield the final conclusion.

Step 1: For $x_0 \in \mathcal{S}$ and $t \in \mathbb{R}$, denote $x_0^t = x_0 + t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|$. Denote $\bar{f} = \pi_1 f_1 + (1 - \pi_1) f_2$ as the Radon-Nikodym derivative with respect to Lebesgue measure of the restriction of $\bar{P}$ to $\mathcal{S}^{\epsilon_n}$ for large n. We need to show that, uniformly for $\boldsymbol{w}_n \in W_{n,\beta}$,

$$\int_{\mathcal{R}\cap\mathcal{S}^{\epsilon_n}} \{\mathbb{P}(S_n(x) < 1/2) - \mathbb{1}\{\eta(x) < 1/2\}\} \, d\bar{P}(x) =$$

$$\int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t) \Big\{\mathbb{P}\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t < 0\}\Big\} \, dt d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\} \quad (3.18)$$

$$\int_{\mathcal{R}\cap\mathcal{S}^{\epsilon_n}} \Big\{\mathbb{P}^2(S_n(x) < 1/2) - \mathbb{1}\{\eta(x) < 1/2\}\Big\} \, d\bar{P}(x) =$$

$$\int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t) \Big\{\mathbb{P}^2\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t < 0\}\Big\} \, dt d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\}(3.19)$$

According to [27], for large n, we define the map $\phi(x_0, t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}) = x_0^t$, and note that

$$\det \dot{\phi}\Big(x_0, t\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}\Big) dt d\mathrm{Vol}^{d-1}(x_0) = \{1 + o(1)\} dt d\mathrm{Vol}^{d-1}(x_0),$$

uniformly in $(x_0, t\dot{\eta}(x_0)/\|\dot{\eta}(x_0)\|)$ for $x_0 \in \mathcal{S}$ and $|t| < \epsilon_n$, where det is the determinant. Then the theory of integration on manifolds [61] implies that, uniformly for $\boldsymbol{w}_n \in W_{n,\beta}$,

$$\int_{\mathcal{S}^{\epsilon_n}} \{\mathbb{P}(S_n(x) < 1/2) - \mathbb{1}\{\eta(x) < 1/2\}\} \, d\bar{P}(x) =$$

$$\int_{\mathcal{S}^{\epsilon_n\epsilon_n}} \int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t) \Big\{\mathbb{P}\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t < 0\}\Big\} \, dt d\mathrm{Vol}^{d-1}(x_0)\{1 + o(1)\}.$$

Furthermore, we can replace $\mathcal{S}^{\epsilon_n}$ with $\mathcal{R}\cap\mathcal{S}^{\epsilon_n}$ since $\mathcal{S}^{\epsilon_n}\backslash\mathcal{R} \subseteq \{x \in \mathbb{R}^d : \mathrm{dist}(x, \partial\mathcal{S}) < \epsilon_n\}$ and the latter has volume $O(\epsilon_n^2)$ by Weyl's tube formula [61]. Similarly, we can safely replace $\mathcal{S}^{\epsilon_n\epsilon_n}$ with $\mathcal{S}$. Therefore, (3.18) holds. Similar arguments imply (3.19).

Step 2: Bound the contribution to CIS from $\mathcal{R}\backslash\mathcal{S}^{\epsilon_n}$. We show that, for all $M > 0$,

$$\sup_{\boldsymbol{w}_n \in W_{n,\beta}} \int_{\mathcal{R}\backslash\mathcal{S}^{\epsilon_n}} \Big\{\mathbb{P}\Big(S_n(x) < 1/2\Big) - \mathbb{1}\{\eta(x) < 1/2\}\Big\} \, d\bar{P}(x) \quad = \quad O(n^{-M}),(3.20)$$

$$\sup_{\boldsymbol{w}_n \in W_{n,\beta}} \int_{\mathcal{R}\backslash\mathcal{S}^{\epsilon_n}} \Big\{\mathbb{P}^2\Big(S_n(x) < 1/2\Big) - \mathbb{1}\{\eta(x) < 1/2\}\Big\} \, d\bar{P}(x) \quad = \quad O(n^{-M}).(3.21)$$

Here (3.20) follows from the fact $|\mathbb{P}(S_n(x) < \frac{1}{2}) - \mathbb{1}\{\eta(x) < 1/2\}| = O(n^{-M})$ for all $M > 0$, uniformly for $\boldsymbol{w}_n \in W_{n,\beta}$ and $x \in \mathcal{R}\backslash\mathcal{S}^{\epsilon_n}$ [27]. Furthermore, (3.21) holds since

$$\Big|\mathbb{P}^2\Big(S_n(x) < 1/2\Big) - \mathbb{1}\{\eta(x) < 1/2\}\Big| \le 2\Big|\mathbb{P}\Big(S_n(x) < 1/2\Big) - \mathbb{1}\{\eta(x) < 1/2\}\Big|.$$

Step 3: In the end, we will show

$$\begin{aligned}
&\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\mathbb{P}\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t<0\}\Big\}dtd\mathrm{Vol}^{d-1}(x_0) \\
&\quad - \int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\mathbb{P}^2\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t<0\}\Big\}dtd\mathrm{Vol}^{d-1}(x_0) \\
= \;& \frac{1}{2}B_3 s_n + o(s_n + t_n). \qquad (3.22)
\end{aligned}$$

We first apply the nonuniform version of Berry-Esseen Theorem to approximate $\mathbb{P}(S_n(x_0^t) < 1/2)$. Let $Z_i = (w_{ni}\mathbb{1}\{Y_{(i)} = 1\} - w_{ni}\mathbb{E}[\mathbb{1}\{Y_{(i)} = 1\}])/\sigma_n(x)$ and $W = \sum_{i=1}^n Z_i$. Note that $\mathbb{E}(Z_i) = 0$, $\mathrm{Var}(Z_i) < \infty$, and $\mathrm{Var}(W) = 1$. Then the nonuniform Berry-Esseen Theorem [62] implies that

$$\Big|\mathbb{P}(W \le y) - \Phi(y)\Big| \le \frac{M_1}{n^{1/2}(1+|y|^3)},$$

where Φ is the standard normal distribution function and M_1 is a constant. Therefore,

$$\sup_{x_0\in\mathcal{S}}\sup_{t\in[-\epsilon_n,\epsilon_n]}\Big|\mathbb{P}\Big(\frac{S_n(x_0^t) - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \le y\Big) - \Phi(y)\Big| \le \frac{M_1}{n^{1/2}(1+|y|^3)}. \qquad (3.23)$$

Thus, we have

$$\begin{aligned}
&\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\mathbb{P}\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t<0\}\Big\}dtd\mathrm{Vol}^{d-1}(x_0) \\
&\quad = \int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\Phi\Big(\frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)}\Big) - \mathbb{1}\{t<0\}\Big\}dtd\mathrm{Vol}^{d-1}(x_0) + o(s_n^2 + t_n^2),
\end{aligned}$$

where the remainder term $o(s_n^2 + t_n^2)$ is due to (3.23) by slightly modifying the proof of A.21 in [27].

