NEAREST NEIGHBOR CLASSIFIER WITH OPTIMAL STABILITY

BY WEI SUN*, XINGYE QIAO† AND GUANG CHENG‡

Purdue University and Binghamton University

MAY 27, 2014

Stability has been of a great concern in statistics: similar statistical conclusions should be drawn based on different data sampled from the same population. In this article, we introduce a general measure of classification instability (CIS) to capture the sampling variability of the predictions made by a classification procedure. The minimax rate of CIS is established for general plug-in classifiers. As a concrete example, we consider the stability of the nearest neighbor classifier. In particular, we derive an asymptotically equivalent form for the CIS of a weighted nearest neighbor classifier. This allows us to develop a novel stabilized nearest neighbor classifier which well balances the trade-off between classification accuracy and stability. The resulting classification procedure is shown to possess the minimax optimal rates in both excess risk and CIS. Extensive experiments demonstrate a significant improvement of CIS over existing nearest neighbor classifiers at an ignorable cost of classification accuracy.

1. Introduction. In the scientific community, stability plays an important role since scientific conclusions should be reproducible and stable with respect to small perturbation of data (Yu, 2013). Stability has received much attention in many areas of statistics. For example, in clustering problems, Ben-Hur et al. (2002) introduced the clustering instability to assess the quality of a clustering algorithm; Wang (2010) used the clustering instability as a criterion to select the number of clusters. In high dimensional regression, Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) proposed stability selection procedures for variable selection; Liu et al. (2010) applied stability for tuning parameter selection. In addition to the stability in these specific contexts, general definitions of stability have been introduced as tools for various other purposes. For example, stability was employed for model selection (Breiman, 1996), for analyzing the effect of bagging (Bühlmann and Yu, 2002), and for deriving generalization error bound

*PhD student. Research Sponsored by NSF CAREER Award DMS-1151692.
†Assistant Professor. Research Sponsored by Simons Foundation #246649.
‡Corresponding Author. Associate Professor. Research Sponsored by NSF CAREER Award DMS-1151692 and Simons Foundation #305266.

AMS 2000 subject classifications: Primary 62H30, 62G20; secondary 68T10, 68Q32

Keywords and phrases: Classification, minimax rate, optimality, stability
(Bousquet and Elisseeff, 2002; Elisseeff et al., 2005). While successes of stability have been reported in the aforementioned works, to the best of our knowledge, there has been little systematic and rigorous theoretical study of stability in the classification context.

Our first contribution is to rigorously define a general measure of stability for a classification procedure. It characterizes the sampling variability of the yielded predictions, and is thus called the classification instability (CIS) throughout the paper. Moreover, we have established the minimax optimal rate of CIS, which is slower than but approaching \( n^{-1} \), for general plug-in classifiers by adapting the theoretical framework of Audibert and Tsybakov (2007).

Our second contribution is to elaborate on CIS for the nearest neighbor classifier (Fix and Hodges, 1951; Cover and Hart, 1967), one of the most popular nonparametric classification methods. In the literature, extensive research have been done to theoretically justify various nearest neighbor classifiers based on their risks (Devroye and Wagner, 1977; Stone, 1977; Györfi, 1981; Devroye et al., 1994; Snapp and Venkatesh, 1998; Biau et al., 2010). For a comprehensive study of the nearest neighbor classifier, we refer the readers to Devroye et al. (1996). Recently, Samworth (2012) has proposed an optimal weighted nearest neighbor (ownn) classifier by minimizing the asymptotic excess risk (regret) of a weighted nearest neighbor (wnn) classifier. As seen above, most existing methods only focus on the regret without paying much attention to the classification stability.

An interesting result we find is that the CIS of a wnn classifier is asymptotically proportional to the Euclidean norm of the weight vector. This rather concise form is crucial in our theoretical analysis. To simultaneously control regret and CIS, we propose a novel method called the stabilized nearest neighbor (snn) classifier. Specifically, we construct the snn procedure by minimizing the CIS of a wnn classifier over an acceptable region of the weight where the regret is small. A tuning parameter is introduced to balance the trade-off between regret and CIS. This new methodology encompasses the ownn classifier as a special case. In theory, we show that the proposed snn procedure achieves the minimax optimal rates in both regret and CIS.

Our last contribution is to offer a comprehensive comparison between snn and the existing nearest neighbor classifiers, through which new theoretical insights are obtained. We theoretically verify that the CIS of our snn procedure is dramatically smaller than those of the others. Extensive experiments further illustrate that the snn classifier has a significant improvement, up to 60%, in CIS with an ignorable loss, up to 2%, in its classification accuracy. This suggests that slight sacrifice of accuracy could greatly improve stability. The above phenomenon can be partially explained by our theoretical finding (Corollary 1) 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. As a by-product, we show that the ownn classifier is more stable than the \( k \)-nearest neighbor (knn) and the bagged nearest neighbor (bnn) classifiers: the CIS ratios of the former to the latter two classifiers are strictly less than 1 and depend asymptotically only on the dimension of the feature vector.
For simplicity, we focus on binary classification in this paper. The generalization to multiclassification is conceptually feasible. The rest of the article is organized as follows. Section 2 defines the classification instability for general classifiers and establishes its minimax rate. Section 3 proposes a novel snn classifier which achieves the minimax optimality in terms of both regret and CIS. Section 4 presents a more refined theoretical comparison of regret and CIS between the proposed classifier and the existing nearest neighbor classifiers. Section 5 illustrates the empirical performance of various nearest neighbor classifiers, and demonstrates that our method has a significant improvement of stability. The appendix is devoted to technical proofs.

2. Classification Instability. Let \((X, Y)\) be a random couple taking values in \(\mathcal{R} \otimes \{1, 2\}\) with a joint distribution \(P\), where \(\mathcal{R} \subset \mathbb{R}^d\). We regard \(X\) as a vector of features corresponding to an object and \(Y\) as the label indicating that the object belongs to one of two classes. Denote the prior class probability for \(Y\) as \(\pi_0 = P(Y = 1) \in (0, 1)\), where \(P\) is the probability with respect to \(P\), and the distribution of \(X\) given \(Y = r\) as \(P_r\) with \(r = 1, 2\). Hence, the marginal distribution of \(X\) can be written as \(\bar{P} = \pi_0 P_1 + (1 - \pi_0) P_2\). For a Borel measurable set \(\mathcal{R}\) and a classifier \(\phi : \mathcal{R} \mapsto \{1, 2\}\), the classification risk of \(\phi\) is defined as \(R(\phi) = P(\phi(X) \neq Y)\). It is well known that the Bayes rule, denoted as \(\phi^{\text{Bayes}}\), minimizes the above classification risk. Specifically, \(\phi^{\text{Bayes}}(x) = 1 + \mathbb{1}\{\eta(x) < 1/2\}\), where \(\eta(x) = P(Y = 1|X = x)\) and \(\mathbb{1}\{\cdot\}\) is the indicator function. In practice, a classification procedure \(\Psi\) is applied to a training data set \(D = \{(X_i, Y_i), i = 1, \ldots, n\}\) to produce a classifier \(\hat{\phi}_n : \mathcal{R} \mapsto \{1, 2\}\). We define its regret as \(\mathbb{E}_{(X, Y), D}[R(\hat{\phi}_n)] - R(\phi^{\text{Bayes}})\), where \(\mathbb{E}_{(X, Y), D}\) denotes the expectation with respect to the distribution of \((X, Y)\) and \(D\).

In general, a classification procedure is reliable if the produced classifiers, which are trained from different samples with the same underlying distribution, yield the same prediction for an input with high probability.

2.1. General Definition. Our first step in formalizing the classification instability is to define the distance between two generic classifiers \(\phi_1\) and \(\phi_2\).

**Definition 1.** (Distance of Classifiers) Define the distance between two classifiers \(\phi_1\) and \(\phi_2\) as \(d(\phi_1, \phi_2) = P_X(\phi_1(X) \neq \phi_2(X))\), where \(P_X\) is the probability with respect to the marginal distribution of \(X\).

The distance defined above measures the level of disagreement between two classifiers. It is a valid distance measure since it is nonnegative, symmetric and satisfies the triangle inequality.

We next define the classification instability that measures the prediction variation of classifiers trained based on different data from the same population. Throughout the paper, we denote \(D_1\) and \(D_2\) as two independent and identically distributed copies of the training data \(D\).
Definition 2. (Classification Instability) Define the classification instability of a classification procedure $\Psi$ as

\begin{equation}
\text{CIS}(\Psi) = \mathbb{E}_{D_1,D_2} \left[ d(\hat{\phi}_{n1}, \hat{\phi}_{n2}) \right]
\end{equation}

where $\hat{\phi}_{n1} = \Psi(D_1)$ and $\hat{\phi}_{n2} = \Psi(D_2)$ are the classifiers obtained by applying the classification procedure $\Psi$ to data $D_1$ and $D_2$.

For ease of notation, we have suppressed the dependence of CIS($\Psi$) on the sample size $n$ of $D$. By definition, $0 \leq \text{CIS}(\Psi) \leq 1$, and small CIS($\Psi$) represents a stable classification procedure $\Psi$.

We next use the knn classification procedure to illustrate the interplay between classification accuracy and classification stability. Given any $x \in \mathcal{R}$, the knn classifier $\hat{\phi}_{n}^{\text{knn}}(x) = 1$ if and only if $k^{-1} \sum_{i=1}^{k} 1\{Y(i) = 1\} \geq 1/2$, where $Y(i)$ is the label of the $i$th nearest neighbor of $x$ in the training sample. Consider a simple Gaussian setup: $X|Y = 1 \sim N(0_2, I_2)$ and $X|Y = 2 \sim N(1_2, I_2)$, where $\mu_p$ is a $p$-dimensional vector of all $\mu \in \mathbb{R}$, and $I_p$ is the $p$ by $p$ identity matrix. Assume $n = 300$ and $\pi_0 = 1/3$. Figure 1 plots the regret versus CIS for knn, calculated according to Proposition 1 and Theorem 3 in Section 3, with $k$ ranging from 1 to 25. As $k$ increases, the classifier becomes more and more stable due to the aggregation effect, while the regret first decreases and then increases. The green cross in the plot, which minimizes the sum of the squared regret and CIS, can be viewed as an instance of good balance between classification accuracy and stability. In view of the knn classifier with the minimal regret, marked as the red triangle in the plot, one may have the impression that a slight sacrifice of risk can lead to a big improvement of stability. This will be confirmed by our theorems and extensive experiments in the sequel.

![Fig 1. Regret and CIS of the k-nearest neighbor classifier. The x-axis is regret and the y-axis is CIS. Each dot represents one choice of $k \in [1, 25]$. The red triangle denotes the classifier obtaining minimal regret; the green cross is the projection of the origin to the path, which minimizes the sum of the squared regret and CIS.](image-url)
2.2. Minimax Properties of General Plug-in Classifiers. We introduce a theoretical framework, motivated by Audibert and Tsybakov (2007), to study the minimax properties of CIS for plug-in classifiers. A plug-in classifier estimates the regression function \( \eta(x) \) by \( \tilde{\eta}_n(x) \), and then plugs it into the form of the Bayes classifier, that is, \( \tilde{\phi}_n(x) = 1 + \mathbb{1}\{\tilde{\eta}_n(x) < 1/2\} \). In this subsection, we first state an upper bound of CIS’s convergence rate and then establish the minimax lower bound showing that the obtained rate cannot be improved.

The following margin condition (Tsybakov, 2004) is assumed for deriving the upper bound, 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 \( \epsilon > 0 \),

\[
\mathbb{P}(0 < |\eta(X) - 1/2| \leq \epsilon) \leq C_0 \epsilon^{\alpha}.
\]  

The second condition is on the smoothness of \( \eta(x) \). Specifically, we assume that \( \eta \) belongs to a H"older class of functions \( \Sigma(\gamma, L, \mathbb{R}^d) \) (for some fixed \( L, \gamma > 0 \)) containing the functions \( g : \mathbb{R}^d \rightarrow \mathbb{R} \) 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 \( \gamma \) and \( g_x \) is the Taylor polynomial series of degree \( \lfloor \gamma \rfloor \) at \( x \).

Our last condition says the marginal distribution \( \bar{P} \) satisfies the strong density assumption if

- for a compact set \( R \subset \mathbb{R}^d \) and constants \( c_0, r_0 > 0 \), \( \bar{P} \) is supported on a compact \( (c_0, r_0) \)-regular set \( A \subset R \) 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 \( \nu_d \) 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 \( r > 0 \);

- for all \( x \in A \), the Lebesgue density \( \bar{f} \) of \( P \) satisfies \( \bar{f}_{\min} \leq \bar{f} \leq \bar{f}_{\max} \) 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 deduce the rate of convergence of CIS by assuming the exponential convergence rate of the corresponding regression function estimator.

