

---

# On Some Fast And Robust Classifiers For High Dimension, Low Sample Size Data

---

**Sarbojit Roy**

Indian Institute of Technology  
Kanpur, India  
sarbojit@iitk.ac.in

**Jyotishka Ray Choudhury**

Indian Statistical Institute  
Kolkata, India  
bs1903@isical.ac.in

**Subhajit Dutta**

Indian Institute of Technology  
Kanpur, India  
duttas@iitk.ac.in

## Abstract

In high dimension, low sample size (HDLSS) settings, *distance concentration* phenomena affects the performance of several popular classifiers which are based on Euclidean distances. The behaviour of these classifiers in high dimensions is completely governed by the first and second order moments of the underlying class distributions. Moreover, the classifiers become absolutely useless for such HDLSS data when the first two moments of the competing distributions are equal, or when the moments do not exist. In this work, we propose robust, computationally efficient and tuning-free classifiers applicable in HDLSS scenarios. As the data dimension increases, these classifiers yield *perfect classification* if the one-dimensional marginals of the underlying distributions are different. We also establish strong theoretical properties for the proposed classifiers in *ultra-high-dimensional* settings. Numerical experiments with a wide variety of simulated examples and analysis of real data exhibit clear and convincing advantages over existing methods.

## 1 INTRODUCTION

Let us consider a classification problem involving two distribution functions  $\mathbf{F}_1$  and  $\mathbf{F}_2$  on  $\mathbb{R}^p$  with  $p \geq 1$ . Suppose  $\chi_1 = \{\mathbf{X}_1, \dots, \mathbf{X}_{n_1}\}$  and  $\chi_2 =$

$\{\mathbf{Y}_1, \dots, \mathbf{Y}_{n_2}\}$  are random samples drawn from  $\mathbf{F}_1$  and  $\mathbf{F}_2$ , respectively and  $\chi = \chi_1 \cup \chi_2$  is the training sample of size  $n = n_1 + n_2$ . The prior probability of  $j$ -th class is given by  $\pi_j > 0$  for  $j = 1, 2$  with  $\pi_1 + \pi_2 = 1$ . Using the training sample, a classifier  $\delta$  assigns a new point  $\mathbf{z} \in \mathbb{R}^p$  to one of the two competing classes. We develop classifiers that yield *perfect classification* under fairly general conditions in high dimension, low sample size (HDLSS) settings, where the sample size  $n$  remains fixed, but the dimension  $p$  increases. The classifier  $\delta$  is said to yield *perfect classification* in HDLSS settings if the misclassification probability of  $\delta$  goes to 0 as  $p \rightarrow \infty$ .

In the classical setting,  $p$  is fixed and  $n \rightarrow \infty$ . Information is accumulated as more samples are collected.

In the HDLSS setting,  $n$  is fixed while  $p \rightarrow \infty$ . Information is accumulated as more features are measured.

### 1.1 Literature Review

In the HDLSS asymptotic regime, Euclidean distance (ED) based classifiers face some natural drawbacks due to *distance concentration* (Aggarwal et al., 2001; Francois et al., 2007). To give a mathematical exposition of this fact, let  $\boldsymbol{\mu}_j$  and  $\Sigma_j$  denote the mean vector and the covariance matrix of  $\mathbf{F}_j$  for  $j = 1, 2$ . We assume that the following limits exist:

$$\begin{aligned} \nu^2 &= \lim_{p \rightarrow \infty} \frac{1}{p} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 \text{ and} \\ \sigma_j^2 &= \lim_{p \rightarrow \infty} \frac{1}{p} \text{tr}(\Sigma_j) \text{ for } j = 1, 2. \end{aligned} \quad (1.1)$$

Here,  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^p$  and  $\text{tr}(M)$  denotes the trace of a matrix  $M$ . The constants  $\nu^2$  and  $|\sigma_1^2 - \sigma_2^2|$  can be interpreted as asymp-

totic measures of the difference between locations and scales of  $\mathbf{F}_1$  and  $\mathbf{F}_2$ , respectively. Hall et al. (2005) studied the consequence of distance concentration on some popular ED based classifiers such as the 1-nearest neighbor (1NN) classifier (Hastie et al., 2009), average distance (AVG) classifier (Chan and Hall, 2009b) and support vector machines (SVM) (Vapnik, 1998). The authors showed that in high dimensions, these methods are incapable of correctly classifying an observation if the location difference between the competing populations gets masked by their difference in scales, i.e.,  $\nu^2 < |\sigma_1^2 - \sigma_2^2|$ . Chan and Hall (2009b); Dutta and Ghosh (2016) proposed some improved classifiers that yield *perfect classification* if  $\nu^2 > 0$ , or  $\sigma_1^2 \neq \sigma_2^2$ . However, these improved methods fail in high dimensions when the competing populations have same location and scale, i.e.,  $\nu^2 = 0$  and  $\sigma_1^2 = \sigma_2^2$ , or when  $\nu^2, \sigma_1^2$  and  $\sigma_2^2$  do not exist. The limitations of these methods stem from the fact that they are based on Euclidean distances, and the behavior of ED in the HDLSS asymptotic regime is completely governed by these constants. As a result, ED based classifiers cannot distinguish between populations that do not have differences in their first two moments. On top of that, these classifiers lack robustness since ED is sensitive to outliers. Chan and Hall (2009a) proposed a robust version of the NN classifier for high-dimensional data, but it is only applicable to a specific type of two-class location problem. Other approaches for classifying high-dimensional data include Globerson and Roweis (2005); Tomašev et al. (2014); Weinberger and Saul (2009). A recent work by Thrampoulidis (2020) discusses the high-dimensional behavior of several classifiers, but under an additional assumption of Gaussianity.

## 1.2 Motivation

Li and Zhang (2020) proposed a method for testing equality of two distributions, where the authors considered a new measure of distance between  $\mathbf{F}_1$  and  $\mathbf{F}_2$  as defined below:

$$\tau = E[h(\mathbf{X}_1, \mathbf{X}_2) + h(\mathbf{Y}_1, \mathbf{Y}_2) - 2h(\mathbf{X}_1, \mathbf{Y}_1)].$$

Here,  $h : \mathbb{R}^p \times \mathbb{R}^p \rightarrow [-1, 1]$  is given by

$$h(\mathbf{u}, \mathbf{v}) = \frac{1}{2\pi} \sin^{-1} \left( \frac{1 + \mathbf{u}^\top \mathbf{v}}{[(1 + \|\mathbf{u}\|^2)(1 + \|\mathbf{v}\|^2)]^{\frac{1}{2}}} \right)$$

for  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$  with  $p \geq 1$ . The authors showed that for a fixed  $p$ ,  $\tau = 0$  iff  $\mathbf{F}_1 = \mathbf{F}_2$ . This property of  $\tau$  is useful for distinguishing one distribution from another, and can be utilized in classification problems

as well. However, a classifier that directly utilizes  $\tau$ , faces certain challenges in the HDLSS setting.