Furthermore, Taylor expansion leads to

$$\bar{f}(x_0^t) = \bar{f}(x_0) + (\dot{\bar{f}}(x_0))^T\frac{\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}t + o(t).$$

Denote $\Delta_0 = \Phi\Big(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\Big) - \mathbb{1}\{t<0\}$. We have

$$\begin{aligned}
&\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\mathbb{P}\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t<0\}\Big\}dtd\mathrm{Vol}^{d-1}(x_0) \qquad (3.24) \\
= \;& \int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0)\Delta_0 dtd\mathrm{Vol}^{d-1}(x_0) + \int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \frac{\dot{\bar{f}}(x_0)^T\dot{\eta}(x_0)t}{\|\dot{\eta}(x_0)\|}\Delta_0 dtd\mathrm{Vol}^{d-1}(x_0) + R_1,
\end{aligned}$$

where $R_1 = R_{11} + R_{12} + o(s_n^2 + t_n^2)$ with

$$\begin{aligned} R_{11} &= \int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0)\Delta dt d\text{Vol}^{d-1}(x_0), \\ R_{12} &= \int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \frac{\dot{\bar{f}}(x_0)^T \dot{\eta}(x_0) t}{\|\dot{\eta}(x_0)\|} \Delta dt d\text{Vol}^{d-1}(x_0), \\ \Delta &= \Phi\Big(\frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)}\Big) - \Phi\Big(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\Big). \end{aligned}$$

Next we show $R_1 = o(s_n + t_n)$. Denote $r_{x_0} = \frac{-a(x_0)t_n}{\|\dot{\eta}(x_0)s_n\|}$. According to (3.16) and (3.17), for a sufficiently small $\epsilon \in (0, \inf_{x_0 \in \mathcal{S}} \|\dot{\eta}(x_0)\|)$ and large n, for all $\boldsymbol{w}_n \in W_{n,\beta}$, $x_0 \in \mathcal{S}$ and $r \in [-\epsilon_n/s_n, \epsilon_n/s_n]$, [27] showed that

$$\Big|\frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})} - [-2\|\dot{\eta}(x_0)\|(r - r_{x_0})]\Big| \le \epsilon^2(|r| + t_n/s_n).$$

In addition, when $|r - r_{x_0}| \le \epsilon t_n/s_n$,

$$\Big|\Phi\Big(\frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})}\Big) - \Phi\Big(-2\|\dot{\eta}(x_0)\|(r - r_{x_0})\Big)\Big| \le 1$$

and when $\epsilon t_n/s_n < |r| < t_n/s_n$,

$$\Big|\Phi\Big(\frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})}\Big) - \Phi\Big(-2\|\dot{\eta}(x_0)\|(r - r_{x_0})\Big)\Big| \le \epsilon^2(|r| + t_n/s_n)\phi(\|\dot{\eta}(x_0)\||r - r_{x_0}|),$$

where ϕ is the density function of standard normal distribution.

Therefore, we have

$$\begin{aligned} |R_{11}| &\le \int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0)\,|\Delta|\, dt d\text{Vol}^{d-1}(x_0) \\ &\le \bar{f}(x_0)s_n \int_{|r - r_{x_0}| \le \epsilon t_n/s_n} dr + \bar{f}(x_0)s_n\epsilon^2 \int_{-\infty}^{\infty} (|r| + t_n/s_n)\phi(\|\dot{\eta}(x_0)\||r - r_{x_0}|)dr \\ &\le \epsilon(t_n + s_n). \end{aligned} \tag{3.25}$$

Similarly,

$$\begin{aligned} &|R_{12}| \\ \le\ & \int_{\mathcal{S}} \int_{-\epsilon_n}^{\epsilon_n} \frac{\dot{\bar{f}}(x_0)^T \dot{\eta}(x_0) t}{\|\dot{\eta}(x_0)\|}\,|\Delta|\, dt d\text{Vol}^{d-1}(x_0) \\ \le\ & \bar{f}(x_0)\epsilon s_n^2 \int_{|r - r_{x_0}| \le \epsilon t_n/s_n} |r| dr + \bar{f}(x_0)s_n^2\epsilon^2 \int_{-\infty}^{\infty} (|r| + t_n/s_n)\phi(\|\dot{\eta}(x_0)\||r - r_{x_0}|)dr \\ \le\ & \epsilon(t_n^2 + s_n^2). \end{aligned}$$

The inequality above, along with with (3.25), leads to $R_1 = o(s_n + t_n)$.

By similar arguments, we have

$$\begin{aligned}
&\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0^t)\Big\{\mathbb{P}^2\Big(S_n(x_0^t) < 1/2\Big) - \mathbb{1}\{t<0\}\Big\}\,dtd\mathrm{Vol}^{d-1}(x_0) \qquad (3.26)\\
=&\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \bar{f}(x_0)\Big\{\Phi^2\Big(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\Big) - \mathbb{1}\{t<0\}\Big\}\,dtd\mathrm{Vol}^{d-1}(x_0)\\
&+\int_{\mathcal{S}}\int_{-\epsilon_n}^{\epsilon_n} \frac{\dot{\bar{f}}(x_0)^T\dot{\eta}(x_0)t}{\|\dot{\eta}(x_0)\|}\Big\{\Phi^2\Big(\frac{-2t\|\dot{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\Big) - \mathbb{1}\{t<0\}\Big\}\,dtd\mathrm{Vol}^{d-1}(x_0)\\
&+o(s_n+t_n).
\end{aligned}$$

Denote $\widetilde{\Delta} = \Phi\Big(- \|\dot{\eta}(x_0)\|u - \frac{2a(x_0)t_n}{s_n}\Big) - \mathbb{1}\{u<0\}$. Finally, after substituting $t = us_n/2$ in (3.24) and (3.26), we have, up to $o(s_n + t_n)$ difference,

$$\begin{aligned}
&\mathrm{CIS(WNN)}/2\\
=\;& \frac{s_n}{2}\int_{\mathcal{S}}\int_{-\infty}^{\infty}\bar{f}(x_0)\widetilde{\Delta}dud\mathrm{Vol}^{d-1}(x_0) + \frac{s_n^2}{4}\int_{\mathcal{S}}\int_{-\infty}^{\infty}\frac{(\dot{\bar{f}}(x_0))^T\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}u\widetilde{\Delta}dtd\mathrm{Vol}^{d-1}(x_0)\\
&-\frac{s_n}{2}\int_{\mathcal{S}}\int_{-\infty}^{\infty}\bar{f}(x_0)\widetilde{\Delta}dud\mathrm{Vol}^{d-1}(x_0) - \frac{s_n^2}{4}\int_{\mathcal{S}}\int_{-\infty}^{\infty}\frac{(\dot{\bar{f}}(x_0))^T\dot{\eta}(x_0)}{\|\dot{\eta}(x_0)\|}u\widetilde{\Delta}dtd\mathrm{Vol}^{d-1}(x_0)\\
=\;& I + II - III - IV.
\end{aligned}$$