Theorem 1. Let \( \mathcal{P} \) be a set of probability distributions supported on \( R \otimes \{1, 2\} \) such that for some constants \( C_1, C_2 > 0 \), some positive sequence \( a_n \to \infty \), and almost all \( x \) with respect to \( \bar{P} \),

\[
\sup_{P \in \mathcal{P}} \mathbb{P}_P\left( |\tilde{\eta}_n(x) - \eta(x)| \geq \delta \right) \leq C_1 \exp(-C_2 a_n \delta^2)
\]

holds for any \( n > 1 \) and \( \delta > 0 \), where \( \mathbb{P}_P \) is the probability with respect to \( P^{\otimes n} \). Furthermore, if all the distributions \( P \in \mathcal{P} \) satisfy the margin condition, then we have

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

for any \( n > 1 \) and some constant \( C > 0 \) depending only on \( \alpha, C_0, C_1 \) and \( C_2 \).
It is worth noting that (2.3) holds for various types of estimators. For example, Theorem 3.2 in Audibert and Tsybakov (2007) showed that the local polynomial estimator fulfils (2.3) with \( a_n = n^{2\gamma/(2\gamma+d)} \) when the bandwidth is of the order \( n^{-1/(2\gamma+d)} \) and \( \eta \in \Sigma(\gamma, L, \mathbb{R}^d) \). In addition, our Theorem 5 in Section 3.3 implies that (2.3) holds for the newly proposed 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 show the minimax lower bound of CIS in Theorem 2. As will be seen, this lower bound implies that the obtained rate of CIS, i.e., \( n^{-\alpha\gamma/(2\gamma+d)} \), cannot be further improved.

**Theorem 2. (Minimax Lower Bound)** Let \( \mathcal{P}_{\alpha,\gamma} \) be a set of probability distributions supported on \( \mathcal{R} \otimes \{1, 2\} \) such that for any \( P \in \mathcal{P}_{\alpha,\gamma} \), \( P \) satisfies the margin condition (2.2), the regression function \( \eta(x) \) belongs to the H"older 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 (2.3) with \( a_n = n^{2\gamma/(2\gamma+d)} \) and \( \alpha\gamma \leq d \). We have

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

for any \( n > 1 \) and some constant \( C' > 0 \) independent of \( n \).

Theorems 1 and 2 establish the minimax optimal rate of the CIS on the set \( \mathcal{P}_{\alpha,\gamma} \). The requirement \( \alpha\gamma \leq d \) in Theorem 2 implies that \( \alpha \) and \( \gamma \) cannot be large simultaneously. As pointed out in Audibert and Tsybakov (2007), this is intuitively true because a very large \( \gamma \) implies a very smooth regression function \( \eta \), while a large \( \alpha \) implies that \( \eta \) cannot stay very long near 1/2, and hence when \( \eta \) hits 1/2, it should take off quickly. Lastly, we note that the minimax rate is slower than \( n^{-1} \), but approaches \( n^{-1} \) as the dimension \( d \) increases when \( \alpha\gamma = d \). As a reminder, Audibert and Tsybakov (2007) have established the minimax rate of regret as \( n^{-(\alpha+1)\gamma/(2\gamma+d)} \) under similar conditions. In next section, we will introduce a new classification procedure which attains the minimax optimal rates in both regret and CIS.

**3. Stabilized Nearest Neighbor Classifier.** We first review the general weighted nearest neighbor (wnn) classifiers and their optimal weights derived by Samworth (2012). For the class of wnn, the general CIS measure defined in Section 2 is shown to be proportional to the Euclidean norm of the weight vector. Based on the explicit expressions of regret and CIS, we propose a novel classification method called the stabilized nearest neighbor (snn) classifier which well balances the trade-off between classification accuracy and stability. Moreover, this new approach is proven to be theoretically optimal for both regret and CIS in the minimax sense.

**3.1. Review of WNN.** For a fixed \( x \in \mathcal{R} \), let \((X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})\) be the samples with increasing distance to \( x \). The wnn classifier \( \hat{\phi}^{wnn}_n(x) \) with a weight vector \( \mathbf{w}_n = (w_{ni})_{i=1}^n \) outputs 1 if and only if \( \sum_{i=1}^n w_{ni} \mathbb{I}\{Y_{(i)} = 1\} \geq 1/2 \), where the nonnegative weights satisfy \( \sum_{i=1}^n w_{ni} = 1 \).
For a smooth function $g$, we denote $\dot{g}(x)$ as its gradient vector at $x$. Samworth (2012) revealed a nice asymptotic expansion of the regret of the wnn classifier under the following assumptions.

(A.1) The set $\mathcal{R}$ is a compact $d$-dimensional manifold with boundary $\partial \mathcal{R}$.

(A.2) 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) $\eta$ is continuous on $U \setminus U_0$ with $U$ an open set containing $\mathcal{R}$; (ii) the restriction of the conditional densities 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 $f_2$.

(A.3) There exists $\rho > 0$ such that $\int_{\mathbb{R}^d} \|x\|^\rho d\tilde{P}(x) < \infty$. Moreover, for sufficiently small $\delta > 0$, $\inf_{x \in \mathcal{R}} \tilde{P}(B_0(x))/(a_d\delta^d) > 0$, where $a_d = \pi^{d/2}/\Gamma(1 + d/2)$ and $\Gamma(\cdot)$ is gamma function.

(A.4) 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 $\partial \eta(x) \neq 0$, where $\partial \eta$ is the restriction of $\eta$ to $\partial \mathcal{R}$.

Define

$$B_1 = \int_{\mathcal{S}} \frac{\tilde{f}(x)}{4\|\tilde{\eta}(x)\|} d\text{Vol}^{d-1}(x), \quad B_2 = \int_{\mathcal{S}} \frac{\tilde{f}(x_0)}{\|\tilde{\eta}(x)\|} a(x)^2 d\text{Vol}^{d-1}(x),$$

where $a(x)$ is defined in Appendix A.IV and $\text{Vol}^{d-1}$ is the natural $(d - 1)$-dimensional volume measure that $\mathcal{S}$ inherits. Proposition 1 below is one main result in Samworth (2012).

**Proposition 1.** (Theorem 1, Samworth (2012)) Under Assumptions (A.1)–(A.4), for each $\beta \in (0, 1/2)$, uniformly for $w_n \in W_{n, \beta}$ with $W_{n, \beta}$ defined in Appendix A.IV, as $n \to \infty$,

$$\text{Regret}(\text{wnn}) = \left\{B_1 \sum_{i=1}^{n} w_{ni}^2 + B_2 \left(\sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/\gamma}}\right)^2\right\} \{1 + o(1)\},$$

(3.1)

where $\alpha_i = i^{1+\frac{2}{\gamma}} - (i - 1)^{1+\frac{2}{\gamma}}$.

**Remark 1.** Assumptions (A.1)–(A.4) have also been employed to show the asymptotic expansion of the regret of the knn classifier (Hall et al., 2008). It is worth pointing out that the condition $\dot{\eta}(x) \neq 0$ in (A.4) is equivalent to the margin condition with $\gamma = 1$; see (2.1) in Samworth (2012). Furthermore, these assumptions ensure that $\tilde{f}(x_0)$ and $\dot{\eta}(x_0)$ are bounded away from zero and infinity on $\mathcal{S}$. Therefore, $B_1$ and $B_2$ are finite with $B_1 > 0$ and $B_2 \geq 0$, where $B_2 = 0$ only when $a(x)$ is identically zero on $\mathcal{S}$.

Samworth (2012) further derived the vector of nonnegative weights that minimizes the asymptotic expansion in (3.1) and called the corresponding nearest neighbor classifier as the optimal weighted nearest neighbor (ownn) classifier. As will be shown in Section 3.3, the ownn classifier is a special case of our snn classifier.
3.2. Asymptotically Equivalent Formulation of CIS. This subsection establishes the asymptotically equivalent form of CIS for the wnn procedure.

Denote the two resulting wnn classifiers trained on $\mathcal{D}_1$ and $\mathcal{D}_2$ as $\hat{\phi}_{n1}^w(x)$ and $\hat{\phi}_{n2}^w(x)$ respectively. In the sequel, we refer to the CIS of the wnn procedure by $\text{CIS}(\text{wnn})$ for simplicity. According to the definition in (2.1), the classification instability of the wnn procedure is

$$
(3.2) \quad \text{CIS}(\text{wnn}) = \mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2,X}(\hat{\phi}_{n1}^w(X) \neq \hat{\phi}_{n2}^w(X)).
$$

Theorem 3 shows that CIS of wnn is asymptotically proportional to the Euclidean norm of the weight vector.

**Theorem 3.** (Asymptotically Equivalent Form of CIS) Under Assumptions (A.1)–(A.4), for each $\beta \in (0, 1/2)$, uniformly for $w_n \in W_{n, \beta}$ with $W_{n, \beta}$ defined in Appendix A.IV, as $n \to \infty$,

$$
(3.3) \quad \text{CIS}(\text{wnn}) = B_3 \left( \sum_{i=1}^{n} w_{ni}^2 \right)^{1/2} \{1 + o(1)\},
$$

where $B_3 = 4B_1/\sqrt{\pi} > 0$.

We sketch the proof of Theorem 3 here. The detailed proof is included in Appendix A.VI. Let $S_n(x) = \sum_{i=1}^{n} w_{ni} 1\{Y_{(i)} = 1\}$. Since $\mathcal{D}_1, \mathcal{D}_2 \overset{i.i.d.}{\sim} \mathcal{D}$, we have

$$
\text{CIS}(\text{wnn}) = \mathbb{E}_X \left[ \mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2} \left( \hat{\phi}_{n1}^w(X) \neq \hat{\phi}_{n2}^w(X) \bigg| X \right) \right]
$$

$$
= \mathbb{E}_X \left[ \mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2} \left( \hat{\phi}_{n1}^w(X) = 1, \hat{\phi}_{n2}^w(X) = 2 \bigg| X \right) \right] + \mathbb{E}_X \left[ \mathbb{P}_{\mathcal{D}_1, \mathcal{D}_2} \left( \hat{\phi}_{n1}^w(X) = 2, \hat{\phi}_{n2}^w(X) = 1 \bigg| X \right) \right]
$$

$$
= \mathbb{E}_X \left[ 2\mathbb{P}_D \left( \hat{\phi}_{n1}^w(X) = 1\big| X \right) \left( 1 - \mathbb{P}_D \left( \hat{\phi}_{n2}^w(X) = 1\big| X \right) \right) \right]
$$

$$
= 2 \int_{\mathcal{R}} \left\{ \mathbb{P}(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\} \right\} d\bar{P}(x) - 2 \int_{\mathcal{R}} \left\{ \mathbb{P}^2(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\} \right\} d\bar{P}(x).
$$

We then study each term in the last line by dividing the region $\mathcal{R}$ into two parts: the area near the boundary $\mathcal{S}$ and the area far away from $\mathcal{S}$.

Theorem 3 demonstrates that CIS of the wnn procedure with the weight $w_n$ is asymptotically proportional to $(\sum_{i=1}^{n} w_{ni}^2)^{1/2}$. For the knn procedure (that is the wnn procedure with $w_{ni} = k^{-1} 1\{1 \leq i \leq k\}$), its CIS is asymptotically $B_3 1/\sqrt{k}$. Therefore, a larger value of $k$ leads to a more stable knn procedure, which is observed in Figure 1. Meanwhile, we observe a trade-off between classification regret and instability in Figure 1. This serves as a strong motivation for us to develop the snn procedure.

**Remark 2.** We note that the asymptotically equivalent form of CIS is related to the first term in the asymptotic expansion of regret (3.1), which was called the variance contribution to the
regret by Samworth (2012). In other words, part of the regret is explained by the CIS. However, to minimize regret does not necessarily lead to a small CIS because the relative weight between the first term and the second term of the regret is unobservable and out of control. On the other hand, the CIS allows a direct control of the classification instability through the proposed snn below.

3.3. Stability Control. We are now ready to propose a novel snn procedure that well balances the classification accuracy and stability. In particular, this procedure is shown to simultaneously achieve the minimax optimal rates in terms of regret and CIS.

Specifically, we determine the weights by minimizing CIS over an acceptable region where the classification regret is less than some constant $c_1 > 0$, that is,

$$\min_{w_n} \text{CIS}^2(wnn)$$

s.t. $\text{Regret}(wnn) \leq c_1$, $\sum_{i=1}^{n} w_{ni} = 1$, $w_n \geq 0$.

The Lagrangian formulation of (3.4) minimizes $\text{Regret}(wnn) + \lambda_0 \cdot \text{CIS}^2(wnn)$ subject to the constraints $\sum_{i=1}^{n} w_{ni} = 1$ and $w_n \geq 0$, where the Lagrangian multiplier $\lambda_0 > 0$. Proposition 1 and Theorem 3 imply the following asymptotically equivalent formulation:

$$\min_{w_n} \left( \sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2$$

s.t. $\sum_{i=1}^{n} w_{ni} = 1$, $w_n \geq 0$,

where $\lambda = (B_1 + \lambda_0 B_2^2)/B_2$ depends on the unknown constants $B_1$ and $B_2$ and the Lagrangian multiplier $\lambda_0$. When $\lambda \to \infty$, (3.5) leads to the most stable knn classifier, i.e., $k = n$. The case $\lambda \downarrow B_1 / B_2$ (i.e., $\lambda_0 \downarrow 0$) approaches the ownn classifier considered in Samworth (2012). Recall that $(n^{-2/d} \sum_{i=1}^{n} \alpha_i w_{ni})^2$ and $\sum_{i=1}^{n} w_{ni}^2$ in (3.5) represent the bias and variance terms of the regret expansion given in Proposition 1. By changing the weights between these two terms through $\lambda$, we are able to balance the classification accuracy and classification instability. As will be seen, such a balance method leads to a novel classifier which achieves the minimax rate optimality in both regret and CIS.