To motivate the problem, we modify the scale-adjusted average distance (SAVG) classifier (Chan and Hall, 2009b) by simply replacing the squared Euclidean norm  $\|\mathbf{u} - \mathbf{v}\|^2$  with  $h(\mathbf{u}, \mathbf{v})$  defined above. A formal definition of this modified classifier (henceforth referred to as  $\delta_0$ ) is given in Section 2 where we also discuss how  $\delta_0$  uses  $\tau$  to classify an observation.

Let us now consider the following examples:

**Example 1**  $X_{1k} \stackrel{i.i.d.}{\sim} N(1, 1)$  and  $Y_{1k} \stackrel{i.i.d.}{\sim} N(1, 2)$ ,

**Example 2**  $X_{1k} \stackrel{i.i.d.}{\sim} N(0, 3)$  and  $Y_{1k} \stackrel{i.i.d.}{\sim} t_3$ ,

for  $1 \leq k \leq p$ . Here,  $N(\mu, \sigma^2)$  denotes the univariate Gaussian distribution with mean  $\mu \in \mathbb{R}$  and standard deviation  $\sigma(> 0)$ , and  $t_\kappa$  denotes the standard Student's  $t$  distribution with  $\kappa(> 0)$  degrees of freedom. In Figure 1, we compare the performance of the classifier  $\delta_0$  with some popular classifiers like 1NN, the usual SAVG, SVM with the linear kernel (SVM-LIN) and SVM with the radial basis function (SVM-RBF) kernel. Detail of the simulation study is given in Section 4.

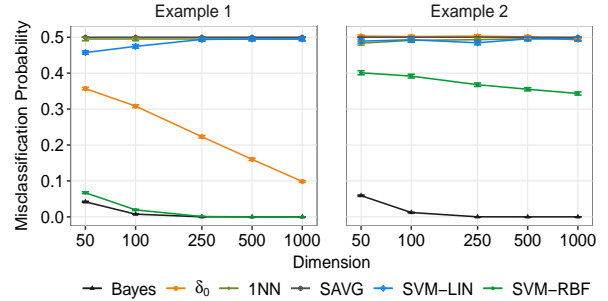


Figure 1: Average Misclassification Rates (along with Standard Errors) of  $\delta_0$  and Some Popular Classifiers. Dimension is in Logarithmic Scale.

In the first example,  $\nu^2 = 0$  (since  $\mu_1 = \mu_2 = 1_p$ ) but  $\sigma_1^2 = 1$  and  $\sigma_2^2 = 2$ . The classifier  $\delta_0$  identifies the difference in scales and yields a moderate performance. Whereas the existing classifiers (except SVM-RBF) misclassify 50% of the observations. SVM-RBF capitalizes on the difference between  $\sigma_1^2$  and  $\sigma_2^2$ , and perfectly classifies the test observations as dimension increases. The problem of classification is more challenging in **Example 2**. Here, we have  $\nu^2 = 0$  (since  $\mu_1 = \mu_2 = \mathbf{0}_p$ ) and  $\sigma_1^2 = \sigma_2^2 = 3$ , i.e., there is no difference between either of the location and scale parameters. As a result,  $\delta_0$  along with existing classifiers fail to correctly classify the test observations. We will revisit these examples again in Sections 3.1.2 and 4.

### 1.3 Our Contribution

In this article, we develop classifiers that are suitable for high dimensional data. The behavior of the proposed classifiers in HDLSS settings do not depend on the existence of the moments. If the one-dimensional marginals of the underlying populations are different, then the proposed classifiers are shown to yield *perfect classification* in HDLSS settings.

The proposed classifiers

- are robust,
- computationally fast,
- free from tuning parameters, and
- have strong theoretical properties.

The rest of the article is organized as follows. In Section 2, we propose a classifier and further modify it to achieve improved classification accuracy under specific conditions. Asymptotic properties of the proposed classifiers are studied in Section 3. A theoretical result is presented in Section 3.1.2 to analyze their relative performances. In Section 3.2, we investigate their behavior when both  $n$  and  $p$  increase. Performance of the classifiers is studied through a numerical exercise involving several simulated data sets in Section 4. We also examine the behavior of the classifiers on real data sets in Section 5. The article ends with concluding remarks in Section 6. All proofs and relevant mathematical details are provided in Supplementary A. Additional details on the numerical performance of the classifiers and a link to related R-codes can be found in Supplementary B.

## 2 METHODOLOGY

Let us recall the classifier  $\delta_0$  mentioned in Section 1.2. Fix  $\mathbf{z} \in \mathbb{R}^p$ . For given random samples  $\chi_1$  and  $\chi_2$  with sizes  $n_1 (\geq 2)$  and  $n_2 (\geq 2)$ , respectively, the classifier  $\delta_0$  is formally defined as

$$\begin{aligned} \delta_0(\mathbf{z}) &= \arg \min_{j \in \{1,2\}} L_j(\mathbf{z}), \text{ where } L_j(\mathbf{z}) = T_{jj} - 2T_j(\mathbf{z}), \\ T_{jj} &= \frac{1}{n_j(n_j - 1)} \sum_{\mathbf{U}, \mathbf{U}' \in \chi_j, \mathbf{U} \neq \mathbf{U}'} h(\mathbf{U}, \mathbf{U}') \text{ and} \\ T_j(\mathbf{z}) &= \frac{1}{n_j} \sum_{\mathbf{U} \in \chi_j} h(\mathbf{U}, \mathbf{z}) \text{ for } j = 1, 2. \end{aligned} \quad (2.1)$$

For a random vector  $\mathbf{Z}$  (independent from  $\chi$ ), the misclassification probability of a classifier  $\delta$  is defined as  $\Delta = P[\delta(\mathbf{Z}) \neq \text{true class label of } \mathbf{Z}]$ . Throughout this article, we will follow this definition

of the misclassification probability. In the previous section, we have introduced the constants  $\nu^2, \sigma_1^2$  and  $\sigma_2^2$ . Now, we define  $\nu_{jj'} = \lim_{p \rightarrow \infty} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'}/p$  for  $j, j' \in \{1, 2\}$ , and further assume the following:

- (i) There exists a constant  $C_0$  such that  $E[|U_k|^4] < C_0 < \infty$  for all  $1 \leq k \leq p$ , where  $\mathbf{U} = (U_1, \dots, U_p)^\top \sim \mathbf{F}_j$  for  $j = 1, 2$ .
- (ii) The constants  $\nu_{jj'}$  and  $\sigma_j^2$  exist for  $j, j' \in \{1, 2\}$ .

Let  $\mathbf{U}$  and  $\mathbf{V}$  be two independent vectors such that  $\mathbf{U} \sim \mathbf{F}_j$  and  $\mathbf{V} \sim \mathbf{F}_{j'}$  for  $j, j' \in \{1, 2\}$ . We also assume that the components of the sequence  $\{U_k V_k, k \geq 1\}$  are weakly dependent. In particular,

- (iii)  $\sum_{1 \leq k < k' \leq p} \text{Corr}(U_k V_k, U_{k'} V_{k'}) = o(p^2)$ .