According to Lemma 5, we have

$$\begin{aligned}
I - III &= \Big[\int_{\mathcal{S}}\frac{\bar{f}(x_0)}{2\sqrt{\pi}\|\dot{\eta}(x_0)\|}d\mathrm{Vol}^{d-1}(x_0)\Big]s_n = \frac{1}{2}B_3s_n\\
II - IV &= -\Big[\int_{\mathcal{S}}\frac{(\dot{\bar{f}}(x_0))^T\dot{\eta}(x_0)a(x_0)}{2\sqrt{\pi}(\|\dot{\eta}(x_0)\|)^3}d\mathrm{Vol}^{d-1}(x_0)\Big]s_nt_n = \frac{1}{2}B_4s_nt_n.
\end{aligned}$$

Therefore, the desirable result is obtained by noting that $B_4s_nt_n = o(s_n + t_n)$. This concludes the proof of Theorem 3.2.1. ■

3.7.2 Proof of Theorem 3.2.2

We first introduce a Lemma for Proving Theorem 3.2.2.

Lemma 6 *Given $\alpha_i = i^{1+2/d} - (i-1)^{1+2/d}$, we have*

$$(1+\frac{2}{d})(i-1)^{\frac{2}{d}} \le \alpha_i \le (1+\frac{2}{d})i^{\frac{2}{d}}, \tag{3.27}$$

$$\sum_{j=1}^{k} \alpha_j^2 = \frac{(d+2)^2}{d(d+4)} k^{1+4/d} \left\{ 1 + O(\frac{1}{k}) \right\}. \tag{3.28}$$

Proof of Lemma 6: First, (3.27) is a direct result from the following two inequalities.

$$(1-\frac{1}{i})^{2/d} \ge 1 - \frac{2}{(i-1)d} \quad \text{and} \quad (1+\frac{1}{i-1})^{2/d} \ge 1 + \frac{2}{id},$$

where i and d are positive integers. These two inequalities hold because both differences $(1-\frac{1}{i})^{2/d} - (1-\frac{2}{(i-1)d})$ and $(1+\frac{1}{i-1})^{2/d} - (1+\frac{2}{id})$ are decreasing in i and the limit equals 0.

Second, (3.28) is due to (3.27) and Faulhaber's formula $\sum_{i=1}^{k} i^p = \frac{1}{p+1}k^{p+1} + O(k^p)$. According to (3.27), we have

$$(1+\frac{2}{d})^2 \sum_{i=1}^{k} (i-1)^{4/d} \le \sum_{j=1}^{k} \alpha_j^2 \le (1+\frac{2}{d})^2 \sum_{i=1}^{k} i^{4/d}.$$

Due to Faulhaber's formula, $\sum_{i=1}^{k} i^{4/d} = \frac{d}{d+4}k^{1+4/d} + O(k^{4/d})$ and $\sum_{i=1}^{k}(i-1)^{4/d} = \frac{d}{d+4}k^{1+4/d} + O(k^{4/d})$, which leads to (3.28). This concludes the proof of Lemma 6. ■

Proof of Theorem 3.2.2: For any weight $\boldsymbol{w}_n$, the Lagrangian of (3.6) is

$$L(\boldsymbol{w}_n) = \Big(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \Big)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2 + \nu (\sum_{i=1}^{n} w_{ni} - 1).$$

Considering the constraint of nonnegative weights, we denote $k^* = \max\{i : w_{ni}^* > 0\}$. Setting derivative of $L(\boldsymbol{w}_n)$ to be 0, we have

$$\frac{\partial L(\boldsymbol{w}_n)}{\partial w_{ni}} = 2n^{-4/d} \alpha_i \sum_{i=1}^{k^*} \alpha_i w_{ni} + 2\lambda w_{ni} + \nu = 0. \tag{3.29}$$

Summing (3.29) from 1 to k^*, and multiplying (3.29) by α_i and then summing from 1 to k^* yields

$$\begin{aligned} 2n^{-4/d}(k^*)^{1+2/d} \sum_{i=1}^{k^*} \alpha_i w_{ni} + 2\lambda + \nu k^* &= 0 \\ 2n^{-4/d} \sum_{i=1}^{k^*} \alpha_i w_{ni} \sum_{i=1}^{k^*} \alpha_i^2 + 2\lambda \sum_{i=1}^{k^*} \alpha_i w_{ni} + \nu (k^*)^{1+2/d} &= 0. \end{aligned}$$

Therefore, we have

$$w_{ni}^* = \frac{1}{k^*} + \frac{(k^*)^{4/d} - (k^*)^{2/d}\alpha_i}{\sum_{i=1}^{k^*}\alpha_i^2 + \lambda n^{4/d} - (k^*)^{1+4/d}} \tag{3.30}$$

Here w_{ni}^* is decreasing in i since α_i is increasing in i and $\sum_{i=1}^{k^*}\alpha_i^2 > (k^*)^{1+4/d}$ from Lemma 6. Next we solve for k^*. According to the definition of k^*, we only need to find k such that $w_{nk}^* = 0$. Using the results from Lemma 6, solving this equation reduces to solving k^* such that

$$(1+\frac{2}{d})(k^*-1)^{2/d} \le \lambda n^{4/d}(k^*)^{-1-2/d} + \frac{(d+2)^2}{d(d+4)}(k^*)^{2/d}\{1+O(\frac{1}{k^*})\} \le (1+\frac{2}{d})(k^*)^{2/d}.$$

Therefore, for large n, we have

$$k^* = \Big\lfloor \Big\{\frac{d(d+4)}{2(d+2)}\Big\}^{\frac{d}{d+4}} \lambda^{\frac{d}{d+4}} n^{\frac{4}{d+4}} \Big\rfloor.$$

Plugging k^* and the result (3.28) in Supplementary into (3.30) yields the optimal weight. ■

3.7.3 Proof of Theorem 3.3.1

Following the proofs of Lemma 3.1 in [21], we consider the sets $A_j \subset \mathcal{R}$

$$\begin{aligned} A_0 &= \{x \in \mathcal{R} : 0 < |\eta(x) - 1/2| \le \delta\}, \\ A_j &= \{x \in \mathcal{R} : 2^{j-1}\delta < |\eta(x) - 1/2| \le 2^j\delta\} \text{ for } j \ge 1. \end{aligned}$$