Theorem 4 gives the optimal weight with respect to the constrained optimization (3.5).

**Theorem 4. (Optimal Weight) For any fixed $\lambda > 0$, the minimizer of (3.5) 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 $\alpha_i = i^{1+\frac{d}{2}} - (i-1)^{1+\frac{d}{2}}$ and $k^* = \left\lfloor \frac{d(d+4)}{2(d+2)} \lambda^{d+4} n^{d+4} \right\rfloor$. 
We define the wnn classifier with the optimal weight \(w_{ni}^*\) as the stabilized nearest neighbor (snn) classifier. The snn classifier encompasses the ownn classifier as a special case when \(\lambda = B_1/B_2\). In practice, \(\lambda\) can be tuned by the cross validation method and the estimated \(\lambda\) is then plugged into the optimal weight formula in Theorem 4; see Algorithm 1 in Section 5.

Theorem 5 below shows that the regret and CIS of the proposed snn procedure achieve the minimax rate optimality. Recall that \(\mathcal{P}_{\alpha,\gamma}\) is defined in Theorem 2.

**Theorem 5.** (Optimal Rates of SNN) Under Assumptions (A.1)–(A.4), for any \(\alpha \geq 0\) and \(\gamma \in (0,2]\), the regret and CIS of the proposed snn procedure with any fixed \(\lambda > 0\) satisfy
\[
\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{Regret}(\text{snn}) \leq \tilde{C} n^{-(\alpha+1)\gamma/(2\gamma+d)},
\]
\[
\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \text{CIS}(\text{snn}) \leq C n^{-\alpha\gamma/(2\gamma+d)},
\]
for any \(n > 1\) and some constants \(\tilde{C}, C > 0\) independent of \(n\).

Corollary 1 investigates the difference between the regret and CIS of the snn procedure (with \(\lambda \neq B_1/B_2\)) and those of the ownn procedure. We write \(a_n \asymp b_n\) if the ratio sequence \(a_n/b_n\) stays away from zero and infinity as \(n \to \infty\).

**Corollary 1.** Under Assumptions (A.1)–(A.4), for any \(\alpha \geq 0\), \(\gamma \in (0,2]\) and \(\alpha\gamma \leq d\), we have, when \(\lambda \neq B_1/B_2\),
\[
\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \left\{ \text{Regret}(\text{snn}) - \text{Regret}(\text{ownn}) \right\} \asymp n^{-(1+\alpha)\gamma/(2\gamma+d)},
\]
\[
\sup_{P \in \mathcal{P}_{\alpha,\gamma}} \left\{ \text{CIS}(\text{ownn}) - \text{CIS}(\text{snn}) \right\} \asymp n^{-\alpha\gamma/(2\gamma+d)}.
\]

Corollary 1 illustrates that the regret of snn approaches that of the optimal ownn (from above) at a faster rate than the rate at which the CIS of ownn approaches that of the snn procedure (from above). This theoretically justifies the observations in Section 5 that the empirical gain in CIS of the snn procedure is more significant than the empirical loss in risk.

**4. Asymptotic Comparisons.** This section starts with the asymptotic comparisons of the CIS among the existing nearest neighbor classification procedures, i.e., knn, bnn and ownn. We further demonstrate in theory that our snn procedure significantly improves the CIS of the ownn procedure by slightly sacrificing the regret.

**4.1. CIS Comparison of Existing Methods.** The \(k\)-nearest neighbor classifier (knn) 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 bagged nearest neighbor (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., \( m \rightarrow \infty \) and \( m/n \rightarrow 0 \), the bnn classifier is shown to be a consistent classifier (Hall and Samworth, 2005). In particular, Hall and Samworth (2005) 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 \( i = 1, \ldots, n \), where \( q \) is the resampling ratio \( m/n \). Lastly, we consider the ownn classifier (achieving the minimal regret) in Samworth (2012).

We denote the CIS of the above classification procedures as \( \text{CIS}(\text{knn}) \), \( \text{CIS}(\text{bnn}) \) and \( \text{CIS}(\text{ownn}) \), respectively. Here \( k \) in the knn classifier is selected as the one minimizing the regret (Hall et al., 2008). The optimal \( q \) in the bnn classifier is calculated based on its asymptotic relationship to the optimal \( k \) in knn, that is, \( q = k^{-1}2^{d/(d+4)}\Gamma(2+2/d)^{2d/(d+4)} \); see (3.5) in Samworth (2012). Due to the difficulty in the direct estimation of \( B_1 \) and \( B_2 \), Samworth (2012) estimates the weight of the ownn classifier in two steps: (i) find \( k \) for the optimal knn classifier; (ii) calculate ownn’s weight according to an asymptotic relationship between ownn and the optimal knn. Corollary 2 gives the pairwise CIS ratios of these classification procedures.

**Corollary 2.** Under Assumptions (A.1)-(A.4) and \( B_2 > 0 \), as \( n \rightarrow \infty \),

\[
\frac{\text{CIS}(\text{ownn})}{\text{CIS}(\text{knn})} \rightarrow 2^{2/(d+4)} \left( \frac{d+2}{d+4} \right)^{(d+2)/(d+4)},
\]

\[
\frac{\text{CIS}(\text{bnn})}{\text{CIS}(\text{knn})} \rightarrow 2^{-2/(d+4)} \Gamma(2+2/d)^{d/(d+4)},
\]

\[
\frac{\text{CIS}(\text{bnn})}{\text{CIS}(\text{ownn})} \rightarrow 2^{-4/(d+4)} \Gamma(2+2/d)^{d/(d+4)} \left( \frac{d+4}{d+2} \right)^{(d+2)/(d+4)}.
\]

It is notable that these ratios merely depend on the feature dimension.

The limiting CIS ratios in Corollary 2 are plotted in Figure 2. A main message is that the ownn procedure is more stable than the knn and bnn procedures for any \( d \). We note that the largest improvement of the ownn procedure over the knn procedure is achieved when \( d = 4 \) and the improvement diminishes as \( d \rightarrow \infty \). The CIS ratio of the bnn procedure over the knn procedure 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 (Samworth, 2012). Therefore, bagging can be used to improve the knn procedure in terms of both accuracy and classification stability when \( d > 2 \). Furthermore, the CIS ratio of the ownn procedure over the bnn procedure is less than 1 for any \( d \) and it quickly converges to 1 for large \( d \). This implies that although the bnn procedure is asymptotically less stable than the ownn procedure, their difference sharply vanishes as \( d \) increases.
4.2. *Regret and CIS Comparisons between SNN and OWN*NN. Corollary 1 implies that ownn and snn share the same rates of regret and CIS (note that ownn is a special case of snn). Hence, it is of great interest to compare their ratios that can reflect finite sample performance. The asymptotic comparisons between snn and ownn in terms of regret and CIS are characterized in Corollary 3.

Denote the regret and CIS of the snn procedure as $\text{Regret}(\text{snn})$ and $\text{CIS}(\text{snn})$ respectively.

**Corollary 3.** Under Assumptions (A.1)-(A.4) and $B_2 > 0$, as $n \to \infty$,\[
\frac{\text{Regret}(\text{snn})}{\text{Regret}(\text{ownn})} \to \left\{ \frac{B_1}{\lambda B_2} \right\}^{d/(d+4)} \left\{ \frac{4 + d \lambda B_2 / B_1}{4 + d} \right\}.
\]
\[
\frac{\text{CIS}(\text{snn})}{\text{CIS}(\text{ownn})} \to \left\{ \frac{B_1}{\lambda B_2} \right\}^{d/(2(d+4))}.
\]

As seen from Corollary 3, both ratios of the snn procedure over the ownn procedure depend on the tuning parameter $\lambda$, and unknown constants $B_1$ and $B_2$. According to the facts that $\lambda = (B_1 + \lambda_0 B_3^2)/B_2$ in (3.5) and $B_3 = 4B_1/\sqrt{\pi}$ in (3.3), we have\[
(4.1) \quad \frac{\text{Regret}(\text{snn})}{\text{Regret}(\text{ownn})} \to \left\{ \frac{1}{1 + 16B_1 \lambda_0 / \pi} \right\}^{d/(d+4)} \left\{ \frac{4 + d(1 + 16B_1 \lambda_0 / \pi)}{4 + d} \right\},
\]
\[
(4.2) \quad \frac{\text{CIS}(\text{snn})}{\text{CIS}(\text{ownn})} \to \left\{ \frac{1}{1 + 16B_1 \lambda_0 / \pi} \right\}^{d/(2(d+4))}.
\]

Clearly, for any Lagrangian multiplier $\lambda_0 > 0$, the snn procedure has an improvement in CIS over the ownn. To show a concrete comparison, we fix $\lambda_0 = 1$ such that the ratios in (4.1) and (4.2) now only depend on $B_1$ and $d$. Figure 3 below shows 3D plots of these ratios as functions of $B_1$ and $d$. As expected, the regret of the ownn procedure is universally smaller than that of...
the snn procedure, while the ownn procedure has a larger CIS. For a fixed $B_1$, as the dimension $d$ increases, the regret of snn approaches that of ownn, while the advantage of snn over ownn in CIS grows. For a fixed dimension $d$, as $B_1$ increases, the ratio between the regret of snn and that of ownn gets larger, but the CIS advantage of the former also grows, partially compensating the loss in the regret.

As shown in Figure 3, snn is always more stable than ownn but with higher regret. Hence, it is of theoretical interest to investigate when the improvement of the snn procedure in CIS is larger than its loss in regret. We thus study the relative gain, which is defined as the ratio of the percentage of snn’s improvement in CIS and the percentage of snn’s loss in regret. The expression \((A.24)\) in Appendix A.XIII gives the relative gain as a function of $B_1$ and $d$. As shown in Figure 7, in most cases the logarithm of relative gain is larger than 0, which means that snn’s improvement in CIS is larger than its loss in regret. See Appendix A.XIII for more discussions on the relative gain. It is worth noting that all the above asymptotic findings are confirmed by our numerical experiments in the next section.

5. Experiments. This section first introduces the algorithms for estimating the risk, the CIS and for tuning the parameter $\lambda$ in the snn classifier, and then illustrates the significant improvements of the snn classifier over the existing methods using simulations and real examples.

5.1. Algorithm. In this subsection, we present an estimation scheme of CIS and risk based on cross validation, and then discuss the tuning parameter selection for the snn classifier.

The training sample $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ is randomly partitioned into five subsets $I_i$, 

![Fig 3. Regret ratio and CIS ratio of snn over ownn as functions of $B_1$ and $d$.](image)
\[ i = 1, \cdots, 5, \text{ with equal size } \lfloor n/5 \rfloor. \] For any fixed \( \lambda \), when \( i = 1 \), the prediction disagreements of the two classifiers, which are respectively trained from \( I_2 \cup I_3 \) and \( I_4 \cup I_5 \), are counted on the subset \( I_1 \). Similar procedure is repeated for \( i = 2, 3, 4 \) and 5. The average disagreement proportion is the estimated CIS. Similarly, the risk is estimated by the average misclassification rate on each subset using the same classifiers trained above. Finally, the selected tuning parameter is the value of \( \lambda \) leading to the minimal estimated \( \text{CIS}^2 + \text{Risk} \) (equivalently, \( \text{CIS}^2 + \text{Regret} \)). The weight \( w^*_n \) of the snn classifier in the numerical study is calculated by applying Theorem 4 using the optimal \( \hat{\lambda} \) found by Algorithm 1 below.

Algorithm 1 below summarizes the tuning of \( \lambda \). We let \( \hat{\phi}_{D}^\lambda \) denote an snn classifier with parameter \( \lambda \) trained from data \( D \).