Assumption (iii) is trivially satisfied if the component variables of the underlying populations are independent. It continues to hold with some additional conditions on their dependence structure. For example, (iii) is satisfied when the sequence  $\{U_k V_k, k \geq 1\}$  has  $\rho$ -mixing property (Bradley, 2005; Hall et al., 2005). Conditions similar to (iii) are frequently considered in the literature for studying high-dimensional behavior of various statistical procedures (Aoshima et al., 2018).

**Lemma 2.1** Suppose assumptions (i)-(iii) are satisfied. Let  $L(\mathbf{Z}) = L_2(\mathbf{Z}) - L_1(\mathbf{Z})$  for a test observation  $\mathbf{Z}$ .

(a) If  $\mathbf{Z} \sim \mathbf{F}_1$ , then  $|L(\mathbf{Z}) - \tau| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ .

(b) If  $\mathbf{Z} \sim \mathbf{F}_2$ , then  $|L(\mathbf{Z}) + \tau| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ .

Lemma 2.1 states that if  $\mathbf{Z} \sim \mathbf{F}_1$  ( $\mathbf{F}_2$ , respectively), then the discriminant of  $\delta_0$  converges in probability to  $\tau$ , a positive (negative, respectively) quantity as  $p \rightarrow \infty$ . Let  $\Delta_0$  denote the misclassification probability of the classifier  $\delta_0$ . The following theorem shows that in HDLSS settings, the asymptotic behavior of  $\delta_0$  is governed by the constants  $\nu_{jj'}$ , and  $\sigma_j^2$  for  $j, j' \in \{1, 2\}$ .

**Theorem 2.2** Suppose that assumptions (i)-(iii) are satisfied, and either of the following conditions hold:

- (a)  $\nu_{11}, \nu_{12}$  and  $\nu_{22}$  are unequal,
- (b)  $\nu_{11} = \nu_{12} = \nu_{22} \neq 0$  and  $\sigma_1^2 \neq \sigma_2^2$ .

For any  $\pi_1 > 0$ ,  $\Delta_0 \rightarrow 0$  as  $p \rightarrow \infty$ .

Theorem 2.2 shows that if  $\mathbf{F}_1$  and  $\mathbf{F}_2$  differ either in their locations and/or scales, then  $\Delta_0$  converges to 0 as dimension increases. Recall **Example 1**, and note that condition (b) of Theorem 2.2 is satisfied in this example since  $|\sigma_1^2 - \sigma_2^2| = 1$ . In **Example 2**, both (a) and (b) are violated and Theorem 2.2 fails to hold. This gives us a clear explanation why the classifier  $\delta_0$  performed well in the first example, but failed in the second one (see Figure 1). Now, we develop classifiers whose asymptotic properties are not governed by the constants  $\nu_{jj'}$ , and  $\sigma_j^2$  for  $j, j' \in \{1, 2\}$ . The proposed classifiers use differences between the one-dimensional marginals of  $\mathbf{F}_1$  and  $\mathbf{F}_2$ , and attain *perfect classification* in high dimensions under fairly general conditions.

## 2.1 A New Measure of Distance

Let  $F_{j,k}$  denote the distribution of the random variable  $U_k$ , where  $\mathbf{U} = (U_1, \dots, U_p)^\top \sim \mathbf{F}_j$  for  $j = 1, 2$  and  $1 \leq k \leq p$ . Suppose,  $\mathbf{X}_1, \mathbf{X}_2 \stackrel{i.i.d.}{\sim} \mathbf{F}_1$  and  $\mathbf{Y}_1, \mathbf{Y}_2 \stackrel{i.i.d.}{\sim} \mathbf{F}_2$ . Fix  $1 \leq k \leq p$  and recall the definition of  $\tau$  in Section 1.2. The distance between  $F_{1,k}$  and  $F_{2,k}$  is given by  $\tau_k = \mathbb{E}[h(X_{1k}, X_{2k}) - 2h(X_{1k}, Y_{1k}) + h(Y_{1k}, Y_{2k})]$ . Here,  $\tau_k \geq 0$  and equality holds iff  $F_{1,k} = F_{2,k}$ . We denote the average of these distances by  $\bar{\tau}_p = \sum_{k=1}^p \tau_k / p$ . Clearly,  $\bar{\tau}_p = 0$  iff  $\tau_k = 0$  for all  $1 \leq k \leq p$ ,

$$\text{i.e., } \bar{\tau}_p = 0 \text{ iff } F_{1,k} = F_{2,k} \text{ for all } 1 \leq k \leq p.$$

This property of  $\bar{\tau}_p$  suggests that it can be used as a *measure of separation* between  $\mathbf{F}_1$  and  $\mathbf{F}_2$ . If  $F_{1,k} \neq F_{2,k}$  for some  $1 \leq k \leq p$ , then  $\bar{\tau}_p$  is strictly positive. This is the fundamental idea that we will use in developing a new classifier.

Recall the definition of  $h$  given in Section 1.2, and consider

$$\bar{h}_p(\mathbf{u}, \mathbf{v}) = \frac{1}{p} \sum_{k=1}^p h(u_k, v_k) \text{ for } \mathbf{u}, \mathbf{v} \in \mathbb{R}^p. \quad (2.2)$$

Using (2.2), we re-write the definition of  $\bar{\tau}_p$  as

$$\bar{\tau}_p = \mathbb{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{X}_2) - 2\bar{h}_p(\mathbf{X}_1, \mathbf{Y}_1) + \bar{h}_p(\mathbf{Y}_1, \mathbf{Y}_2)].$$

Let  $\bar{\tau}_p(1, 1), \bar{\tau}_p(1, 2) (= \bar{\tau}_p(2, 1))$  and  $\bar{\tau}_p(2, 2)$  denote the quantities  $\mathbb{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{X}_2)]$ ,  $\mathbb{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{Y}_1)]$  and  $\mathbb{E}[\bar{h}_p(\mathbf{Y}_1, \mathbf{Y}_2)]$ , respectively. Observe that

$$\bar{\tau}_p = \bar{\tau}_p(1, 1) - 2\bar{\tau}_p(1, 2) + \bar{\tau}_p(2, 2). \quad (2.3)$$

Fix  $\mathbf{z} \in \mathbb{R}^p$ . We define the following:

$$\bar{T}_{jj} = \frac{1}{n_j(n_j - 1)} \sum_{\substack{\mathbf{U}, \mathbf{U}' \in \chi_j \\ \mathbf{U} \neq \mathbf{U}'}} \bar{h}_p(\mathbf{U}, \mathbf{U}'),$$

$$\bar{T}_j(\mathbf{z}) = \frac{1}{n_j} \sum_{\mathbf{U} \in \chi_j} \bar{h}_p(\mathbf{U}, \mathbf{z}), \quad \bar{L}_j(\mathbf{z}) = \bar{T}_{jj} - 2\bar{T}_j(\mathbf{z})$$

$$\text{for } j = 1, 2 \text{ and } \bar{L}(\mathbf{z}) = \bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z}). \quad (2.4)$$