For the classification procedure $\Psi(\cdot)$, we have

$$\text{CIS}(\Psi) = \mathbb{E}[\mathbb{1}\{\widehat{\phi}_{n1}(X) \ne \widehat{\phi}_{n2}(X)\}],$$

where $\widehat{\phi}_{n1}$ and $\widehat{\phi}_{n2}$ are classifiers obtained by applying $\Psi(\cdot)$ to two independently and identically distributed samples $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Denote the Bayes classifier ϕ^{Bayes}, we have

$$\begin{aligned} \text{CIS}(\Psi) &= 2\mathbb{E}[\mathbb{1}\{\widehat{\phi}_{n1}(X) = \phi^{\text{Bayes}}(X), \widehat{\phi}_{n2}(X) \ne \phi^{\text{Bayes}}(X)\}] \\ &= 2\mathbb{E}[\{1 - \mathbb{1}\{\widehat{\phi}_{n1}(X) \ne \phi^{\text{Bayes}}(X)\}\}\mathbb{1}\{\widehat{\phi}_{n2}(X) \ne \phi^{\text{Bayes}}(X)\}] \\ &= 2\mathbb{E}_X[\mathbb{P}_{\mathcal{D}_1}(\widehat{\phi}_{n1}(X) \ne \phi^{\text{Bayes}}(X)|X) - \{\mathbb{P}_{\mathcal{D}_1}(\widehat{\phi}_{n1}(X) \ne \phi^{\text{Bayes}}(X)|X)\}^2] \\ &\le 2\mathbb{E}[\mathbb{1}\{\widehat{\phi}_{n1}(X) \ne \phi^{\text{Bayes}}(X)\}], \end{aligned}$$

where the last equality is due to the fact that $\mathcal{D}_1$ and $\mathcal{D}_2$ are independently and identically distributed. For ease of notation, we will denote $\widehat{\phi}_{n1}$ as $\widehat{\phi}_n$ from now on. We further have

$$\begin{aligned}\mathrm{CIS}(\Psi) &\leq 2\sum_{j=0}^{\infty}\mathbb{E}[\mathbb{1}\{\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)\}\mathbb{1}\{X \in A_j\}] \\ &\leq 2\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + 2\sum_{j\geq 1}\mathbb{E}[\mathbb{1}\{\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)\}\mathbb{1}\{X \in A_j\}].\end{aligned}$$

Given the event $\{\widehat{\phi}_n \neq \phi^{\mathrm{Bayes}}\} \cap \{|\eta - 1/2| > 2^{j-1}\delta\}$, we have $|\widehat{\eta}_n - \eta| \geq 2^{j-1}\delta$. Therefore, for any $j \geq 1$, we have

$$\begin{aligned}&\mathbb{E}[\mathbb{1}\{\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)\}\mathbb{1}\{X \in A_j\}] \\ \leq\ & \mathbb{E}[\mathbb{1}\{|\widehat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta\}\mathbb{1}\{2^{j-1}\delta < |\eta(X) - 1/2| \leq 2^j\delta\}] \\ \leq\ & \mathbb{E}_X[\mathbb{P}_{\mathcal{D}}(|\widehat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta|X)\mathbb{1}\{0 < |\eta(X) - 1/2| \leq 2^j\delta\}] \\ \leq\ & C_1\exp(-C_2 a_n(2^{j-1}\delta)^2)\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq 2^j\delta) \\ \leq\ & C_1\exp(-C_2 a_n(2^{j-1}\delta)^2)C_0(2^j\delta)^{\alpha},\end{aligned}$$

where the last inequality is due to margin assumption (3.7) and condition (3.8).

Taking $\delta = a_n^{-1/2}$, we have

$$\mathrm{CIS}(\Psi) \leq C_0 a_n^{-\alpha/2} + C_0 C_1 a_n^{-\alpha/2}\sum_{j\geq 1}2^{\alpha j+1}e^{-C_2 4^{j-1}} \leq C a_n^{-\alpha/2},$$

for some $C > 0$ depending only on α, C_0, C_1 and C_2. ■

3.7.4 Proof of Theorem 3.3.2

Before we prove Theorem 3.3.2, we introduce a useful lemma. In particular, we adapt the Assouad's lemma to prove the lower bound of CIS. This lemma is of independent interest.

We first introduce an important definition called (m, w, b, b')-hypercube that is slightly modified from [63]. We observe independently and identically distributed training samples $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ with $X_i \in \mathcal{X} = \mathcal{R}$ and $Y_i \in \mathcal{Y} = \{1, -1\}$. Let $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ denote the set of all measurable functions mapping from $\mathcal{X}$ into $\mathcal{Y}$. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For the distribution function P, we denote its corresponding probability and expectation as $\mathbb{P}$ and $\mathbb{E}$, respectively.

Definition 4 *[63] Let m be a positive integer, $w \in [0,1]$, $b \in (0,1]$ and $b' \in (0,1]$. Define the (m, w, b, b')-hypercube $\mathcal{H} = \{P_{\vec{\sigma}} : \vec{\sigma} \overset{\Delta}{=} (\sigma_1, \ldots, \sigma_m) \in \{-1, +1\}^m\}$ of probability distributions $P_{\vec{\sigma}}$ of (X, Y) on $\mathcal{Z}$ as follows.*

For any $P_{\vec{\sigma}} \in \mathcal{H}$, the marginal distribution of X does not depend on $\vec{\sigma}$ and satisfies the following conditions. There exists a partition $\mathcal{X}_0, \ldots, \mathcal{X}_m$ of $\mathcal{X}$ satisfying,
(i) for any $j \in \{1, \ldots, m\}$, $\mathbb{P}_X(X \in \mathcal{X}_j) = w$;
(ii) for any $j \in \{0, \ldots, m\}$ and any $X \in \mathcal{X}_j$, we have

$$\mathbb{P}_{\vec{\sigma}}(Y = 1|X) = \frac{1 + \sigma_j \psi(X)}{2}$$

with $\sigma_0 = 1$ and $\psi : \mathcal{X} \to (0,1]$ satisfies for any $j \in \{1, \ldots, m\}$,

$$\begin{aligned} b &\overset{\Delta}{=} \left[1 - \left(\mathbb{E}_{\vec{\sigma}}[\sqrt{1 - \psi^2(X)}|X \in \mathcal{X}_j]\right)^2\right]^{1/2}, \\ b' &\overset{\Delta}{=} \mathbb{E}_{\vec{\sigma}}[\psi(X)|X \in \mathcal{X}_j]. \end{aligned}$$

Lemma 7 *If a collection of probability distributions $\mathcal{P}$ contains a (m, w, b, b')-hypercube, then for any measurable estimator $\widehat{\phi}_n$ obtained by applying Ψ to the training sample $\mathcal{D}$, we have*

$$\sup_{P \in \mathcal{P}} \mathbb{E}^{\otimes n}[\mathbb{P}_X(\widehat{\phi}_n(X) \neq \phi^{Bayes}(X))] \geq \frac{mw}{2}[1 - b\sqrt{nw}]. \tag{3.31}$$

where $\mathbb{E}^{\otimes n}$ is the expectation with respect to $P^{\otimes n}$.