**Algorithm 1:**

1. **Step 1.** Randomly partition \( D \) into five subsets \( I_i, i = 1, \cdots, 5 \).
2. **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 \( \hat{\phi}_{I_2 \cup I_3}^\lambda (X_j) \) and \( \hat{\phi}_{I_4 \cup I_5}^\lambda (X_j) \) respectively for each \( X_j \in I_1 \). Estimate the CIS and risk of classifier \( \hat{\phi}_{D}^\lambda \) by
   \[
   \hat{\text{CIS}}_1(\lambda) = \frac{1}{|I_1|} \sum_{(X_j, Y_j) \in I_1} \mathbb{1}\{\hat{\phi}_{I_2 \cup I_3}^\lambda (X_j) \neq \hat{\phi}_{I_4 \cup I_5}^\lambda (X_j)\},
   \]
   \[
   \hat{\text{Risk}}_1(\lambda) = \frac{1}{2|I_1|} \sum_{(X_j, Y_j) \in I_1} \left\{ \mathbb{1}\{\hat{\phi}_{I_2 \cup I_3}^\lambda (X_j) \neq Y_j\} + \mathbb{1}\{\hat{\phi}_{I_4 \cup I_5}^\lambda (X_j) \neq Y_j\} \right\}.
   \]
3. **Step 3.** Repeat the **Step 2** for \( i = 2, 3, 4 \) and 5 with the corresponding training and test sets. The estimated CIS and risk are
   \[
   \hat{\text{CIS}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \hat{\text{CIS}}_i(\lambda),
   \]
   \[
   \hat{\text{Risk}}(\lambda) = \frac{1}{5} \sum_{i=1}^{5} \hat{\text{Risk}}_i(\lambda).
   \]
4. **Step 4.** Grid search for the optimal tuning parameter \( \hat{\lambda} \) from a given range which minimizes
   \[
   \hat{\text{CIS}}^2(\lambda) + \hat{\text{Risk}}(\lambda).
   \]

In Steps 1 – 3, the estimation scheme based on cross-validation can be replaced by other data resampling strategies such as bootstrap or random weighting. In Step 4, equal weight for \( \text{CIS}^2 \) and Risk leads to a tuning criterion without any a priori preference. In practice, other weights may be employed depending on specific preferences. The optimal tuning parameter \( \hat{\lambda} \) from Algorithm 1 is consistent according to the consistency of the cross validation procedure; see Yang (2007). Alternatively, one may perform a two-stage approach by first selecting a set of \( \lambda \) values whose risks are below some pre-specified threshold, and then choosing the optimal \( \lambda \) among this set as the one that achieves the minimal classification instability.
5.2. Validation of Asymptotically Equivalent Forms. This subsection aims to support the asymptotically equivalent forms of CIS for the snn and ownn classifiers derived in Theorem 3 and the CIS and regret ratios of the snn procedure over the ownn procedure given in Corollary 3. We focus on a multivariate normal example in which regret and CIS have explicit expressions.

Assume that the underlying distributions of both classes are \( f_1 = N(0_2, I_2) \) and \( f_2 = N(1_2, I_2) \) and the prior class probability \( \pi_0 = 1/3 \). 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 are independently generated. The estimated risk and CIS are calculated based on 100 replications. In this example, some simple calculations lead to \( B_1 = 0.1299, B_2 = 10.68 \) and \( B_3 = 0.2931 \). According to Proposition 1, Theorems 3 and 4, we obtain that

\[
\begin{align*}
\text{Regret(snn)} &= \frac{0.1732}{k^*} + \frac{4.7467(k^*)^2}{n^2} \\
\text{CIS(snn)} &= \frac{0.3385}{(k^*)^{1/2}}
\end{align*}
\]

with \( k^* = [1.5^{1/3} \lambda^{1/3} n^{2/3}] \). Here we choose \( \lambda = (B_1 + B_3^2)/B_2 \), which corresponds to \( \lambda_0 = 1 \), so we have \( k^* = [0.3118 n^{2/3}] \). This amounts to minimizing \( \text{CIS}^2 + \text{Risk} \); see the line below (3.4). Recall that Algorithm 1 tunes \( \lambda \) by minimizing \( \tilde{\text{CIS}}^2(\lambda) + \tilde{\text{Risk}}(\lambda) \), which also corresponds to \( \lambda_0 = 1 \). Hence, our choice of \( \lambda \) above leads to a fair comparison. Similarly, the asymptotic regret and CIS of the ownn classifier are (5.1) and (5.2) with \( k^* = [0.2633 n^{2/3}] \) due to (2.4) in Samworth (2012).

In Figures 4, we plot the asymptotic value of CIS and risk of the snn and ownn classifiers computed using the above formula, along with the estimated CIS and risk based on the test data (see Step 3 of Algorithm 1). 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 classifier is 0.078 (0.085) while the estimated CIS is 0.079 (0.086). In Figure 5, we plot the asymptotic risks, that is, the asymptotic regret (5.1) plus the true Bayes error 0.215, of the snn and ownn classifiers, along with the estimated ones. Here we compute the Bayes error by Monte Carlo integration. Again the difference of the estimated risk and asymptotic risk decreases as the sample size grows, although the convergence is slower than that of CIS. This “slower convergence” phenomenon has empirically supported the theoretical finding in Corollary 1.

Furthermore, according to (4.2), the asymptotic CIS ratio of the snn classifier over the ownn classifier is 0.9189 in this example, and the 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 (4.1), 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 error 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 error is 0.215 (see Figure 5). Similar issue was reported in Samworth (2012) as well.

\[
B_1 = \frac{\sqrt{2\pi}}{3\pi \mu d} \exp \left( -\frac{\mu d/2 - \ln 2/\mu}{2d} \right).
\]

Therefore, 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 is generated by setting $n = 200$, $d = 2$ or 5, $f_1 \sim 0.5N(0_d, I_d) + 0.5N(\mu_d, I_d)$, $f_2 \sim 0.5N(1.5d, I_d) + 0.5N(4.5d, 2I_d)$, and $\pi_0 = 1/2$ or 1/3.

Simulation 3 has the same setting as Simulation 2, except that $f_1 \sim 0.5N(0_d, \Sigma) + 0.5N(3d, 2\Sigma)$ and $f_2 \sim 0.5N(1.5d, \Sigma) + 0.5N(4.5d, 2\Sigma)$, where $\Sigma$ is the Toeplitz matrix whose $j$th entry of its first row is $0.6^{j-1}$.

5.3. Simulations. This subsection shows the improvement of CIS of the snn procedure over the existing nearest neighbor procedures in three simulated experiments.

For the knn, ownn and bnn classifiers, the parameter $k$ is tuned from 20 equally spaced grid points from 5 to $n/2$. For a fair comparison, in the snn classifier, the parameter $\lambda$ is tuned so that the corresponding parameter $k$ (see Theorem 4) are equally spaced and falls into the same range.

In Simulation 1, we assume that the underlying distributions of the two classes are $f_1 = N(0_d, I_d)$ and $f_2 = N(\mu_d, I_d)$ with the prior probability $\pi_0 = 1/3$. We set sample size $n = 200$ and choose $\mu$ such that the resulting $B_1$ is fixed as 0.1 for different $d$. Specifically, in Appendix A.XII we show that

\[
(5.3) \quad B_1 = \frac{\sqrt{2\pi}}{3\pi \mu d} \exp \left( -\frac{\mu d/2 - \ln 2/\mu}{2d} \right).
\]
Simulation 1 is a relatively easy classification problem. It aims to empirically verify the asymptotic comparisons in Section 4.2. Simulation 2 examines the bimodal effect and Simulation 3 combines bimodality with dependence between the components. The latter two settings have also been considered in Samworth (2012). In each simulation setting, a test data set of size 1000 is independently generated and the average CIS and test error over 100 replications are calculated on the test data.

Figure 6 shows the average regret and CIS in Simulation 1. The regrets of the snn procedure are not significantly different from those of the other classification procedures. In some cases, the regret of the ownn procedure is even not the smallest, which might be explained by the Monte Carlo error (Samworth, 2012). However, the snn procedure always has a dramatically smaller CIS than other methods over all cases. In particular, as $d$ increases, the improvements of the snn procedure over all other procedures increase. This agrees with the asymptotic findings in Section 4.2. In particular, when $d = 10$, the CIS of the snn procedure is 0.0183, while that of the knn, bnn, and ownn procedures are 0.0683, 0.0937 and 0.0853, respectively. This phenomenon illustrates the advantages of the snn procedure: with an ignorable loss in prediction accuracy, we can achieve a big improvement in classification stability.

Table 1 summarizes the test error and CIS in Simulations 2 and 3. Again, the snn procedure obtains the minimal CIS among all 8 cases. Moreover, it achieves the minimal error among 4 out of the 8 cases. The ownn procedure achieves minimal error in the rest 4 out of 8 scenarios, but it has a significantly larger CIS than snn. The largest improvements of the snn procedure in...
both simulations are more than 50%. In Table 1, the column of $\Delta$ shows the percentage change of the error or the CIS of the snn procedure compared with the ownn procedure (see Appendix A.XIII). A negative $\Delta$ value means the snn procedure is better than the ownn procedure. Clearly, with a slight sacrifice of classification accuracy (less than 2%), our snn procedure has a significant improvement in CIS (about 10% $\sim$ 60%). These observations are consistent with the minimax results in Corollary 1.

5.4. Real Examples. In this subsection, we investigate the performance of the knn, bnn, ownn and snn procedures on 7 datasets publicly available in the UCI Machine Learning Repository (Bache and Lichman, 2013). The seven datasets are Haberman survival data (haberman), liver disorder data (liver), appendicitis data (appendicitis), Pima Indians diabetes data (pima), Stalog heart data (stalog), Australian credit approval data (credit) and SPECT heart data (spect). The summary information of these datasets is given in the first three columns of Table 2. For each data set, we randomly split it into training and test sets with equal size. The same procedure for tuning parameter in the simulation is applied here. We compute the classification error and CIS on each test set. These procedures are repeated 100 times and the average error and CIS are reported in Table 2.

Similar to the simulation results, the snn procedure obtains the minimal CIS over all cases. Specifically, on the data sets haberman, appendicitis and spect, the snn procedure has a significant improvement of CIS, about 30%, over all the other procedures. On the data sets liver and pima, the improvement in CIS of the snn procedure is about 10%. On the other two data sets, this improvement is about 5%. In addition, when $d$ is relatively small, both the knn and
ownn procedures have smaller CIS than the bnn procedure; when $d$ is relatively large, the knn procedure has a larger CIS compared with the bnn procedure, whose CIS is very close to that of the ownn procedure. These comparisons agree with the theoretical findings in Section 4. Finally, the snn procedure achieves the smallest error among 4 out of the 7 cases and only has slightly larger errors in the other 3 cases. This further illustrates the advantage of the snn procedure, that is, with slight sacrifice of classification accuracy, it can have a significant improvement in classification stability.

Acknowledgements: We would like to thank Statistical and Applied Mathematical Sciences

Table 1
Error and CIS (both multiplied by 100) of the knn, bnn, ownn, and snn procedures in Simulations 2 and 3. The smallest value in each case is given in bold. Standard errors are given in subscript. $\Delta$ refers to the percentage change of the error or the CIS of the snn procedure compared with the ownn procedure. A negative $\Delta$ value means the snn procedure is better than the ownn procedure.

| Sim | $d$ | $\pi_0$ | knn Error | bnn Error | ownn Error | snn Error | $\Delta$ |
|-----|-----|---------|-----------|-----------|------------|-----------|---------|
| 2   | 1/2 | Bayes 26.83 | 30.13_{0.167} | 29.85_{0.162} | 29.75_{0.176} | 30.14_{0.174} | 1.31% |
|     |     |          | 31.80_{0.978} | 30.48_{0.873} | 30.06_{0.833} | 17.82_{0.76} | -40.72% |
| 2   | 1/3 | Bayes 22.76 | 23.79_{0.111} | 23.85_{0.131} | 23.68_{0.113} | 23.91_{0.075} | 0.97% |
|     |     |          | 14.90_{0.517} | 13.99_{0.508} | 14.99_{0.503} | 6.90_{0.394} | -53.97% |
| 5   | 1/2 | Bayes 11.61 | 16.50_{0.132} | 16.00_{0.142} | 15.91_{0.131} | 15.51_{0.118} | -2.51% |
|     |     |          | 17.02_{0.414} | 16.19_{0.391} | 16.15_{0.449} | 14.43_{0.332} | -10.65% |
| 5   | 1/3 | Bayes 10.58 | 15.14_{0.115} | 15.00_{0.101} | 14.88_{0.102} | 15.01_{0.110} | 0.87% |
|     |     |          | 11.57_{0.332} | 12.52_{0.324} | 11.99_{0.324} | 10.57_{0.276} | -11.84% |

| Sim | $d$ | $\pi_0$ | knn Error | bnn Error | ownn Error | snn Error | $\Delta$ |
|-----|-----|---------|-----------|-----------|------------|-----------|---------|
| 2   | 1/2 | Bayes 32.45 | 35.07_{0.144} | 35.01_{0.139} | 35.05_{0.139} | 34.36_{0.104} | -1.97% |
|     |     |          | 33.49_{0.457} | 32.73_{0.057} | 32.55_{0.099} | 15.25_{0.021} | -53.15% |
| 2   | 1/3 | Bayes 25.30 | 26.13_{0.136} | 26.29_{0.158} | 26.14_{0.136} | 26.25_{0.078} | 0.42% |
|     |     |          | 14.72_{0.686} | 15.47_{0.549} | 15.83_{0.543} | 6.03_{0.424} | -61.91% |
| 5   | 1/2 | Bayes 25.58 | 31.05_{0.118} | 30.82_{0.136} | 30.78_{0.132} | 30.21_{0.126} | -1.85% |
|     |     |          | 23.82_{0.668} | 25.39_{0.629} | 25.19_{0.650} | 20.31_{0.669} | -19.37% |
| 5   | 1/3 | Bayes 21.98 | 24.82_{0.093} | 24.92_{0.122} | 24.81_{0.105} | 24.57_{0.059} | -8.48% |
|     |     |          | 7.08_{0.324} | 7.97_{0.592} | 7.49_{0.549} | 6.23_{0.38} | -16.82% |
Table 2