It follows from the above definitions that

$$\begin{aligned} \mathbb{E}[\bar{T}_j(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_{j'}] &= \bar{\tau}_p(j, j') \text{ and} \\ \mathbb{E}[\bar{T}_{jj}] &= \bar{\tau}_p(j, j) \text{ for } j, j' \in \{1, 2\}. \end{aligned} \quad (2.5)$$

Consequently, we obtain

$$\begin{aligned} \mathbb{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] &= \bar{\tau}_p \geq 0 \text{ and} \\ \mathbb{E}[\bar{L}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_2] &= -\bar{\tau}_p \leq 0. \end{aligned} \quad (2.6)$$

It is clear from the above equation that  $\mathbb{E}[\bar{L}(\mathbf{Z})]$  is an indicator to whether a test observation  $\mathbf{Z}$  belongs to the first or the second class. This observation motivates us to use  $\bar{L}(\mathbf{Z})$  as a discriminant.

### 2.1.1 A Classifier Based on $\bar{\tau}_p$

Based on (2.6), we propose the following classifier:

$$\delta_1(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{L}(\mathbf{z}) > 0, \\ 2, & \text{otherwise,} \end{cases} \quad (2.7)$$

The classifier  $\delta_1(\mathbf{z})$  can also be expressed as  $\arg \min_{j \in \{1, 2\}} \bar{L}_j(\mathbf{z})$ . For given random samples  $\chi_1, \dots, \chi_J$  with  $J \geq 2$ , we propose  $\delta_1(\mathbf{z}) = \arg \min_{1 \leq j \leq J} \bar{L}_j(\mathbf{z})$ , where  $\bar{L}_j(\mathbf{z}), \bar{T}_j(\mathbf{z})$  and  $\bar{T}_{jj}$  are as defined in (2.4) for  $1 \leq j \leq J$ . We denote the misclassification probability of  $\delta_1$  by  $\Delta_1$ .

## 2.2 Limitations of Using $\bar{\tau}_p$

To classify a test point, the classifier  $\delta_1$  leverages on the quantity  $\bar{\tau}_p$ , the average of distances between  $F_{1,k}$  and  $F_{2,k}$  for  $1 \leq k \leq p$ . However,  $\bar{\tau}_p$  has some limitations. Recall that

$$\begin{aligned} \bar{\tau}_p &= \bar{\tau}_p(1, 1) - 2\bar{\tau}_p(1, 2) + \bar{\tau}_p(2, 2) \\ &= \{\bar{\tau}_p(1, 1) - \bar{\tau}_p(1, 2)\} + \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 2)\}. \end{aligned}$$

Since  $\bar{\tau}_p \geq 0$ , we always have  $\bar{\tau}_p(1, 2) \leq \{\bar{\tau}_p(1, 1) + \bar{\tau}_p(2, 2)\}/2$ . Without loss of generality, let us assume that  $\bar{\tau}_p(1, 1) < \bar{\tau}_p(2, 2)$ . If  $\bar{\tau}_p(1, 2)$  lies between  $\bar{\tau}_p(1, 1)$  and  $\bar{\tau}_p(2, 2)$ , i.e.,  $\bar{\tau}_p(1, 1) < \bar{\tau}_p(1, 2) < \bar{\tau}_p(2, 2)$ , then  $\bar{\tau}_p(1, 1) - \bar{\tau}_p(1, 2) < 0$  and  $\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 2) > 0$ . Adding them up may cancel each other. As a result,  $\bar{\tau}_p$  may become small despite  $\mathbf{F}_1$  and  $\mathbf{F}_2$  being different. One way to rectify this problem is to square the two quantities before adding them up. This eliminates the possibility of cancellations. Define

$$\bar{\psi}_p = \{\bar{\tau}_p(1, 1) - \bar{\tau}_p(1, 2)\}^2 + \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 2)\}^2.$$

It is easy to check that  $\bar{\psi}_p = 0$  iff  $F_{1,k} = F_{2,k}$  for all  $1 \leq k \leq p$ . Hence,  $\bar{\psi}_p$  can be viewed as a measure of separation between  $\mathbf{F}_1$  and  $\mathbf{F}_2$ . This new measure can also be expressed as

$$\bar{\psi}_p = \frac{1}{2} [\bar{\tau}_p^2 + \{\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)\}^2]. \quad (2.8)$$

Observe that if  $\bar{\tau}_p(1,2)$  lies between  $\bar{\tau}_p(1,1)$  and  $\bar{\tau}_p(2,2)$ , then  $|\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)| > \bar{\tau}_p$ . As a result,

$$\bar{\psi}_p = \frac{1}{2} [\bar{\tau}_p^2 + \{\bar{\tau}_p(1,1) - \bar{\tau}_p(2,2)\}^2] > \frac{1}{2} [\bar{\tau}_p^2 + \bar{\tau}_p^2] = \bar{\tau}_p^2.$$

Clearly,  $\bar{\tau}_p \geq 1$  implies  $\bar{\psi}_p > \bar{\tau}_p$ , making  $\bar{\psi}_p$  a better choice than  $\bar{\tau}_p$  (in terms of measuring separation between two distributions). In general, if the underlying distributions  $\mathbf{F}_1$  and  $\mathbf{F}_2$  are such that  $\bar{\tau}_p(1,2) > \min\{\bar{\tau}_p(1,1), \bar{\tau}_p(2,2)\}$ , then a classifier that utilizes  $\bar{\psi}_p$  is shown to have better classification accuracy than the classifier  $\delta_1$  (see Section 3.1.2 for more details). The modification proposed in (2.8) is similar to what Biswas and Ghosh (2014) had suggested for improving the power of some energy based tests for HDLSS data.