Proof of Lemma 7: Let $\vec{\sigma}_{j,r} \overset{\Delta}{=} (\sigma_1, \ldots, \sigma_{j-1}, r, \sigma_{j+1}, \ldots, \sigma_m)$ for any $r \in \{-1, 0, +1\}$. The distribution $P_{\vec{\sigma}_{j,0}}$ satisfies $\mathbb{P}_{\vec{\sigma}_{j,0}}(dX) = \mathbb{P}_X(dX)$, $\mathbb{P}_{\vec{\sigma}_{j,0}}(Y = 1|X) = 1/2$ for any $X \in \mathcal{X}_j$ and $\mathbb{P}_{\vec{\sigma}_{j,0}}(Y = 1|X) = \mathbb{P}_{\vec{\sigma}}(Y = 1|X)$ otherwise. Let ν denote the distribution of a Rademacher variable σ such that $\nu(\sigma = +1) = \nu(\sigma = -1) = 1/2$. Denote the variational distance between two probability distributions P_1 and P_2 as

$$V(P_1, P_2) = 1 - \int \Big(\frac{dP_1}{dP_0} \wedge \frac{dP_2}{dP_0}\Big) dP_0,$$

where $a \wedge b$ means the minimal of a and b, and P_1 and P_2 are absolutely continuous with respect to some probability distribution P_0.

Lemma 5.1 in [63] showed that the variational distance between two distribution functions $P^{\otimes n}_{-1,1,\ldots,1}$ and $P^{\otimes n}_{1,1,\ldots,1}$ is bounded above. Specifically,

$$V(P^{\otimes n}_{-1,1,\ldots,1}, P^{\otimes n}_{1,1,\ldots,1}) \leq b\sqrt{nw}.$$

Note that $\mathcal{P}$ contains a (m,w,b,b')-hypercube and for $X \in \mathcal{X}_j$, $\phi^{\text{Bayes}}(X) = 1 - 2\mathbb{1}\{\eta(X) < 1/2\} = 1 - 2\mathbb{1}\{(1+\sigma_j\psi(X))/2 < 1/2\} = \sigma_j$ since $\psi(X) \neq 0$. Therefore, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{E}^{\otimes n}[\mathbb{P}_X(\widehat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X))] \\
&\geq \sup_{\vec{\sigma}\in\{-1,+1\}^m} \Big\{\mathbb{E}^{\otimes n}_{\vec{\sigma}} \mathbb{P}_X(\mathbb{1}\{\widehat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)\})\Big\} && (3.32)\\
&\geq \sup_{\vec{\sigma}\in\{-1,+1\}^m} \Bigg\{\mathbb{E}^{\otimes n}_{\vec{\sigma}} \Big(\sum_{j=1}^m \mathbb{P}_X[\mathbb{1}\{\widehat{\phi}_n(X) \neq \sigma_j; X \in \mathcal{X}_j\}]\Big)\Bigg\} \\
&\geq \mathbb{E}_{\nu^{\otimes m}} \sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}} \Big(\mathbb{P}_X[\mathbb{1}\{\widehat{\phi}_n(X) \neq \sigma_j; X \in \mathcal{X}_j\}]\Big) && (3.33)\\
&= \mathbb{E}_{\nu^{\otimes m}} \sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \Big(\frac{dP^{\otimes n}_{\vec{\sigma}}}{dP^{\otimes n}_{\vec{\sigma}_{j,0}}} \mathbb{P}_X[\mathbb{1}\{\widehat{\phi}_n(X) \neq \sigma_j; X \in \mathcal{X}_j\}]\Big) \\
&= \mathbb{E}_{\nu^{\otimes(m-1)}(d\vec{\sigma}_{-j})} \sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \mathbb{E}_{\nu(d\sigma_j)} \Big(\frac{dP^{\otimes n}_{\vec{\sigma}}}{dP^{\otimes n}_{\vec{\sigma}_{j,0}}} \mathbb{P}_X[\mathbb{1}\{\widehat{\phi}_n(X) \neq \sigma_j; X \in \mathcal{X}_j\}]\Big) \\
&\geq \mathbb{E}_{\nu^{\otimes(m-1)}(d\vec{\sigma}_{-j})} \sum_{j=1}^m \mathbb{E}^{\otimes n}_{\vec{\sigma}_{j,0}} \Big[\Big(\frac{dP^{\otimes n}_{\vec{\sigma}_{j,-1}}}{dP^{\otimes n}_{\vec{\sigma}_{j,0}}} \wedge \frac{dP^{\otimes n}_{\vec{\sigma}_{j,+1}}}{dP^{\otimes n}_{\vec{\sigma}_{j,0}}}\Big) \\
&\qquad \mathbb{E}_{\nu(d\sigma_j)}\Big(\mathbb{P}_X[\mathbb{1}\{\widehat{\phi}_n(X) \neq \sigma_j; X \in \mathcal{X}_j\}]\Big)\Big] && (3.34)\\
&= \mathbb{E}_{\nu^{\otimes(m-1)}(d\vec{\sigma}_{-j})} \sum_{j=1}^m \frac{1}{2}\mathbb{P}_X[\mathbb{1}\{X \in \mathcal{X}_j\}]\Big[1 - V(P^{\otimes n}_{\vec{\sigma}_{j,-1}}, P^{\otimes n}_{\vec{\sigma}_{j,+1}})\Big] \\
&= \frac{mw}{2}\Big[1 - V(P^{\otimes n}_{-1,1,\ldots,1}, P^{\otimes n}_{1,1,\ldots,1})\Big] \\
&\geq \frac{mw}{2}[1 - b\sqrt{nw}],
\end{aligned}$$

where (3.32) is due to the assumption that $\mathcal{P}$ contains a (m,w,b,b')-hypercube, (3.33) is because the supremum over the m Rademacher variables is no less than the corresponding expected value. Finally, the inequality (3.34) is due to $dP_{\vec{\sigma}}^{\otimes n} \geq \{dP_{\vec{\sigma}_{j,+1}}^{\otimes n} \wedge dP_{\vec{\sigma}_{j,-1}}^{\otimes n}\}$ and the latter is not random with respect to $\nu(d\sigma_j)$. This ends the proof of Lemma 7. ■

Proof of Theorem 3.3.2: According to the proof of Theorem 3.3.1, we have

$$\mathrm{CIS}(\Psi) = 2\left\{\mathbb{E}_X[\mathbb{P}_{\mathcal{D}}(\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)|X)] - \mathbb{E}_X[\{\mathbb{P}_{\mathcal{D}}(\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)|X)\}^2]\right\}.$$