Error and CIS (both multiplied by 100) of the knn, bnn, ownn, and snn procedures in real data sets. The smallest value in each case is given in bold. Standard errors are given in subscript. ∆ refers to the percentage change of the error or the CIS of the snn procedure compared with the ownn procedure. A negative ∆ value means the snn procedure is better than the ownn procedure.

| Data     | n     | d | knn  | bnn   | ownn | snn   | ∆       |
|----------|-------|---|------|-------|------|-------|---------|
| haberman | 306   | 3 | 26.08 | 26.60 | 26.30 | 26.56 | 0.99%   |
|          |       |   | 0.281 | 0.268 | 0.275 | 0.260 |         |
|          |       |   |       | 6.03  | 5.25  | 3.92  | -25.33% |
|          |       |   |       |       |       |       |         |
| liver    | 345   | 6 | 38.76 | 38.61 | 37.50 | 38.27 | 2.05%   |
|          |       |   | 0.356 | 0.488 | 0.360 | 0.399 |         |
|          |       |   |       | 39.86 | 39.38 | 33.20 | -15.69% |
| appendicitis | 106  | 7 | 15.36 | 17.91 | 15.92 | 15.19 | -4.59% |
|          |       |   | 0.477 | 0.786 | 0.533 | 0.493 |         |
|          |       |   |       | 18.43 | 14.36 | 9.38  | -34.68% |
| pima     | 768   | 8 | 26.08 | 25.92 | 25.83 | 26.04 | 0.81%   |
|          |       |   | 0.212 | 0.198 | 0.192 | 0.205 |         |
|          |       |   |       | 14.36 | 14.11 | 12.64 | -10.42% |
| stalog   | 270   | 13| 17.44 | 17.64 | 17.37 | 16.97 | -2.30%  |
|          |       |   | 0.236 | 0.297 | 0.245 | 0.238 |         |
|          |       |   |       | 12.72 | 11.94 | 11.28 | -5.53%  |
| credit   | 690   | 14| 14.55 | 14.63 | 14.60 | 14.54 | -0.41%  |
|          |       |   | 0.144 | 0.144 | 0.146 | 0.144 |         |
|          |       |   |       | 6.85  | 6.77  | 6.41  | -5.32%  |
| spect    | 267   | 22| 20.66 | 20.41 | 20.34 | 20.25 | -0.44%  |
|          |       |   | 0.330 | 0.402 | 0.310 | 0.298 |         |
|          |       |   |       | 12.90 | 11.09 | 6.86  | -38.14% |

Institute (SAMSI) for warmly hosting all three authors in the 2012–2013 Massive Dataset program where part of this work was done. We thank Professor Richard Samworth for personal communications.

APPENDIX

A.I. Proof of Theorem 1. Following the proofs of Lemma 3.1 in Audibert and Tsybakov (2007), we consider the sets $A_j \subset \mathcal{R}$, where

$$A_0 = \{ x \in \mathcal{R} : 0 < |\eta(x) - 1/2| \leq \delta \},$$

$$A_j = \{ x \in \mathcal{R} : 2^{j-1}\delta < |\eta(x) - 1/2| \leq 2^j\delta \} \text{ for } j \geq 1.$$

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

$$\text{CIS}(\Psi) = E[I(\hat{\phi}_{n1}(X) \neq \hat{\phi}_{n2}(X))],$$

where $\hat{\phi}_{n1}$ and $\hat{\phi}_{n2}$ 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 \( \phi_{\text{Bayes}} \), we have

\[
\text{CIS}(\Psi) = 2\mathbb{E}[\mathbb{I}\{\hat{\phi}_{n1}(X) = \phi_{\text{Bayes}}(X), \hat{\phi}_{n2}(X) \neq \phi_{\text{Bayes}}(X)\}]
\]

\[
= 2\mathbb{E}[\{1 - \mathbb{I}\{\hat{\phi}_{n1}(X) \neq \phi_{\text{Bayes}}(X)\}\} \mathbb{I}\{\hat{\phi}_{n2}(X) \neq \phi_{\text{Bayes}}(X)\}]
\]

\[
= 2\mathbb{E}_X[\mathbb{P}_{\mathcal{D}_1}(\hat{\phi}_{n1}(X) \neq \phi_{\text{Bayes}}(X)|X) - \mathbb{P}_{\mathcal{D}_1}(\hat{\phi}_{n1}(X) \neq \phi_{\text{Bayes}}(X)|X)^2]
\]

\[
\leq 2\mathbb{E}[\mathbb{I}\{\hat{\phi}_{n1}(X) \neq \phi_{\text{Bayes}}(X)\}],
\]

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 \( \hat{\phi}_{n1} \) for some \( \hat{\phi}_n \) from now on. We further have

\[
\text{CIS}(\Psi) \leq 2 \sum_{j=0}^{\infty} \mathbb{E}[\mathbb{I}\{\hat{\phi}_n(X) \neq \phi_{\text{Bayes}}(X)\} \mathbb{I}\{X \in A_j\}]
\]

\[
\leq 2\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + 2 \sum_{j \geq 1} \mathbb{E}[\mathbb{I}\{\hat{\phi}_n(X) \neq \phi_{\text{Bayes}}(X)\} \mathbb{I}\{X \in A_j\}].
\]

Given the event \( \{\hat{\phi}_n \neq \phi_{\text{Bayes}}\} \cap \{|\eta - 1/2| > 2^{j-1}\delta\} \), we have \( |\hat{\eta}_n - \eta| \geq 2^{j-1}\delta \). Therefore, for any \( j \geq 1 \), we have

\[
\mathbb{E}[\mathbb{I}\{\hat{\phi}_n(X) \neq \phi_{\text{Bayes}}(X)\} \mathbb{I}\{X \in A_j\}]
\]

\[
\leq \mathbb{E}[\mathbb{I}\{|\hat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta\} \mathbb{I}\{2^{j-1}\delta < |\eta(X) - 1/2| \leq 2^j\delta\}]
\]

\[
\leq \mathbb{E}[\mathbb{P}_X(|\hat{\eta}_n(X) - \eta(X)| \geq 2^{j-1}\delta|X) \mathbb{I}\{0 < |\eta(X) - 1/2| \leq 2^j\delta\}]
\]

\[
\leq C_1 \exp(-C_2a_n(2^{j-1}\delta)^2)\mathbb{P}_X(0 < |\eta(X) - 1/2| \leq 2^j\delta)
\]

\[
\leq C_1 \exp(-C_2a_n(2^{j-1}\delta)^2)\mathbb{P}_X(|\eta(X) - 1/2| \leq 2^j\delta),
\]

where the last inequality is due to margin assumption (2.2) and condition (2.3).

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

\[
\text{CIS}(\Psi) \leq C_0a_n^{-\alpha/2} + C_0C_1a_n^{-\alpha/2}\sum_{j \geq 1} 2^{\alpha j + 1}e^{-C_24^{j-1}}
\]

\[
\leq Ca_n^{-\alpha/2},
\]

for some \( C > 0 \) depending only on \( \alpha, C_0, C_1 \) and \( C_2 \).

\[\blacksquare\]

A.II. A Lemma for proving Theorem 2 (Lower Bound of CIS). Before proving Theorem 2, we adapt the Assouad’s lemma to prove the lower bound of CIS. This lemma is also of independent interest.

We first introduce an important definition called \((m, w, b, b')\)-hypercube defined in Audibert (2004). 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 \( Y_i \in \mathcal{Y} = \{1, 2\}. \) Let \( \mathcal{F}(\mathcal{X}, \mathcal{Y}) \) denote the set of all measurable functions mapping from \( \mathcal{X} \) into \( \mathcal{Y}. \) Let \( \mathcal{Z} = \mathcal{X} \otimes \mathcal{Y}. \) For the distribution function \( P, \) we denote its corresponding probability and expectation as \( P \) and \( E, \) respectively.

**Definition 3.** (Audibert, 2004) 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 \( \mathcal{H} = \{ P_\bar{\sigma} : \bar{\sigma} \Delta (\sigma_1, \ldots, \sigma_m) \in \{-1, +1\}^m \} \) of probability distributions \( P_\bar{\sigma} \) of \((X, Y) \) on \( \mathcal{Z} \) as follows.

For any \( P_\bar{\sigma} \in \mathcal{H}, \) the marginal distribution of \( X \) does not depend on \( \bar{\sigma} \) and its distribution \( \theta \) satisfies the following conditions. There exists a partition \( \mathcal{X}_0, \ldots, \mathcal{X}_m \) of \( \mathcal{X} \) satisfying,

(i) for any \( j \in \{1, \ldots, m\}, \) \( \theta(X_j) = w; \)

(ii) for any \( j \in \{0, \ldots, m\} \) and any \( X \in \mathcal{X}_j, \) we have

\[
P_\bar{\sigma}(Y = 1|X) = \frac{1 + \sigma_j \psi(X)}{2}
\]

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

\[
b \Delta \left[ 1 - \left( \theta(\sqrt{1 - \psi^2(X)}|X \in \mathcal{X}_j) \right)^2 \right]^{1/2},
\]

\[
b' \Delta \theta(\psi(X)|X \in \mathcal{X}_j).
\]

**Lemma A.1.** If a collection of probability distributions \( \mathcal{P} \) contains a \((m, w, b, b')\)-hypercube, then for any measurable estimator \( \hat{\phi}_n \) obtained by applying \( \Psi \) to the training sample \( \mathcal{D}, \) we have

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

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

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

\[
V(P_1, P_2) = 1 - \int \left( \frac{P_1}{P_0} \wedge \frac{P_2}{P_0} \right) dP_0,
\]

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 Audibert (2004) 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 + \( \mathbb{1}\{\eta(X) < 1/2\} = 1 + \mathbb{1}\{(1 + \sigma_j \psi(X))/2 < 1/2\} = (3 - \sigma_j)/2 \), a.s. Therefore, we have

\[
\text{A.2} \quad \sup_{P \in \mathcal{P}} \mathbb{E}^\otimes n [P_X(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X))]
\geq \sup_{\delta \in \{-1,+1\}^m} \left\{ P^\otimes n_{\delta} \mathbb{P}_{\delta}(\mathbb{1}\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)\}) \right\}
\geq \sup_{\delta \in \{-1,+1\}^m} \left\{ \sum_{j=1}^m \theta[\mathbb{1}\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)\} \geq \frac{3 - \sigma_j}{2}; X \in \mathcal{X}_j]\right\}
\]

\(\text{(A.3)} \quad \mathbb{E}_{\nu \otimes m} \sum_{j=1}^m P^\otimes n_{\delta j,0} \left[ \theta[\mathbb{1}\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)\} \geq \frac{3 - \sigma_j}{2}; X \in \mathcal{X}_j]\right]\)

\(\text{(A.4)} \quad \mathbb{E}_{\nu \otimes (m-1)(d \sigma - j)} \sum_{j=1}^m P^\otimes n_{\delta j,0} \mathbb{E}_{\nu(\sigma_j)} \left[ \frac{P^\otimes n_{\delta j,0}}{P^\otimes n_{\delta j,1}} \theta[\mathbb{1}\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)\} \geq \frac{3 - \sigma_j}{2}; X \in \mathcal{X}_j]\right]\)

\(\text{(A.5)} \quad \mathbb{E}_{\nu \otimes (m-1)(d \sigma - j)} \sum_{j=1}^m \frac{1}{2} \theta[\mathbb{1}\{X \in \mathcal{X}_j]\} \left[ 1 - V(P^\otimes n_{\delta j,1}, P^\otimes n_{\delta j,1}) \right]\)

\[
\geq \frac{m w}{2} \left[ 1 - b \sqrt{m w} \right]
\]

where \(\text{(A.2)}\) is due to the assumption that \( \mathcal{P} \) contains a \((m, w, b, b')\)-hypercube, \(\text{(A.3)}\) is because the supremum over the \( m \) Rademacher variables is no less than the corresponding expected value, \(\text{(A.4)}\) is because we separate the space of the expectation into two parts: \( \nu(\sigma_j) \) and \( \nu \otimes (m-1)(d \sigma - j) \). Finally, the inequality \(\text{(A.5)}\) is due to \( P^\otimes n_{\delta j,0} \geq \{ P^\otimes n_{\delta j,1} \wedge P^\otimes n_{\delta j,-1} \} \) and the latter is not random with respect to \( \nu(\sigma_j) \). This ends the proof of Lemma \text{A.1}. \(\blacksquare\)

\text{A.III. Proof of Theorem 2.} According to the proof of Theorem 1, we have

\[
\text{CIS}(\Psi) = 2 \left\{ \mathbb{E}_X [P_D(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X] - \mathbb{E}_X [\mathbb{P}_D(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X)]^2 \right\}.
\]

We bound each part separately.