### 2.2.1 A Classifier Based on $\bar{\psi}_p$

We now develop a classifier that leverages the amplified measure of dissimilarity  $\bar{\psi}_p$ . First, let us estimate  $\bar{\tau}_p(1,2)$  by

$$\bar{T}_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \bar{h}_p(\mathbf{X}_i, \mathbf{Y}_j). \quad (2.9)$$

Fix  $\mathbf{z} \in \mathbb{R}^p$ . We define

$$\begin{aligned} \bar{\theta}(\mathbf{z}) &= \frac{1}{2} \{\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}\} \{\bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z})\} \\ &\quad + \frac{1}{2} \{\bar{T}_{22} - \bar{T}_{11}\} \{\bar{L}_2(\mathbf{z}) + \bar{L}_1(\mathbf{z}) + 2\bar{T}_{12}\}. \end{aligned} \quad (2.10)$$

We will prove that  $|\bar{\theta}(\mathbf{Z})|$  is a consistent estimator of  $\bar{\psi}_p$  where  $\mathbf{Z}$  is a test observation. In particular,  $\bar{\theta}(\mathbf{Z})$  converges in probability to  $\bar{\psi}_p$  as  $p \rightarrow \infty$  if  $\mathbf{Z} \sim \mathbf{F}_1$ , and to  $-\bar{\psi}_p$  if  $\mathbf{Z} \sim \mathbf{F}_2$  (see Section 3 for more details). This motivates us to propose the following classifier:

$$\delta_2(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}(\mathbf{z}) > 0, \\ 2, & \text{otherwise.} \end{cases} \quad (2.11)$$

Let  $\Delta_2$  denote the misclassification probability of the classifier  $\delta_2$ . Unlike  $\delta_1$ , the classifier  $\delta_2$  cannot be readily extended to deal with  $J$ -class problems when  $J > 2$ . For multi-class problems, we implement the idea of ‘majority voting’ (Friedman et al., 2001).

**Examples 1 and 2** establish the necessity of the modified measure  $\bar{\psi}_p$  and the advantage of using  $\delta_2$  over  $\delta_1$ . In Figure 2, we see that  $\delta_2$  has substantial improvement over  $\delta_1$  in terms of misclassification probability. This improvement is due to the fact that

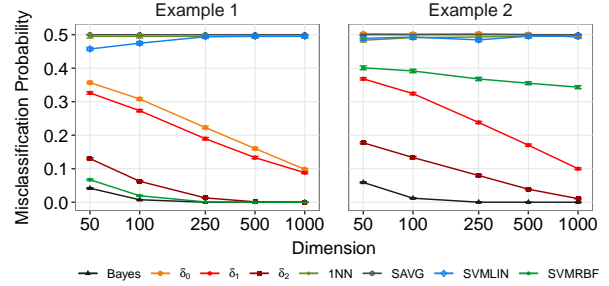


Figure 2: Average Misclassification Rates (along with Standard Errors) of the Proposed Classifiers.

$\bar{T}_{12}$  lies between  $\bar{T}_{11}$  and  $\bar{T}_{22}$  in both examples (see Table 2 in the Supplementary B). A theoretical result on the relative performance of the classifiers  $\delta_1$  and  $\delta_2$  is presented in Section 3.1.2.

## 3 ASYMPTOTIC PROPERTIES

In HDLSS settings,  $n$  is fixed and  $p \rightarrow \infty$ , whereas in the *ultra-high-dimensional* setting,  $p$  grows simultaneously with  $n$ . The behavior of the classifiers  $\delta_1$  and  $\delta_2$  is investigated in both asymptotic regimes. We first show that the classifiers yield *perfect classification* in HDLSS settings under fairly general conditions.

### 3.1 Asymptotic Behavior in HDLSS Settings

Suppose  $\mathbf{U}$  and  $\mathbf{V}$  are two independent vectors such that  $\mathbf{U} = (U_1, \dots, U_p)^\top \sim \mathbf{F}_j$  and  $\mathbf{V} = (V_1, \dots, V_p)^\top \sim \mathbf{F}_{j'}$  for  $j, j' \in \{1, 2\}$ . We assume that the component variables are weakly dependent. In particular, we assume

$$A1. \quad \sum_{1 \leq k < k' \leq p} \text{Corr}(h(U_k, V_k), h(U_{k'}, V_{k'})) = o(p^2),$$

where  $h$  is defined in Section 1.2. Assumption A1 is trivially satisfied if the component variables of the underlying distributions are independently distributed and it continues to hold when the components have weak dependence among them. For example, A1 is satisfied when the sequence  $\{h(U_k, V_k), k \geq 1\}$  has  $\rho$ -mixing property. Note that if the sequences  $\{U_k, k \geq 1\}$  and  $\{V_k, k \geq 1\}$



have  $\rho$ -mixing property, then  $\{h(U_k, V_k), k \geq 1\}$  has  $\rho$ -mixing property for every measurable function  $h$  (see Theorem 6.6-II of Bradley (2007)).

Recall assumption (iii) introduced in Section 2. Both (iii) and A1 require the component variables to be weakly dependent. However, A1 is weaker between the two since, unlike (iii), it does not require the existence of first and second order moments of the distributions. The boundedness of  $h$  enables A1 to deal with heavy-tailed distribution.

**Lemma 3.1** *If A1 is satisfied, then for a test observation  $\mathbf{Z}$ , we have*

$$(a) \text{ If } \mathbf{Z} \sim \mathbf{F}_1, \text{ then } |\bar{L}(\mathbf{Z}) - \bar{\tau}_p| \xrightarrow{P} 0 \text{ and } |\bar{\theta}(\mathbf{Z}) - \bar{\psi}_p| \xrightarrow{P} 0 \text{ as } p \rightarrow \infty.$$

$$(b) \text{ If } \mathbf{Z} \sim \mathbf{F}_2, \text{ then } |\bar{L}(\mathbf{Z}) + \bar{\tau}_p| \xrightarrow{P} 0 \text{ and } |\bar{\theta}(\mathbf{Z}) + \bar{\psi}_p| \xrightarrow{P} 0 \text{ as } p \rightarrow \infty.$$

Similar results on distance concentration can be derived for independently distributed sub-Gaussian components. See Theorem 3.1.1 of Vershynin (2018) for further details. Lemma 3.1 is stronger than existing results in the sense that it holds even when the components are not necessarily independent, or sub-Gaussian.

Lemma 3.1 states that both the discriminants converge in probability to a non-negative value if  $\mathbf{Z} \sim \mathbf{F}_1$ , while they converge in probability to a value which is not positive, when  $\mathbf{Z} \sim \mathbf{F}_2$ . We expect  $\delta_1$  and  $\delta_2$  to have a good performance if  $\bar{\tau}_p$  and  $\bar{\psi}_p$  do not vanish with increasing dimension. Clearly,  $\bar{\tau}_p = \bar{\psi}_p = 0$  iff  $F_{1,k} = F_{2,k}$  for all  $1 \leq k \leq p$ . Hence, it is reasonable to assume the following:

$$\text{A2. } \liminf_p \bar{\tau}_p > 0.$$

A2 implies that the separation between  $\mathbf{F}_1$  and  $\mathbf{F}_2$  is asymptotically non-negligible. Observe that the assumption is satisfied if the component variables of  $\mathbf{U} \sim \mathbf{F}_j$  are identically distributed for  $j = 1, 2$ . Then,  $\tau_k = \tau_1 > 0$  for all  $k \geq 1$ , making  $\bar{\tau}_p (= \tau_1)$  free of  $p$ . It follows from the definition of  $\bar{\psi}_p$  in (2.8) that A2 implies  $\liminf_p \bar{\psi}_p > 0$ .