[21] showed that when $\alpha\gamma \leq d$, the set of probability distribution $\mathcal{P}_{\alpha,\gamma}$ contains a (m, w, b, b')-hypercube with $w = C_3 q^{-d}$, $m = \lfloor C_4 q^{d-\alpha\gamma} \rfloor$, $b = b' = C_5 q^{-\gamma}$ and $q = \lfloor C_6 n^{1/(2\gamma+d)} \rfloor$, with some constants $C_i \geq 0$ for $i = 3, \ldots, 6$ and $C_6 \leq 1$. Therefore, Lemma 7 implies that the first part is bound, that is,

$$\begin{aligned}
& \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathbb{E}_X[\mathbb{P}_{\mathcal{D}}(\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)|X)] \\
= & \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathbb{E}_{\mathcal{D}}[\mathbb{P}_X(\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X))] \\
\geq & \frac{mw}{2}[1 - b\sqrt{nw}] \\
= & (1 - C_6) C_3 C_4 C_5 n^{-\alpha\gamma/(2\gamma+d)}.
\end{aligned}$$

To bound the second part, we again consider the sets A_j defined in Appendix 3.7.3. On the event $\{\widehat{\phi}_n \neq \phi^{\mathrm{Bayes}}\} \cap \{|\eta - 1/2| > 2^{j-1}\delta\}$, we have $|\widehat{\eta}_n - \eta| \geq 2^{j-1}\delta$. Letting $\delta = a_n^{-1/2}$ leads to

$$\begin{aligned}
& \mathbb{E}_X[\{\mathbb{P}_{\mathcal{D}}(\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)|X)\}^2] \\
= & \sum_{j=0}^{\infty} \mathbb{E}_X[\{\mathbb{P}_{\mathcal{D}}(\{\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)\}|X)\}^2 \mathbb{1}\{X \in A_j\}] \\
\leq & \mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + \sum_{j=1}^{\infty} \mathbb{E}_X[\{\mathbb{P}_{\mathcal{D}}(\{\widehat{\phi}_n(X) \neq \phi^{\mathrm{Bayes}}(X)\}|X)\}^2 \mathbb{1}\{X \in A_j\}] \\
\leq & \mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + \sum_{j \geq 1} C_1 e^{-2C_2 4^{j-1}} \mathbb{P}_X(0 < |\eta(x) - 1/2| \leq 2^j \delta) \\
\leq & C_0 a_n^{-\alpha/2} + C_0 C_1 a_n^{-\alpha/2} \sum_{j \geq 1} 2^{\alpha j} e^{-2C_2 4^{j-1}} \\
\leq & C_7 a_n^{-\alpha/2},
\end{aligned}$$

for some positive constant C_7 depending only on α, C_0, C_1, C_2. When $a_n = n^{2\gamma/(2\gamma+d)}$, we have

$$\mathbb{E}_X[(\mathbb{P}_{\mathcal{D}}(\widehat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X))^2] \leq C_7 n^{-\alpha\gamma/(2\gamma+d)}.$$

By properly choosing constants C_i such that $(1-C_6)C_3C_4C_5 - C_7 > 0$, we have

$$\text{CIS}(\Psi) \geq 2[(1-C_6)C_3C_4C_5 - C_7]n^{-\alpha\gamma/(2\gamma+d)} \geq C' n^{-\alpha\gamma/(2\gamma+d)},$$

for a constant $C' > 0$. This concludes the proof of Theorem 3.3.2. ■

3.7.5 Proof of Theorem 3.3.3

According to our Theorem 3.3.1 and the proof of Theorem 1 in the supplementary of [27], it is sufficient to show that for any $\alpha \geq 0$ and $\gamma \in (0, 2]$, there exist positive constants C_1, C_2 such that for all $\delta > 0$, $n \geq 1$ and $\bar{P}$-almost all x,

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathbb{P}_{\mathcal{D}}\Big(|S_n^*(x) - \eta(x)| \geq \delta\Big) \leq C_1 \exp(-C_2 n^{2\gamma/(2\gamma+d)}\delta^2). \tag{3.35}$$

where $S_n^*(x) = \sum_{i=1}^n w_{ni}^* \mathbb{1}\{Y_{(i)} = 1\}$ with the optimal weight w_{ni}^* defined in Theorem 3.2.2 and $k^* \asymp n^{2\gamma/(2\gamma+d)}$.

According to Lemma 6, we have

$$\sum_{i=1}^{k^*} (w_{ni}^*)^2 = \frac{2(d+2)}{(d+4)k^*}\{1 + O((k^*)^{-1})\} \leq C_8 n^{-2\gamma/(2\gamma+d)},$$

for some constant $C_8 > 0$.

Denote $\mu_n^*(x) = \mathbb{E}\{S_n^*(x)\}$. According to the proof of Theorem 1 in the supplement of [27], there exist $C_9, C_{10} > 0$ such that for all $P \in \mathcal{P}_{\alpha,\gamma}$ and $x \in \mathcal{R}$,

$$\begin{aligned} |\mu_n^*(x) - \eta(x)| &\leq \left|\sum_{i=1}^n w_{ni}^* \mathbb{E}\{\eta(X_{(i)}) - \eta_x(X_{(i)})\}\right| + \left|\sum_{i=1}^n w_{ni}^* \mathbb{E}\{\eta_x(X_{(i)})\} - \eta(x)\right| \\ &\leq L\sum_{i=1}^n w_{ni}^* \mathbb{E}\{\|X_{(i)} - x\|^\gamma\} + \left|\sum_{i=1}^n w_{ni}^* \mathbb{E}\{\eta_x(X_{(i)})\} - \eta(x)\right| \\ &\leq C_9 \sum_{i=1}^n w_{ni}^* \Big(\frac{i}{n}\Big)^{\gamma/d} \\ &\leq C_{10} n^{-\gamma/(2\gamma+d)}. \end{aligned} \tag{3.36}$$

The Hoeffding's inequality says that if $Z_1, \ldots, Z_n$ are independent and $Z_i \in [a_i, b_i]$ almost surely, then we have

$$\mathbb{P}\left(\Big|\sum_{i=1}^{n} Z_i - \mathbb{E}\Big[\sum_{i=1}^{n} Z_i\Big]\Big| \geq t\right) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Let $Z_i = w_{ni}^* \mathbb{1}\{Y_{(i)} = 1\}$ with $a_i = 0$ and $b_i = w_{ni}^*$. According to (3.36), we have that for $\delta \geq 2C_{10} n^{-\gamma/(2\gamma+d)}$ and for $\bar{P}$-almost all x,

$$\begin{aligned} \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathbb{P}_{\mathcal{D}}\Big(|S_n^*(x) - \eta(x)| \geq \delta\Big) &\leq \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \mathbb{P}_{\mathcal{D}}\Big(|S_n^*(x) - \mu_n^*(x)| \geq \delta/2\Big) \\ &\leq 2\exp\{-n^{2\gamma/(2\gamma+d)}\delta^2/(2C_8)\}, \end{aligned}$$

which implies (3.35) directly. ■

3.7.6 Proof of Corollary 3

According to Theorems 3.3.1 and 3.3.2, we have, for any $\gamma \in (0, 2]$,

$$\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS(SNN)} \asymp n^{-\alpha\gamma/(2\gamma+d)}.$$