\text{Audibert and Tsybakov (2007)} showed that when \( \alpha \gamma \leq d \), the set of probability distribution \( \mathcal{P}_{\alpha, \gamma} \) contains a \((m, w, b, b')\)-hypercube with \( w = C_3 q^{-d}, m = [C_4 q^{d-\alpha \gamma}], b = b' = C_5 q^{-\gamma} \) and
$q = [C_6n^{1/(2\gamma + d)}]$, with some properly chosen constants $C_i \geq 0$ for $i = 3, \ldots, 6$ and $C_6 \leq 1$. Therefore, Lemma A.1 implies that the first part is bound, that is,

$$
\begin{align*}
\sup_{P \in P_{\alpha,\gamma}} \mathbb{E}_X[\mathbb{P}_D(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X)] \\
= \sup_{P \in P_{\alpha,\gamma}} \mathbb{E}_D[\mathbb{P}_X(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X))] \\
\geq \frac{m_w}{2}[1 - b\sqrt{nw}] \\
= (1 - C_6)C_3C_4C_5n^{-\alpha\gamma/(2\gamma + d)}.
\end{align*}
$$

To bound the second part, we again consider the sets $A_j \subset \mathcal{R}$, defined in Appendix A.I. Furthermore, on the event $\{\hat{\phi}_n \neq \phi^{\text{Bayes}}\} \cap \{|\eta - 1/2| > 2^{i-1}\}$, we have $|\hat{\eta}_n - \eta| \geq 2^{i-1}\delta$. Therefore, letting $\delta = a_n^{1/2}$, we have

$$
\mathbb{E}_X[\mathbb{P}_D(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X)]^2 \\
= \sum_{j=0}^{\infty} \mathbb{E}_X[\mathbb{P}_D(\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X\})^2 1\{X \in A_j\}] \\
\leq \mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + \sum_{j=1}^{\infty} \mathbb{E}_X[\mathbb{P}_D(\{\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X\})^2 1\{X \in A_j\}] \\
\leq \mathbb{P}_X(0 < |\eta(X) - 1/2| \leq \delta) + \sum_{j=1}^{\infty} C_1e^{-2C_24^{j-1}} \mathbb{P}_X(0 < |\eta(x) - 1/2| \leq 2^j\delta) \\
\leq C_0a_n^{-\alpha/2} + C_0C_1a_n^{-\alpha/2} \sum_{j=1}^{\infty} 2^{\alpha j}e^{-2C_24^{j-1}} \\
\leq C_7a_n^{-\alpha/2},
$$

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

$$
\mathbb{E}_X[\mathbb{P}_D(\hat{\phi}_n(X) \neq \phi^{\text{Bayes}}(X)|X)]^2 \leq C_7n^{-\alpha\gamma/(2\gamma + d)}.
$$

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 2. $\blacksquare$

A.IV. Definitions of $a(x)$ and $W_{n,\beta}$. For a smooth function $g: \mathbb{R}^d \rightarrow \mathbb{R}$, we let $\hat{g}(x)$ denote its gradient vector at $x$, and $g_j(x)$ its $j$th partial derivative at $x$. Similarly, we let $\hat{g}(x)$ denote the Hessian matrix at $x$, and $g_{jk}(x)$ the $(j, k)$th element of $\hat{g}(x)$. Let $c_{j,d} = \int_{v : |v| \leq 1} v_j^2 dv$. We define

$$
a(x) = \sum_{j=1}^{d} c_{j,d} \{\eta_j(x)f_j(x) + 1/2\eta_{jj}(x)f(x)\}/a_j^{1+2/d}f(x)^{1+2/d}.
$$
After substitute (A.6), we only show the second equality. The first equality can be shown by similar approach. Note that
\[
\int_{-\infty}^{\infty} u \{ G(-bu-a) - 1 \{u < 0 \} \} du = \int_{-\infty}^{0} u \{ G(-bu-a) - 1 \} du + \int_{0}^{\infty} uG(-bu-a) du
\]
(A.6)

After substitute \( t = -bu-a \) for each term, we have
\[
\int_{-\infty}^{0} u \{ G(-bu-a) - 1 \} du = \frac{1}{b^2} \int_{-\alpha}^{\infty} (t+a)(1-G(t))dt
\]
\[
\int_{0}^{\infty} uG(-bu-a) du = \frac{1}{b^2} \int_{-\infty}^{-\alpha} (t+a)(-G(t))dt
\]

Plugging these two into (A.6), we have
\[
\int_{-\infty}^{\infty} u \{ G(-bu-a) - 1 \{u < 0 \} \} du
\]
\[
= \frac{1}{b^2} \left\{ - \int_{-\infty}^{-\alpha} tG(t)dt - a \int_{-\alpha}^{0} G(t)dt + \int_{-\alpha}^{\infty} t(1-G(t))dt + a \int_{-\alpha}^{\infty} (1-G(t))dt \right\}
\]
\[
= \frac{1}{b^2} \left\{ I + II + III + IV \right\}.
\]
Applying integration by part, we can calculate
\[ I = -\frac{1}{2} \left[ a^2 G(-a) - \int_{-\infty}^{-a} t^2 dG(t) \right] \]
\[ II = a \left[ a G(-a) + \int_{-\infty}^{-a} t dG(t) \right] \]
\[ III = \frac{1}{2} \left[ -a^2 (1 - G(-a)) + \int_{-a}^{\infty} t^2 dG(t) \right] \]
\[ IV = a \left[ a(1 - G(-a)) + \int_{-a}^{\infty} t dG(t) \right] \]

Plugging I-IV into (A.6) leads to desirable equality. This concludes the proof of Lemma A.2. ■

A.VI. **Proof of Theorem 3.** The CIS defined in (3.2) can be expressed in the following way.

\[
\text{CIS(wnn)} = \mathbb{E}_X \left[ \mathbb{P}_{D_1,D_2} \left( \hat{\phi}_{D_1}^{wn}(X) \neq \hat{\phi}_{D_2}^{wn}(X) \right| X \right] 
\]
\[
= \mathbb{E}_X \left[ \mathbb{P}_{D_1,D_2} \left( \hat{\phi}_{D_1}^{wn}(X) = 1, \hat{\phi}_{D_2}^{wn}(X) = 2 \right| X \right] + \mathbb{E}_X \left[ \mathbb{P}_{D_1,D_2} \left( \hat{\phi}_{D_1}^{wn}(X) = 2, \hat{\phi}_{D_2}^{wn}(X) = 1 \right| X \right] 
\]
\[
= \mathbb{E}_X \left[ 2\mathbb{P}_{D_1} \left( \hat{\phi}_{D_1}^{wn}(X) = 1\right) \left( 1 - \mathbb{P}_{D_1} \left( \hat{\phi}_{D_1}^{wn}(X) = 1\right) \right) \right],
\]
where the last equality is valid because $D_1$ and $D_2$ are independently and identically distributed 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 $\|X(1) - x\| \leq \|X(2) - x\| \leq \ldots \leq \|X(n) - x\|$ 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
\[
\mathbb{E}_X \left[ \mathbb{P} \left( \hat{\phi}_{D}^{wn}(X) = 1\right) \right] = \int_{\mathcal{R}} \mathbb{P} \left( S_n(x) \geq 1/2 \right) d\bar{P}(x),
\]
\[
\mathbb{E}_X \left[ \mathbb{P}^2 \left( \hat{\phi}_{D}^{wn}(X) = 1\right) \right] = \int_{\mathcal{R}} \mathbb{P}^2 \left( S_n(x) \geq 1/2 \right) d\bar{P}(x),
\]
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
\[
\text{CIS(wnn)/2} = \int_{\mathcal{R}} \mathbb{P} \left( S_n(x) \geq 1/2 \right) \left( 1 - \mathbb{P} \left( S_n(x) \geq 1/2 \right) \right) d\bar{P}(x)
\]
\[
= \int_{\mathcal{R}} \left\{ \mathbb{P} \left( S_n(x) < 1/2 \right) - \mathbb{1} \{ \eta(x) < 1/2 \} \right\} d\bar{P}(x)
\]
\[
- \int_{\mathcal{R}} \left\{ \mathbb{P}^2 \left( S_n(x) < 1/2 \right) - \mathbb{1} \{ \eta(x) < 1/2 \} \right\} d\bar{P}(x)
\]

Denote the boundary $\mathcal{S} = \{x \in \mathcal{R} : \eta(x) = 1/2\}$. For $\epsilon > 0$, let $\mathcal{S}^{\epsilon} = \{x \in \mathbb{R}^d : \eta(x) = 1/2 \text{ and } \text{dist}(x, \mathcal{S}) < \epsilon\}$, where $\text{dist}(x, \mathcal{S}) = \inf_{x_0 \in \mathcal{S}} \|x - x_0\|$. We will focus on the set
\[
\mathcal{S}^{\epsilon} = \left\{ x_0 + t \frac{\eta(x_0)}{\|\eta(x_0)\|} : x_0 \in \mathcal{S}^{\epsilon}, |t| < \epsilon \right\}.
\]
Let $\mu_n(x) = \mathbb{E}\{S_n(x)\}$, $\sigma_n^2(x) = \text{Var}\{S_n(x)\}$, and $\epsilon_n = n^{-\beta/d^4}$. Denote $s_n^2 = \sum_{i=1}^n w_{ni}^2$ and $t_n = n^{-2/d}\sum_{i=1}^n \alpha_i w_{ni}$. Samworth (2012) showed that

\begin{equation}
\sup_{w_n \in W_{n,\beta}} \sup_{x \in S^n} |\mu_n(x) - \eta(x) - a(x)t_n| = o(t_n),
\end{equation}

\begin{equation}
\sup_{w_n \in W_{n,\beta}} \sup_{x \in S^n} \left| \frac{\sigma_n^2(x)}{s_n^2} - \frac{1}{4} \right| = o(s_n^2).
\end{equation}

We organize our proof in three steps. In Step 1, we focus on analyzing on the set $\mathcal{R} \cap S^{\epsilon_n}$; in Step 2, we focus on the complement set $\mathcal{R} \setminus S^{\epsilon_n}$; 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 $\tilde{f} = \pi_0 f_1 + (1 - \pi_0) f_2$ as the Radon-Nikodym derivative with respect to Lebesgue measure of the restriction of $\tilde{P}$ to $S^{\epsilon_n}$ for large $n$. We need to show that, uniformly for $w_n \in W_{n,\beta},$

\begin{equation}
\int_{\mathcal{R} \cap S^{\epsilon_n}} \{\mathbb{P}(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\}\} d\tilde{P}(x) = \int_{\mathcal{R} \cap S^{\epsilon_n}} \{\mathbb{P}(S_n(x_0^t) < 1/2) - 1\{t < 0\}\} dt d\text{Vol}^{d-1}(x_0)\{1 + o(1)\};
\end{equation}

\begin{equation}
\int_{\mathcal{R} \cap S^{\epsilon_n}} \{\mathbb{P}^2(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\}\} d\tilde{P}(x) = \int_{\mathcal{R} \cap S^{\epsilon_n}} \{\mathbb{P}^2(S_n(x_0^t) < 1/2) - 1\{t < 0\}\} dt d\text{Vol}^{d-1}(x_0)\{1 + o(1)\}.
\end{equation}

According to Samworth (2012), 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

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

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 (Gray, 2004) implies that, uniformly for $w_n \in W_{n,\beta},$

\begin{equation}
\int_{S^n} \{\mathbb{P}(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\}\} d\tilde{P}(x) = \int_{S^n} \{\mathbb{P}(S_n(x_0^t) < 1/2) - 1\{t < 0\}\} dt d\text{Vol}^{d-1}(x_0)\{1 + o(1)\}.
\end{equation}

Furthermore, we can replace $S^{\epsilon_n}$ with $\mathcal{R} \cap S^{\epsilon_n}$ since $S^{\epsilon_n} \cap \mathcal{R} \subseteq \{x \in \mathbb{R}^d : \text{dist}(x, \partial S) < \epsilon_n\}$ and the latter has volume $O(\epsilon_n^2)$ by Weyl’s tube formula (Gray, 2004). Similarly, we can safely replace $S^{\epsilon_n} \cap \mathcal{S}$ with $\mathcal{S}$. Therefore, (A.9) holds. Similar arguments imply (A.10).