### 3.1.1 Asymptotic Properties of $\delta_1$ and $\delta_2$ in HDLSS Settings

Now, we discuss behavior of the classifiers  $\delta_1$  and  $\delta_2$  in HDLSS settings. We show that under fairly

general conditions, the proposed classifiers  $\delta_1$  and  $\delta_2$  perfectly classify an observation as the dimension increases.

**Theorem 3.2** *If A1 and A2 are satisfied, then for any  $\pi_1 > 0$ ,*

- (a)  $\Delta_1 \rightarrow 0$  as  $p \rightarrow \infty$ , and
- (b)  $\Delta_2 \rightarrow 0$  as  $p \rightarrow \infty$ .

Observe that the asymptotic behavior of the classifiers are no longer governed by the constants  $\nu_{jj'}$ ,  $\sigma_j^2$  for  $j, j' \in \{1, 2\}$ . In fact, their behavior do not depend on the existence of moments. In this sense, the classifiers  $\delta_1$  and  $\delta_2$  are robust.

Asymptotic behavior of the proposed classifiers is free of moment conditions.

The classifiers yield *perfect classification* under weak conditions.

One should observe that assumptions A1 and A2 are fairly general, and Theorem 3.2 is stronger than what exists in the current literature.

### 3.1.2 Comparison Between $\delta_1$ and $\delta_2$

We have seen that both the proposed classifiers yield *perfect classification* under the same set of assumptions. The next result provides a set of sufficient conditions under which one classifier performs better than the other.

First, let us consider the following assumption:

$$\text{A3. There exists a } p_0 \in \mathbb{N} \text{ such that } \bar{\tau}_p(1, 2) > \min\{\bar{\tau}_p(1, 1), \bar{\tau}_p(2, 2)\} \text{ for all } p \geq p_0.$$

If assumption A3 is satisfied, then one of  $\bar{\tau}_p(1, 1) - \bar{\tau}_p(1, 2)$  and  $\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 2)$  is positive, while the other is negative.  $\bar{\tau}_p$  may take a small value (see the discussion in Section 2.2). The next result suggests that under such circumstances,  $\delta_2$  leads to better performance than  $\delta_1$ .

**Theorem 3.3** *If assumptions (A1) – (A3) are satisfied, then there exists an integer  $p'_0$  such that*

$$\Delta_2 \leq \Delta_1 \text{ for all } p \geq p'_0.$$

Recall that  $\bar{T}_{11}, \bar{T}_{12}$  and  $\bar{T}_{22}$  are unbiased estimators of  $\bar{\tau}_p(1, 1), \bar{\tau}_p(1, 2)$  and  $\bar{\tau}_p(2, 2)$ , respectively (see (2.5)). We use these estimators to understand the

relative performance of the classifiers. In **Examples 1** and **2**,  $\bar{T}_{12}$  is observed to be lying between  $\bar{T}_{11}$  and  $\bar{T}_{22}$  (see Table 2 in Supplementary B). Following Theorem 3.3, we expect  $\Delta_2$  to be smaller than  $\Delta_1$  in these examples. Figure 2 shows that the estimated misclassification probability of the classifier  $\delta_2$  is indeed lower than that of  $\delta_1$  in both the examples.

### 3.2 Asymptotic Properties of $\delta_1$ and $\delta_2$ when Sample Size Increases

In this section, we assess the performance of our classifiers in the *ultra-high-dimensional* asymptotic regime, when the dimension  $p$  ( $\equiv p_n$ ) is allowed to grow with  $n$  (in non-polynomial order). In particular, we assume the following:

A4. There exists a  $0 \leq \beta < 1$  such that

$$\log p_n = O(n^\beta).$$

Recall that in the classical asymptotic regime,  $p$  is fixed and  $n \rightarrow \infty$ . Therefore, the classical setting is a special case of the *ultra-high-dimensional* regime with  $\beta = 0$ . We also assume that  $\lim_{n \rightarrow \infty} n_1/n = \pi_1 > 0$ .

We first present the ‘oracle’ versions of our classifiers  $\delta_1$  and  $\delta_2$ . Fix  $\mathbf{z} \in \mathbb{R}^p$ . If  $\mathbf{F}_1$  and  $\mathbf{F}_2$  are known, then the ‘oracle’ version of  $\delta_1$  is defined as follows:

$$\delta_1^0(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{L}^0(\mathbf{z}) > 0, \\ 2, & \text{otherwise,} \end{cases} \quad (3.1)$$

where  $\bar{L}^0(\mathbf{z}) = \bar{L}_2^0(\mathbf{z}) - \bar{L}_1^0(\mathbf{z})$ , with  $\bar{L}_j^0(\mathbf{z}) = \bar{\tau}_p(j, j) - 2E[\bar{h}_p(\mathbf{U}, \mathbf{z})]$  for  $\mathbf{U} \sim \mathbf{F}_j$  and  $j = 1, 2$ . Similarly, we define  $\delta_2^0$ , the ‘oracle’ version of  $\delta_2$  as follows:

$$\delta_2^0(\mathbf{z}) = \begin{cases} 1, & \text{if } \bar{\theta}^0(\mathbf{z}) > 0, \\ 2, & \text{otherwise,} \end{cases} \quad (3.2)$$

where  $2\bar{\theta}^0(\mathbf{z}) = \bar{\tau}_p \bar{L}^0(\mathbf{z}) + \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\} \times \{\bar{L}_2^0(\mathbf{z}) + \bar{L}_1^0(\mathbf{z}) + 2\bar{\tau}_p(1, 2)\}$ . Note that  $\bar{L}(\mathbf{z})$  and  $\bar{\theta}(\mathbf{z})$  (defined in (2.4) and (2.10)) are in fact estimators of  $\bar{L}^0(\mathbf{z})$  and  $\bar{\theta}^0(\mathbf{z})$ , respectively.

Let  $\Delta_j^0$  denote the misclassification probability of the classifier  $\delta_j^0$   $j = 1, 2$ . In this section, we derive an upper bound on the difference  $\Delta_j - \Delta_j^0$  for  $j = 1, 2$ . Furthermore, we show that in the classical setting (i.e.,  $p_n (= p)$  is fixed), if the competing distributions are absolutely continuous, then  $\Delta_j - \Delta_j^0$  converges to 0 for  $j = 1, 2$  as  $n \rightarrow \infty$ . We first look into the convergence of the discriminants  $\bar{L}(\mathbf{z})$  and  $\bar{\theta}(\mathbf{z})$ .

**Lemma 3.4** *Suppose assumption A4 is satisfied for some  $0 \leq \beta < 1$ . For any  $\pi_1 > 0$  and  $0 < \gamma < (1 - \beta)/2$ , there exist positive constants  $B_0$  and  $B_1$  such that*

$$(a) \ P[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] \leq O\left(e^{-B_0\{n^{1-2\gamma} - n^\beta\}}\right),$$

$$(b) \ P[|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})| > n^{-\gamma}] \leq O\left(e^{-B_1\{n^{1-2\gamma} - n^\beta\}}\right)$$

for all  $\mathbf{z} \in \mathbb{R}^p$ .