Therefore, when $\lambda \neq B_1/B_2$, we have

$$\begin{aligned} &\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \Big\{\text{CIS(SNN)} - \text{CIS(OWNN)}\Big\} \\ \geq\; &\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS(SNN)} - \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS(OWNN)} \\ \geq\; &C_{11} n^{-\alpha\gamma/(2\gamma+d)}. \end{aligned}$$

for some constant $C_{11} > 0$. Here $C_{11} = 0$ if and only if $\lambda = B_1/B_2$. On the other hand, we have

$$\begin{aligned} &\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \Big\{\text{CIS(SNN)} - \text{CIS(OWNN)}\Big\} \\ \leq\; &\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS(SNN)} + \sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS(OWNN)} \\ \leq\; &C_{12} n^{-\alpha\gamma/(2\gamma+d)}, \end{aligned}$$

for some constant $C_{12} > 0$.

Furthermore, according to Theorem 3.3.3, we have

$$\sup_{P\in\mathcal{P}_{\alpha,\gamma}} \text{Regret(SNN)} \asymp n^{-\gamma(1+\alpha)/(2\gamma+d)}.$$

Similar to above arguments in CIS, we have

$$\sup_{P\in\mathcal{P}_{\alpha,\gamma}} \Big\{\text{Regret(SNN)} - \text{Regret(OWNN)}\Big\} \asymp n^{-\gamma(1+\alpha)/(2\gamma+d)}.$$

This concludes the proof of Corollary 3. ■

3.7.7 Proof of Corollaries 4 and 5

For the OWNN classifier, the optimal k^{**} is a function of k^{opt} of k-nearest neighbor classifier [27]. Specifically,

$$k^{**} = \Big\lfloor \Big\{\frac{2(d+4)}{d+2}\Big\}^{\frac{d}{d+4}} k^{\text{opt}} \Big\rfloor.$$

According to Theorem 3.2.2 and Lemma 6, we have

$$\sum_{i=1}^{k^*} (w_{ni}^*)^2 = \frac{2(d+2)}{(d+4)k^*}\{1+O((k^*)^{-1})\}.$$

Therefore,

$$\frac{\text{CIS(OWNN)}}{\text{CIS}(k\text{NN})} \to 2^{2/(d+4)}\Big(\frac{d+2}{d+4}\Big)^{(d+2)/(d+4)}.$$

Furthermore, for large n,

$$\frac{\text{CIS(SNN)}}{\text{CIS(OWNN)}} = \frac{B_3\Big(\sum_{i=1}^{k^*} w_{ni}^{*2}\Big)^{1/2}}{B_3\Big(\sum_{i=1}^{k^{**}} w_{ni}^{**2}\Big)^{1/2}} = \Big\{\frac{B_1}{\lambda B_2}\Big\}^{d/(2(d+4))}.$$

The rest limit expressions in Corollaries 4 and 5 can be shown in similar manners. ■

3.7.8 Calculation of B_1 in Section 3.6.2

According to the definition,

$$B_1 = \int_{\mathcal{S}} \frac{\bar{f}(x_0)}{4\|\dot{\eta}(x_0)\|} d\text{Vol}^{d-1}(x_0).$$

When $f_1 = N(0_d, \mathbb{I}_d)$ and $f_2 = N(\mu, \mathbb{I}_d)$ with the prior probability $\pi_1 = 1/3$, we have

$$\bar{f}(x_0) = \pi_1 f_1 + (1 - \pi_1) f_2 = 2(2\pi)^{-2/d} \exp\{-x_0^T x_0/2\}/3,$$

and

$$\eta(x) = \frac{\pi_1 f_1}{\pi_1 f_1 + (1 - \pi_1) f_2} = \Big(1 + 2\exp\{\mu^T x - \mu^T \mu/2\}\Big)^{-1}.$$

Hence, the decision boundary is

$$\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\} = \{x \in \mathcal{R} : 1_d^T x = (\mu d)/2 - (\ln 2)/\mu\},$$

where 1_d is a d-dimensional vector of all elements 1.

Therefore, for $x_0 \in \mathcal{S}$, we have $\dot{\eta}(x_0) = -\mu/4$ and hence

$$\begin{aligned} B_1 &= \frac{2}{3\mu(2\pi)^{d/2}\sqrt{d}} \int_{\mathcal{S}} \exp\{-x_0^T x_0/2\} d\text{Vol}^{d-1}(x_0). \\ &= \frac{\sqrt{2\pi}}{3\pi\mu d} \exp\left\{-\frac{(\mu d/2 - \ln 2/\mu)^2}{2d}\right\}. \end{aligned}$$

4. SUMMARY

Stability is an important and desirable property of a statistical procedure. It provides a foundation for the reproducibility, and reflects the credibility of those who use the procedure. To our best knowledge, our work is the first to propose a measure of classification instability to calibrate this quantity. In this thesis, we first introduce a decision boundary instability (DBI). This allows us to propose a two-stage classifier selection procedure based on GE and DBI. It selects the classifier with the most stable decision boundary among those classifiers with relatively small estimated GEs. We then propose a novel SNN classification procedure to improve the nearest neighbor classifier. It enjoys increased classification stability with almost unchanged classification accuracy. Our SNN is shown to achieve the minimax optimal convergence rate in regret and a sharp convergence rate in CIS, which is also established in this article. Extensive experiments illustrate that SNN attains a significant improvement of CIS over existing nearest neighbor classifiers, and sometimes even improves the accuracy.

For simplicity, we focus on the binary classification in this article. The concept of DBI or CIS is quite general, and its extension to a broader framework, e.g., multicategory classification [28,64–67] or high-dimensional classification [68], is an interesting topic to pursue in the future.

Stability for the high-dimensional, low-sample size data is another important topic. Our classification stability can be used as a criterion for tuning parameter selection in high-dimensional classification. There exists work in the literature which uses variable selection stability to select tuning parameter [14]. Classification stability and variable selection stability complement each other to provide a description of the reliability of a statistical procedure.

Finally, in analyzing a big data set, a popular scheme is divide-and-conquer. It is an interesting research question how to divide the data and choose the parameter wisely to ensure the optimal stability of the combined classifier.

REFERENCES

REFERENCES

[1] V. Stodden, F. Leisch, and R. Peng. *Implementing reproducible research.* CRC Press, 2014.

[2] B. Yu. Stability. *Bernoulli*, 19:1484–1500, 2013.