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

\begin{equation}
\sup_{w_n \in W_{n,\beta}} \int_{\mathcal{R} \setminus S^{\epsilon_n}} \{\mathbb{P}(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\}\} d\tilde{P}(x) = O(n^{-M}),
\end{equation}

\begin{equation}
\sup_{w_n \in W_{n,\beta}} \int_{\mathcal{R} \setminus S^{\epsilon_n}} \{\mathbb{P}^2(S_n(x) < 1/2) - 1\{\eta(x) < 1/2\}\} d\tilde{P}(x) = O(n^{-M}).
\end{equation}
Here (A.11) follows from the fact \( |\mathbb{P}(S_n(x) < \frac{1}{2}) - \mathbb{I}\{\eta(x) < 1/2\}| = O(n^{-M}) \) for all \( M > 0 \), uniformly for \( w_n \in W_{n, \beta} \) and \( x \in \mathcal{R} \setminus \mathcal{S}^{\varepsilon_n} \) (Samworth, 2012). Furthermore, (A.12) holds since
\[
\left| \mathbb{P}(S_n(x) < 1/2) - \mathbb{I}\{\eta(x) < 1/2\}\right| \leq 2 \mathbb{P}\left(S_n(x) < 1/2\right) - \mathbb{I}\{\eta(x) < 1/2\}. \tag{A.13}
\]

**Step 3:** In the end, we will show
\[
\int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \mathbb{P}\left(S_n(x_0^i) < 1/2\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) \]
\[= \int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \mathbb{P}\left(S_n(x_0^i) < 1/2\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) + B_3 s_n + o(s_n + t_n). \tag{A.14}
\]

We first apply the nonuniform version of Berry-Esseen Theorem to approximate \( \mathbb{P}(S_n(x_0^i) < 1/2) \). Let \( Z_i = (w_n \mathbb{I}\{Y(i) = 1\} - w_n \mathbb{E}\mathbb{I}\{Y(i) = 1\})/\sigma_n(x) \) and \( W = \sum_{i=1}^{n} Z_i \). Note that \( Z_i \)'s are independent, \( \mathbb{E}(Z_i) = 0, \mathbb{V}(Z_i) < \infty \), and \( \mathbb{V}(W) = 1 \). Then the nonuniform Berry-Esseen Theorem implies that
\[
\left| \mathbb{P}(W \leq y) - \Phi(y) \right| \leq \frac{M_1}{n^{1/2}(1 + |y|^3)},
\]
where \( \Phi \) is the standard normal distribution function and \( M_1 \) is a constant. Therefore,
\[
\mathbb{P}\left(\frac{S_n(x_0^i) - \mu_n(x_0^i)}{\sigma_n(x_0^i)} \leq y\right) - \Phi(y) \leq \frac{M_1}{n^{1/2}(1 + |y|^3)}.
\]

Thus, we have
\[
\int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \mathbb{P}\left(S_n(x_0^i) < 1/2\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) \]
\[= \int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \Phi\left(\frac{1/2 - \mu_n(x_0^i)}{\sigma_n(x_0^i)} - \mathbb{I}\{t < 0\}\right) \right\} dt \text{Vol}^{d-1}(x_0) + o(s_n^2 + t_n^2),
\]
where the remainder term \( o(s_n^2 + t_n^2) \) is due to (A.14) by slightly modifying the proof of A.21 in Samworth (2012).

Furthermore, Taylor expansion leads to
\[
\tilde{f}(x_0) = f(x_0) + (\hat{f}(x_0))^T \frac{\hat{\eta}(x_0)}{\|\hat{\eta}(x_0)\|} t + o(t).
\]
Therefore,
\[
\int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \mathbb{P}\left(S_n(x_0^i) < 1/2\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) \]
\[= \int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} f(x_0) \left\{ \Phi\left(\frac{-2t\|\hat{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) \]
\[+ \int_{-\varepsilon_n}^{\varepsilon_n} \int_{\mathcal{S}} \frac{\hat{f}(x_0)^T \hat{\eta}(x_0) t}{\|\hat{\eta}(x_0)\|} \left\{ \Phi\left(\frac{-2t\|\hat{\eta}(x_0)\| - 2a(x_0)t_n}{s_n}\right) - \mathbb{I}\{t < 0\}\right\} dt \text{Vol}^{d-1}(x_0) + R_1,
\]
The inequality above, along with with

\[ \text{OPTIMAL STABILITY} \]

\( R_1 = \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0) \left\{ \Phi \left( \frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \right) - \Phi \left( \frac{-2t||\dot{\eta}(x_0)|| - 2a(x_0)t_n}{s_n} \right) \right\} dtdVol^{d-1}(x_0) \)

\[ + \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0)^T \dot{\eta}(x_0)t \left\{ \Phi \left( \frac{1/2 - \mu_n(x_0)}{\sigma_n(x_0)} \right) - \Phi \left( \frac{-2t||\dot{\eta}(x_0)|| - 2a(x_0)t_n}{s_n} \right) \right\} dtdVol^{d-1}(x_0) \]

\[ + o(s_n^2 + t_n^2) \]

\[ \Delta R_{11} + R_{12} + o(s_n^2 + t_n^2). \]

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 (A.7) and (A.8), for a sufficiently small \( \epsilon \in (0, \inf_{x_0 \in \mathcal{S}} ||\dot{\eta}(x_0)||) \) and large \( n, \) for all \( w_n \in W_{\nu, \beta}, \) \( x_0 \in \mathcal{S} \) and \( r \in [-\epsilon_n/s_n, \epsilon_n/s_n], \) Samworth (2012) showed that

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

In addition, when \( |r - r_{x_0}| \leq \epsilon t_n/s_n, \)

\[ \left| \Phi \left( \frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})} \right) - \Phi \left( -2||\dot{\eta}(x_0)|| (r - r_{x_0}) \right) \right| \leq 1 \]

and when \( \epsilon t_n/s_n < |r| < t_n/s_n, \)

\[ \left| \Phi \left( \frac{1/2 - \mu_n(x_0^{rs_n})}{\sigma_n(x_0^{rs_n})} \right) - \Phi \left( -2||\dot{\eta}(x_0)|| (r - r_{x_0}) \right) \right| \leq \epsilon^2 (|r| + t_n/s_n) \phi(||\dot{\eta}(x_0)|| |r - r_{x_0}|), \]

where \( \phi \) is the density function of standard normal distribution.

Therefore, we have

\[ |R_{11}| \leq \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0) \left| \Phi \left( \frac{1/2 - \mu_n(x_0^t)}{\sigma_n(x_0^t)} \right) - \Phi \left( \frac{-2t||\dot{\eta}(x_0)|| - 2a(x_0)t_n}{s_n} \right) \right| dtdVol^{d-1}(x_0) \]

\[ \leq \tilde{f}(x_0)s_n \int_{|r - r_{x_0}| \leq \epsilon t_n/s_n} dr + \tilde{f}(x_0)s_n \epsilon^2 \int_{-\infty}^{\infty} (|r| + t_n/s_n) \phi(||\dot{\eta}(x_0)|| |r - r_{x_0}|) dr \]

(A.16) \[ \leq \epsilon (t_n + s_n). \]

Similarly,

\[ |R_{12}| \leq \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0)^T \dot{\eta}(x_0)t \left| \Phi \left( \frac{1/2 - \mu_n(x_0)}{\sigma_n(x_0)} \right) - \Phi \left( \frac{-2t||\dot{\eta}(x_0)|| - 2a(x_0)t_n}{s_n} \right) \right| dtdVol^{d-1}(x_0) \]

\[ \leq \tilde{f}(x_0)s_n^2 \int_{|r - r_{x_0}| \leq \epsilon t_n/s_n} r dr + \tilde{f}(x_0)s_n^2 \epsilon^2 \int_{-\infty}^{\infty} (|r| + t_n/s_n) \phi(||\dot{\eta}(x_0)|| |r - r_{x_0}|) dr \]

\[ \leq \epsilon (t_n^2 + s_n^2). \]

The inequality above, along with with (A.16), leads to \( R_1 = o(s_n + t_n). \)
By similar arguments, we have

\[(A.17) \quad \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0) \left\{ \Phi^2 \left( \frac{-2t\|\tilde{\eta}(x_0)\| - 2a(x_0)t_n}{s_n} \right) - 1 \{ t < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ = \int_S \int_{-\epsilon_n}^{\epsilon_n} \tilde{f}(x_0) \left\{ \Phi^2 \left( -2t\|\tilde{\eta}(x_0)\| - 2a(x_0)t_n \right) s_n - 1 \{ t < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ + \int_S \int_{-\epsilon_n}^{\epsilon_n} \hat{f}(x_0) \tilde{\eta}(x_0)t \left\{ \Phi^2 \left( -2t\|\tilde{\eta}(x_0)\| - 2a(x_0)t_n \right) s_n - 1 \{ t < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ + o(s_n + t_n). \]

Finally, after substituting \( t = us_n/2 \) in (A.15) and (A.17), we have, up to \( o(s_n + t_n) \) difference,

\[ \text{CIS}(\text{wnn})/2 \]

\[ = \frac{s_n}{2} \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) \left\{ \Phi \left( -\|\tilde{\eta}(x_0)\|u - \frac{2a(x_0)t_n}{s_n} \right) - 1 \{ u < 0 \} \right\} du dVol^{d-1}(x_0) \]

\[ + \frac{s_n^2}{4} \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) \tilde{\eta}(x_0) \left\{ \Phi \left( -\|\tilde{\eta}(x_0)\|u - \frac{2a(x_0)t_n}{s_n} \right) - 1 \{ u < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ - \frac{s_n}{2} \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) \left\{ \Phi^2 \left( -\|\tilde{\eta}(x_0)\|u - \frac{2a(x_0)t_n}{s_n} \right) - 1 \{ u < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ - \frac{s_n^2}{4} \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) \tilde{\eta}(x_0) \left\{ \Phi^2 \left( -\|\tilde{\eta}(x_0)\|u - \frac{2a(x_0)t_n}{s_n} \right) - 1 \{ u < 0 \} \right\} dt dVol^{d-1}(x_0) \]

\[ = I + II - III - IV. \]

According to Lemma A.2, we have

\[ I - III = \left[ \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) dVol^{d-1}(x_0) \right] s_n = \frac{1}{2} B_3 s_n \]

\[ II - IV = -\left[ \int_S \int_{-\infty}^{\infty} \tilde{f}(x_0) \tilde{\eta}(x_0) a(x_0) \right] dVol^{d-1}(x_0) s_n t_n = \frac{1}{2} B_4 s_n t_n. \]

Therefore, the desirable result is obtained by noting that \( B_4 s_n t_n = o(s_n + t_n) \). This concludes the proof of Theorem 3. \( \blacksquare \)

A.VII. A Lemma for Proving Theorem 4 (Optimal Weight).

**Lemma A.3.** Given \( \alpha_i = i^{1+2/d} - (i - 1)^{1+2/d} \), we have

\[(A.18) \quad (1 + \frac{2}{d})(i - 1)^{\frac{2}{d}} \leq \alpha_i \leq (1 + \frac{2}{d})i^{\frac{2}{d}}, \]

\[(A.19) \quad \sum_{j=1}^{k} \alpha_j^2 = \frac{(d + 2)^2}{d(d + 4)} k^{1+4/d} \left\{ 1 + O\left( \frac{1}{k} \right) \right\}. \]
Proof of Lemma A.3: First, (A.18) is a direct result from the following two inequalities.

\[
(1 - \frac{1}{i})^{2/d} \geq 1 - \frac{2}{(i-1)d} \quad \text{and} \quad (1 + \frac{1}{i-1})^{2/d} \geq 1 + \frac{2}{id},
\]

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, (A.19) is due to (A.18) and Faulhaber’s formula \(\sum_{k=1}^{n} i^p = \frac{1}{p+1} k^{p+1} + O(k^{p}).\) According to (A.18), we have

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

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 (A.19). This concludes the proof of Lemma A.3.

A.VIII. Proof of Theorem 4. For any weight \(w_n\), the Lagrangian of (3.5) is

\[
L(w_n) = \left( \sum_{i=1}^{n} \frac{\alpha_i w_{ni}}{n^{2/d}} \right)^2 + \lambda \sum_{i=1}^{n} w_{ni}^2 + \nu \left( \sum_{i=1}^{n} w_{ni} - 1 \right).
\]

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

\[
\frac{\partial L(w_n)}{\partial w_{ni}} = 2n^{-4/d} \alpha_i \sum_{k=1}^{k^*} \alpha_i w_{ni} + 2\lambda w_{ni} + \nu = 0.
\]

Summing (A.20) from 1 to \(k^*\), and multiplying (A.20) by \(\alpha_i\) and then summing from 1 to \(k^*\) yields

\[
2n^{-4/d} k^* (k^* + 1)^{1+2/d} \sum_{i=1}^{k^*} \alpha_i w_{ni} + 2\lambda k^* + \nu k^* = 0
\]

\[
2n^{-4/d} \sum_{i=1}^{k^*} \alpha_i w_{ni} \sum_{i=1}^{k^*} \alpha_i^2 + 2\lambda \sum_{i=1}^{k^*} \alpha_i w_{ni} + \nu (k^* + 1)^{1+2/d} = 0.
\]

Therefore, we have

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

Here \(w_{ni}^*\) is decreasing in \(i\) since \(\alpha_i\) is increasing in \(i\) and \(\sum_{i=1}^{k^*} \alpha_i^2 > (k^*)^{1+4/d}\) from Lemma A.3.