Since  $1 - 2\gamma > \beta$ , we have  $e^{-\{n^{1-2\gamma} - n^\beta\}} \rightarrow 0$  as  $n \rightarrow \infty$ . The above result shows that  $|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})|$  and  $|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})|$  converge to 0, at an exponential rate as  $n$  increases. As a consequence of Lemma 3.4, we have the next result.

**Theorem 3.5** *Suppose assumption A4 is satisfied for some  $0 \leq \beta < 1$ . For any  $\pi_1 > 0$  and  $0 < \gamma < (1 - \beta)/2$ , there exist positive constants  $B_0$  and  $B_1$  such that*

$$(a) \ \Delta_1 - \Delta_1^0 \leq O\left(e^{-B_0\{n^{1-2\gamma} - n^\beta\}}\right) + P[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}],$$

$$(b) \ \Delta_2 - \Delta_2^0 \leq O\left(e^{-B_1\{n^{1-2\gamma} - n^\beta\}}\right) + P[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}].$$

Clearly,  $e^{-B_0\{n^{1-2\gamma} - n^\beta\}}$  and  $e^{-B_1\{n^{1-2\gamma} - n^\beta\}}$  go to 0 as  $n \rightarrow \infty$  for all  $0 < \gamma < (1 - \beta)/2$ . Additionally, if  $P[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$  and  $P[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$  decrease to 0 then Theorem 3.5 suggests that  $\Delta_j - \Delta_j^0 \rightarrow 0$  as  $n \rightarrow \infty$  for  $j = 1, 2$ . Observe that in the classical setting, where  $p$  is fixed (i.e.,  $\beta = 0$ ), if  $\mathbf{F}_1, \mathbf{F}_2$  are absolutely continuous, then  $P[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$  and  $P[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$  go to 0 as  $n \rightarrow \infty$ . Suppose, A4 is satisfied for  $\beta > 0$ , i.e.,  $p$  grows with  $n$ . One can prove that if assumptions A1 and A2 are satisfied, then  $P[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}]$ ,  $P[|\bar{\theta}^0(\mathbf{Z})| < n^{-\gamma}]$  go to 0 as  $\min\{n, p_n\} \rightarrow \infty$ . Moreover,  $\Delta_1^0$  and  $\Delta_2^0$  decay to 0 under the same conditions. As a result,  $\Delta_j \rightarrow 0$  as  $\min\{n, p_n\} \rightarrow \infty$  for  $j = 1, 2$ . We do not provide with mathematical details of the arguments as it is similar to that of Theorem 3.2.

### 3.3 Computational Complexity

Computing  $\bar{T}_{jj'}$  and  $\bar{T}_j(\mathbf{z})$  for  $\mathbf{z} \in \mathbb{R}^p$  requires  $O(n^2p)$  and  $O(np)$  operations, respectively, for  $j, j' \in \{1, 2\}$ . Thus, the overall complexity of classifying an observation using  $\delta_1$  and  $\delta_2$  is  $O(n^2p)$ . Clearly, the complexity increases linearly with respect to  $p$ . This makes the methods advantageous in analyzing high-dimensional data sets. We report average time taken by these classifiers to classify a

test observation in Table 2 of the Supplementary B. It clearly shows the advantage of using  $\delta_1$  and  $\delta_2$  over some popular classifiers.

## 4 SIMULATION STUDY

In this section, we analyze some simulated data sets to compare the classifiers  $\delta_0, \delta_1$  and  $\delta_2$  with some popular classifiers like GLMNET (Hastie et al., 2009), the usual 1NN, NN based on the random projection method (NN-RAND) (Deegalla and Bostrom, 2006), neural networks (NNET)(Bishop, 1995), SVM-LIN and SVM-RBF. All numerical exercises are performed on an Intel Xeon Gold 6140 CPU (2.30GHz, 2295 Mhz) using the statistical software R. Details about the used packages and parameters related to implementation of the popular classifiers are provided in Supplementary B.

Recall **Examples 1** and **2** introduced in Section 1. Three more examples are considered to conduct a comparative study of the performance of these classifiers.

**Example 3**  $X_{1k} \stackrel{i.i.d.}{\sim} C(0, 1)$  and  $Y_{1k} \stackrel{i.i.d.}{\sim} C(1, 1)$ ,

**Example 4**  $X_{1k} \stackrel{i.i.d.}{\sim} C(0, 1)$  and  $Y_{1k} \stackrel{i.i.d.}{\sim} C(0, 2)$ ,

**Example 5**  $X_{1k} \stackrel{i.i.d.}{\sim} \text{Par}(1, 1)$  and  $Y_{1k} \stackrel{i.i.d.}{\sim} \text{Par}(1.25, 1)$ ,

for  $1 \leq k \leq p$ . Here,  $C(\mu, \sigma)$  denotes the Cauchy distribution with location  $\mu \in \mathbb{R}$  and scale  $\sigma > 0$ , while  $\text{Par}(\theta, s)$  denotes the Pareto distribution with  $\theta > 0$  and scale  $s > 0$ .

**Examples 3, 4** and **5** correspond to location, scale and location-scale problem, respectively. All three examples involve heavy-tailed distributions. In each example, we simulated data for  $p = 50, 100, 250, 500$  and  $1000$ . The training sample was formed with

20 observations from each class and a test set of size 200 (100 from each class) was used. This process was repeated 100 times to estimate the misclassification probabilities, which are reported in Figure 3 along with the standard errors.

The discussion on the performance of  $\delta_0$  in **Examples 1** and **2** is already presented in Section 2. Figure 3 shows that  $\delta_0$  fails miserably in **Examples 3-5**. Recall Theorem 2.2 and observe that assumption (iii) is violated for these examples since the competing distributions are heavy-tailed. Hence, the poor performance of  $\delta_0$ .

The classifiers  $\delta_1$  and  $\delta_2$  lead to promising results in all examples. Assumption A1 is satisfied in these examples since the component variables are independently distributed. Also, the marginals are identical, i.e.,  $F_{1,k} = F_{1,1}$  and  $F_{2,k} = F_{2,1}$  for all  $1 \leq k \leq p$ . Therefore,  $\bar{\tau}_p (= \tau_1 > 0)$  is free of both  $n$  and  $p$ . Hence, A2 is satisfied. Consequently, Theorem 3.2 holds for all examples.