[3] P. Kraft, E. Zeggini, and J.P.A. Ioannidis. Replication in genome-wide association studies. *Statistical Science*, 24:561–573, 2009.

[4] R.D. Peng. Reproducible research and biostatistics. *Biostatistics*, 10:405–408, 2009.

[5] D.L. Donoho, A. Maleki, M. Shahram, I.U. Rahman, and V. Stodden. Reproducible research in computational harmonic analysis. *IEEE Computing in Science and Engineering*, 11:8–18, 2009.

[6] J.P.A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:696–701, 2005.

[7] A. Gershoff, A. Mukherjee, and A. Mukhopadhyay. Consumer acceptance of online agent advice: Extremity and positivity effects. *Journal of Consumer Psychology*, 13:161–170, 2003.

[8] L. Van Swol and J. Sniezek. Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, 44:443–461, 2005.

[9] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

[10] J. Wang. Consistent selection of the number of clusters via cross validation. *Biometrika*, 97:893–904, 2010.

[11] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:414–473, 2010.

[12] R. Shah and R. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society, Series B*, 75:55–80, 2013.

[13] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection for high-dim graphical models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[14] W. Sun, J. Wang, and Y. Fang. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14:3419–3440, 2013.

[15] L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383, 1996.

[16] P. Bühlmann and B. Yu. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2002.

[17] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[18] A. Elisseeff, T. Evgeniou, and M. Pontil. Stability for randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

[19] T. Lim, W.Y. Loh, and Y.S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40:203–229, 2000.

[20] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of American Statistical Association*, 102:974–983, 2007.

[21] J. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35:608–633, 2007.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag: New York, 2009.

[23] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–279, 1995.

[24] Y. Freund and R. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[25] J. S. Marron, M. Todd, and J. Ahn. Distance weighted discrimination. *Journal of American Statistical Association*, 102:1267–1271, 2007.

[26] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11:111–130, 2010.

[27] R. Samworth. Optimal weighted nearest neighbor classifiers. *Annals of Statistics*, 40:2733–2763, 2012.

[28] Y. Liu and M. Yuan. Reinforced multicategory support vector machines. *Journal of Computational and Graphical Statistics*, 20:901–919, 2011.

[29] V. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, New York, 1998.

[30] I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

[31] X. Qiao and Y. Liu. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics*, 65:159–168, 2009.

[32] J. Wang. Boosting the generalized margin in cost-sensitive multiclass classification. *Journal of Computational and Graphical Statistics*, 22:178–192, 2013.

[33] G. Valentini and T. Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, 5:725–775, 2004.

[34] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–134, 2004.

[35] J. Wang and X. Shen. Estimation of generalization error: Random and fixed inputs. *Statistica Sinica*, 16:569–588, 2006.

[36] B. Jiang, X. Zhang, and T. Cai. Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, 9:521–540, 2008.

[37] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199, 1991.

[38] G. Rocha, X. Wang, and B. Yu. Asymptotic distribution and sparsistency for $l1$ penalized parametric m-estimators, with applications to linear svm and logistic regression. *Technical Report*, 2009.

[39] Y. Park and L.J. Wei. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90:717–723, 2003.

[40] J. Hoffmann-Jorgensen. Stochastic processes on polish spaces. *Technical Report*, 1984.

[41] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[42] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[43] Y. Liu, H. Zhang, and Y. Wu. Hard or soft classification? large-margin unified machines. *Journal of American Statistical Association*, 106:166–177, 2011.

[44] G. Wahba. Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences*, 99:16524–16530, 2002.

[45] J. Wang, X. Shen, and Y. Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95:149–167, 2008.

[46] K. Bache and M. Lichman. Uci machine learning repository. *Irvine, CA: University of California*, 2013.

[47] W.H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences*, pages 9193–9196, 1990.

[48] R. Hable. Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117, 2012.

[49] R. Hable. Asymptotic confidence sets for general nonparametric regression and classification by regularized kernel methods. *Technical Report*, 2014.

[50] E. Fix and J. L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. *Project 21-49-004, Report No.4, Randolph Field, Texas*, 6, 2005.

[51] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[52] L. Devroye and T. J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Annals of Statistics*, 5:536–540, 1977.

[53] R. R. Snapp and S. S. Venkatesh. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.

[54] L. Györfi. The rate of convergence of k-nn regression estimates and classification rules. *IEEE Transactions on Information Theory*, 27:362–364, 1981.

[55] L. Devroye, L. Györfi, A. Krzyak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385, 1994.

[56] R. R. Snapp and S. S. Venkatesh. Asymptotic expansion of the k nearest neighbor risk. *Annals of Statistics*, 26:850–878, 1998.

[57] G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11:687–712, 2010.

[58] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[59] P. Hall, B. Park, and R. Samworth. Choice of neighbor order in nearest neighbor classification. *Annals of Statistics*, 36:2135–2152, 2008.

[60] P. Hall and K. Kang. Bandwidth choice for nonparametric classification. *Annals of Statistics*, 33:284–306, 2005.

[61] L. Devroye, L. Györfi, and G. Lugosi. *Tubes.* Progress in Mathematics, Birkhäuser, Basel, 2004.

[62] S. Bjerve. Error bounds for linear combinations of order statistics. *Annals of Statistics*, 5:357–369, 1977.

[63] J. Audibert. Classification under polynomial entropy and margin assumptions and randomized estimators. *Preprint 905, Laboratoire de Probabilites et Modeles Aleatoires, Univ. Paris VI and VII*, 2004.

[64] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of American Statistical Association*, 99:67–81, 2004.

[65] Y. Liu and X. Shen. Multicategory psi-learning. *Journal of American Statistical Association*, 101:500–509, 2006.

[66] X. Shen and L. Wang. Generalization error for multi-class margin classification. *Electronic Journal of Statistics*, 1:307–330, 2007.

[67] C. Zhang and Y. Liu. Multicategory large-margin unified machines. *Journal of Machine Learning Research*, 14:1349–1386, 2013.

[68] J. Fan, Y. Feng, and X. Tong. A road to classification in high dimensional space. *Journal of the Royal Statistical Society, Series B*, 74:745–771, 2012.

VITA

VITA

Wei Sun was born in Jiangsu, China in 1988. He received a bachelor's degree in Statistics at Nankai University, China in 2009 and a master's degree in Statistics at University of Illinois at Chicago in 2011. Then, he joined the PhD program of Statistics at Purdue University with research supported by Lynn fellowship. He earned a joint master's degree in Statistics and Computer Science in 2014 and a doctoral degree in Statistics in 2015. Under supervision of Prof. Guang Cheng, his PhD thesis addressed the stability of machine learning algorithms. Aside from this, during his PhD study, he has worked on exciting projects on sparse tensor decompositions, sparse tensor regressions, statistical and computational tradeoffs of non-convex optimizations, and high-dimensional clustering with applications to personalized medicine and computational advertising.