Next we solve for \(k^*\). According to the definition of \(k^*\), we only need to find \(k\) such that \(w_{ni}^* = 0\).

Using the results from Lemma A.3, solving this equation reduces to solving \(k^*\) such that

\[
(1 + \frac{2}{d})(k^* - 1)^{2/d} \leq \lambda n^{4/d} (k^*)^{-1+2/d} + \frac{(d+2)^2}{(d+4)} (k^*)^{2/d} \left( 1 + O\left( \frac{1}{k^*} \right) \right) \leq (1 + \frac{2}{d})(k^*)^{2/d}.
\]

Therefore, for large \(n\), we have

\[
k^* = \left\lfloor \frac{d(d+4)}{2(d+2)} \right\rfloor \frac{\lambda n^{4/d}}{\pi^2} \frac{1}{\alpha_i} \frac{1}{n \alpha_i}.
\]

Plugging \(k^*\) and (A.19) into (A.21) yields the optimal weight.
A.IX. Proof of Theorem 5. According to our Theorem 1 and the proof of Theorem 1 in the supplementary of Samworth (2012), 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 $\tilde{P}$-almost all $x$,

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

(A.22)

where $S_n^*(x) = \sum_{i=1}^{n} w_n^* 1\{Y_i = 1\}$ with the optimal weight $w_n^*$ defined in Theorem 4 and $k^* \asymp n^{2\gamma/(2\gamma + d)}$.

According to Lemma A.3, we have

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

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 Samworth (2012), there exist $C_9, C_{10} > 0$ such that for all $P \in \mathcal{P}_{\alpha, \gamma}$ and $x \in \mathcal{R}$,

$$|\mu_n^*(x) - \eta(x)| \leq \sum_{i=1}^{n} w_n^* \mathbb{E}\{\eta(X_{(i)}) - \eta_x(X_{(i)})\} + \sum_{i=1}^{n} w_n^* \mathbb{E}\{\eta_x(X_{(i)})\} - \eta(x) \leq L \sum_{i=1}^{n} w_n^* \mathbb{E}\{\|X_{(i)} - x\|^\gamma\} + \sum_{i=1}^{n} w_n^* \mathbb{E}\{\eta_x(X_{(i)})\} - \eta(x) \leq C_9 \sum_{i=1}^{n} w_n^* \left(\frac{i}{n}\right)^{\gamma/d} \leq C_{10} n^{-\gamma/(2\gamma + d)}.\quad (A.23)$$

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

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

Let $Z_i = w_n^* 1\{Y_i = 1\}$ with $a_i = 0$ and $b_i = w_n^*$. According to (A.23), we have that for $\delta \geq 2C_{10} n^{-\gamma/(2\gamma + d)}$ and for $\tilde{P}$-almost all $x$,

$$\sup_{P \in \mathcal{P}_{\alpha, \gamma}} \mathbb{P}_D \left( |S_n^*(x) - \eta(x)| \geq \delta \right) \leq \sup_{P \in \mathcal{P}_{\alpha, \gamma}} \mathbb{P}_D \left( |S_n^*(x) - \mu_n^*(x)| \geq \delta/2 \right) \leq 2 \exp\left\{-n^{2\gamma/(2\gamma + d)} \frac{\delta^2}{(2C_8)}\right\},$$

which implies (A.22) directly.
A.X. **Proof of Corollary 1.** According to Theorems 1 and 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
\[
\sup_{P \in \mathcal{P}_{\alpha, \gamma}} \left\{ \text{CIS}(snn) - \text{CIS}(ownn) \right\}
\geq \sup_{P \in \mathcal{P}_{\alpha, \gamma}} \text{CIS}(snn) - \sup_{P \in \mathcal{P}_{\alpha, \gamma}} \text{CIS}(ownn)
\geq C_{11} n^{-\alpha\gamma/(2\gamma+d)},
\]
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
\[
\sup_{P \in \mathcal{P}_{\alpha, \gamma}} \left\{ \text{CIS}(snn) - \text{CIS}(ownn) \right\}
\leq \sup_{P \in \mathcal{P}_{\alpha, \gamma}} \text{CIS}(snn) + \sup_{P \in \mathcal{P}_{\alpha, \gamma}} \text{CIS}(ownn)
\leq C_{12} n^{-\alpha\gamma/(2\gamma+d)},
\]
for some constant $C_{12} > 0$.

Furthermore, according to Theorem 5, 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
\[
\sup_{P \in \mathcal{P}_{\alpha, \gamma}} \left\{ \text{Regret}(snn) - \text{Regret}(ownn) \right\} \asymp n^{-\gamma(1+\alpha)/(2\gamma+d)}.
\]
This concludes the proof of Corollary 1. \hfill \blacksquare

A.XI. **Proof of Corollaries 2 and 3.** For the ownn classifier, the optimal $k^{**}$ is a function of
$k^{opt}$ of $k$-nearest neighbor classifier (Samworth, 2012). Specifically,
\[
k^{**} = \left\lfloor \frac{2(d+2)}{d+2} \sum_{i=1}^{k^*} w_{ni}^* \right\rfloor_{k^{opt}}.
\]
According to Theorem 4 and Lemma A.3, we have
\[
\sum_{i=1}^{k^*} (w_{ni}^*)^2 = \frac{2(d+2)}{(d+4)k^*} \{ 1 + O((k^*)^{-1}) \}.
\]
Therefore,
\[
\frac{\text{CIS}(ownn)}{\text{CIS}(knn)} \rightarrow 2^{2/(d+4)} \left( \frac{d+2}{d+4} \right)^{(d+2)/(d+4)}.
\]
Furthermore, for large $n$,
\[
\frac{\text{CIS}(snn)}{\text{CIS}(ownn)} = \frac{B_3 \left( \sum_{i=1}^{k^*} w_{ni}^* \right)^{1/2}}{B_3 \left( \sum_{i=1}^{k^{**}} w_{ni}^{**} \right)^{1/2}} = \left\{ \frac{B_1}{\lambda B_2} \right\}^{d/(2(d+4))}.
\]
The rest limit expressions in Corollaries 2 and 3 can be shown in similar manners. \hfill \blacksquare
A.XII. Calculation of (5.3). According to the definition,
\[ B_1 = \int_S \frac{\bar{f}(x_0)}{4\|\bar{\eta}(x_0)\|} d\text{Vol}^{d-1}(x_0). \]

When \( f_1 = N(0_d, I_d) \) and \( f_2 = N(\mu, I_d) \) with the prior probability \( \pi_0 = 1/3 \), we have
\[ \bar{f}(x_0) = \pi_0 f_1 + (1 - \pi_0) f_2 = 2(2\pi)^{-d/2} \exp\{-x_0^T x_0/2\}/3, \]
and
\[ \eta(x) = \frac{\pi_0 f_1}{\pi_0 f_1 + (1 - \pi_0) f_2} = \left(1 + 2 \exp\{\mu^T x - \mu^T \mu/2\}\right)^{-1}. \]

Hence, the decision boundary is
\[ S = \{x \in \mathcal{R} : \eta(x) = 1/2\} = \{x \in \mathcal{R} : 1^T_d x = (\mu d)/2 - (\ln 2)/\mu\}, \]
where \( 1_d \) is a \( d \)-dimensional vector of all elements 1.

Therefore, for \( x_0 \in S \), we have \( \bar{\eta}(x_0) = -\mu/4 \) and hence
\[ B_1 = \frac{2}{3\mu(2\pi)^{d/2}\sqrt{d}} \int_S \exp\{-x_0^T x_0/2\} d\text{Vol}^{d-1}(x_0). \]
\[ = \frac{\sqrt{2\pi}}{3\pi \mu d} \exp\left\{-\frac{(\mu d/2 - \ln 2/\mu)^2}{2d}\right\}. \]

A.XIII. Relative Gain of SNN Over OWN. Define the relative gain of snn over ownn as the absolute ratio of the percentages of CIS reduction and regret increment, \( i.e., |\Delta\text{CIS}/\Delta\text{Regret}| \), where
\[ \Delta\text{CIS} = \frac{\text{CIS}(\text{snn}) - \text{CIS}(\text{ownn})}{\text{CIS}(\text{ownn})} \quad \text{and} \quad \Delta\text{Regret} = \frac{\text{Regret}(\text{snn}) - \text{Regret}(\text{ownn})}{\text{Regret}(\text{ownn})}. \]

According to (4.1) and (4.2) with \( \lambda_0 = 1 \), we have
\[ (A.24) \quad \text{Relative Gain} \rightarrow \frac{1 - \{1 + 16B_1/\pi\}^{d/(2d+8)}}{\{1 + 16B_1/\pi\}^{4/(d+4)} - 1}. \]

Figure 7 shows the logarithm of the relative gain of snn over ownn as a function of \( B_1 \) and \( d \). Clearly, in most combinations of \( B_1 \) and \( d \), the logarithm of relative gain is greater than 0, which means that snn’s improvement in CIS is larger than its loss in regret. For example, when \( B_1 \leq 0.3 \), the logarithm of relative gain is larger than 0 for any \( d \); when \( B_1 = 1 \), the logarithm of relative gain is larger than 0 for any \( d > 5 \). This indicates that generally snn is particularly effective in trading relatively small accuracy for stability.
Fig 7. Log of relative gain of snn over own as a function of $B_1$ and $d$.

References.

Audibert, J. (2004). Classification under Polynomial Entropy and Margin Assumptions and Randomized Estimators. Preprint 905, Laboratoire de Probabilites et Modeles Aleatoires, Univ. Paris VI and VII.

Audibert, J. and Tsybakov, A. (2007). Fast Learning Rates for Plug-in Classifiers. Annals of Statistics, 35, 608–633.

Ben-Hur, A., Eliseeff, A., and Guyon, I. (2002). A Stability Based Method for Discovering Structure in Clustered Data. Pacific Symposium on Biocomputing, 6–17.

Biau, G., Cérou, F., and Guyader, A. (2010). On the Rate of Convergence of the Bagged Nearest neighbor Estimate. Journal of Machine Learning Research, 11, 687–712.

Bousquet, O. and Elisseeff, A. Stability and Generalization. Journal of Machine Learning Research, 2, 499-526.

Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection. Annals of Statistics, 24, 2350–2383.

Bühlmann, P. and Yu, B. (2002). Analyzing Bagging. Annals of Statistics, 30, 927–961.

Cover, T. M. and Hart, P. E. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13, 21–27.

Devroye, L., Györfi, L., Krzyak, A. and Lugosi, G. (1994). On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. Annals of Statistics, 22, 1371–1385.

Devroye, L., Györfi, L., and Lugosi, G. (1996). A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.

Devroye, L. and Wagner, T. J. (1977). The Strong Uniform Consistency of Nearest Neighbor Density Estimates. Annals of Statistics, 5, 536–540.

Elisseeff, A., Evgeniou, T., and Pontil, M. (2005). Stability for Randomized Learning Algorithms. Journal of Machine Learning Research, 6, 55–79.

Fix, E. and Hodges, J. L., Jr. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Randolph Field, Texas, Project 21-49-004, Report No.4.

Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.
Gray, A. (2004). Tubes. *Progress in Mathematics, 221*, Birkhäuser, Basel.

Györfi, L. (1981). The Rate of Convergence of k-NN Regression Estimates and Classification Rules. *IEEE Transactions on Information Theory, 27*, 362–364.

Hall, P. and Kang, K. (2005). Bandwidth Choice for Nonparametric Classification. *Annals of Statistics, 33*, 284–306.

Hall, P., Park, B., and Samworth, R. (2008). Choice of Neighbor Order in Nearest Neighbor Classification. *Annals of Statistics, 36*, 2135–2152.

Hall, P. and Samworth, R. (2005). Properties of Bagged Nearest neighbor Classifiers. *Journal of the Royal Statistical Society, Series B, 67*, 363–379.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High-Dim Graphical Models. *Advances in Neural Information Processing Systems, 23*.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society, Series B, 72*, 414–473.

Samworth, R. (2012). Optimal Weighted Nearest neighbor Classifiers. *Annals of Statistics, 40*, 2733–2763.

Shah, R. and Samworth, R. (2013). Variable Selection with Error Control: Another Look at Stability Selection. *Journal of the Royal Statistical Society, Series B, 75*, 55–80.

Snapp, R. R. and Venkatesh, S. S. (1998). Asymptotic Expansion of the K Nearest Neighbor Risk. *Annals of Statistics, 26*, 850–878.

Stone, C. J. (1977). Consistent Nonparametric Regression. *Annals of Statistics, 5*, 595–645.

Tsybakov, A. (2004). Optimal Aggregation of Classifiers in Statistical Learning. *Annals of Statistics, 32*, 135–166.

Wang, J. (2010). Consistent Selection of the Number of Clusters via Cross Validation. *Biometrika, 97*, 893–904.

Yang, Y. (2004). Consistency of Cross Validation for Comparing Regression Procedures. *Annals of Statistics, 35*, 2450–2473.

Yu, B. (2013). Stability. *Bernoulli, 19*, 1484–1500.