Figure 3 shows that the misclassification probability of  $\delta_2$  is smaller than that of  $\delta_1$  in **Examples 1, 2, 4** and **5**. Whereas,  $\delta_1$  outperformed  $\delta_2$  in **Example 3**. Recall the discussion on relative performance of the proposed classifiers in Section 3.1.2. Following Theorem 3.3, we estimated  $\bar{\tau}_p(1, 1)$ ,  $\bar{\tau}_p(1, 2)$ , and  $\bar{\tau}_p(2, 2)$  by  $\bar{T}_{11}$ ,  $\bar{T}_{12}$ , and  $\bar{T}_{22}$ , respectively, for all examples and found that  $\bar{T}_{12} < \min\{\bar{T}_{11}, \bar{T}_{22}\}$  in **Example 3** while  $\bar{T}_{12} > \min\{\bar{T}_{11}, \bar{T}_{22}\}$  in other examples (see Table 2 of Supplementary B). The numerical evidence clearly support our claim in Theorem 3.3.

In general, all the popular classifiers exhibit poor performance in general, except for a few instances. In **Example 1**, only SVM-RBF identified the difference between scales of the competing populations and yielded *perfect classification*. The rest of the methods failed miserably and misclassified nearly

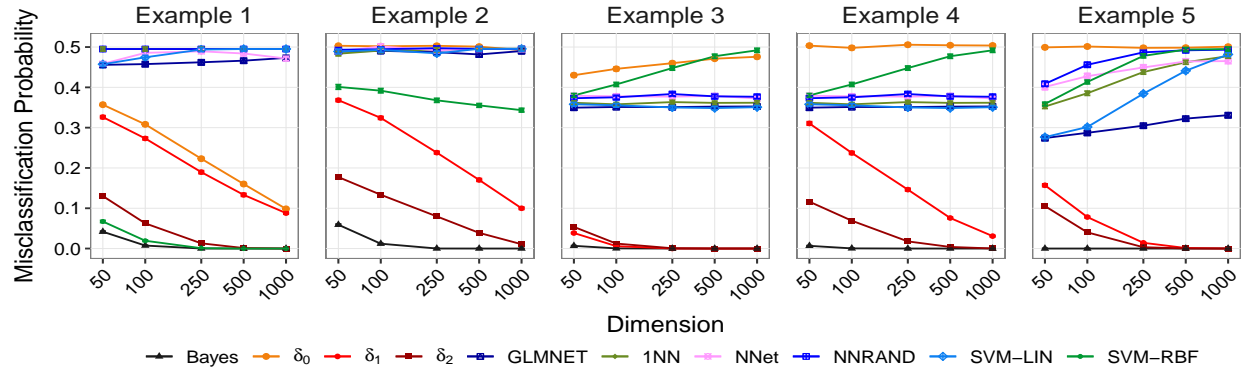


Figure 3: Average Misclassification Rates of Classifiers for Fixed  $n = 40(20 + 20)$  and Varying  $p$ .



50% of the test observations. In **Example 2**, none of the classifiers had satisfactory results since they are unable to look beyond differences between first two moments in HDLSS settings. In **Examples 3-5**, the competing distributions are heavy-tailed and we observe deteriorating performances of all these classifiers.

## 5 REAL DATA ANALYSIS

We study the performance of the proposed classifiers in two real data sets, namely, **Computers** and **SmoothSubspace** available at the UCR Time Series Archive (see [Dau et al., 2018](#)). These data sets have *fixed training and test sets*. For our analysis, we combined the training and test data. We randomly selected 50% of the observations from the combined set to form a new set of training observations, while keeping the proportions of observations from different classes consistent. The remaining observations were considered as the test set. This procedure was repeated 100 times to obtain stable estimates of the misclassification probabilities.

The **Computers** (Comp) data contains readings on electricity consumption from 251 households in UK, sampled in two-minute intervals over a month. Each observation is of length 720 making the data high-dimensional. Classes are ‘Desktop’ and ‘Laptop’ with 250 (125 training and 125 test) samples in each. From Table 1, we observe that  $\delta_0$  performed quite poorly, misclassifying almost half of the test observations. The misclassification probability of  $\delta_2$  is smaller than that of  $\delta_1$  for this data. To understand the relative performance of the classifiers  $\delta_1$  and  $\delta_2$ , we obtained  $\bar{T}_{11} = 0.972$ ,  $\bar{T}_{12} = 1.043$ , and  $\bar{T}_{22} = 1.155$ . Observe that  $\bar{T}_{12}$  lies between  $\bar{T}_{11}$  and  $\bar{T}_{22}$ . This explains the superior performance of  $\delta_2$  over  $\delta_1$ . In fact,  $\delta_2$  outperforms the rest of the classifiers. The regularized linear classifier GLMNET secured the third position with a competitive performance, closely followed by SVM-RBF whereas INN, NN RAND, NNET and SVM-LIN all four classifiers misclassify more than 40% of the observations.

The second data set **SmoothSubspace** (SSub) is about testing the ability of a clustering algorithm to extract smooth subspaces for clustering time series data. This data set has 3 classes with 100 (50 train and 50 test) observations each. The observations have dimension 15. We observe in Table 1 that the classifier  $\delta_0$  misclassified 18% of the test observation. It is also the worst among all.  $\delta_1$  yielded the lowest misclassification rate, while  $\delta_2$  had the second best performance in this

data set. We obtained  $\bar{T}_{11} = 1.384$ ,  $\bar{T}_{22} = 1.378$ ,  $\bar{T}_{33} = 1.386$ ,  $\bar{T}_{12} = 1.340$ ,  $\bar{T}_{13} = 1.326$ , and  $\bar{T}_{23} = 1.314$ . Observe that  $\bar{T}_{jj} < \min\{\bar{T}_{jj'}, \bar{T}_{j'j'}\}$  making both  $\bar{T}_{jj} - \bar{T}_{jj'}$  and  $\bar{T}_{j'j'} - \bar{T}_{jj}$  positive for all  $1 \leq j \neq j' \leq 3$ . Thus, squaring the differences to amplify the measure of separation between  $\mathbf{F}_j$  and  $\mathbf{F}_{j'}$  becomes unnecessary for all  $j \neq j'$ . Among the existing methods, NNET produced the highest misclassification probability. The linear classifiers GLMNET and SVM-LIN also performed poorly while non-linear classifiers like NN, NN RAND and SVM-RBF yielded better misclassification rates. In particular, SVM-RBF yielded the lowest misclassification rate among the popular classifiers, followed by NN-RAND. Although, their performance is six times worse than that of  $\delta_1$ .

Table 1: Average Misclassification Rates of Classifiers (in %) with Standard Errors in Parentheses

| Data    | $\delta_0$ | $\delta_1$  | $\delta_2$   | GLM<br>NET | INN    | NN<br>RAND | NNet   | SVM<br>LIN | SVM<br>RBF |
|---------|------------|-------------|--------------|------------|--------|------------|--------|------------|------------|
| Comp    | 47.09      | 36.40       | <b>35.47</b> | 39.10      | 42.67  | 42.04      | 46.80  | 46.16      | 39.95      |
| $J = 2$ | (0.24)     | (0.22)      | (0.21)       | (0.24)     | (0.28) | (0.27)     | (0.28) | (0.34)     | (0.27)     |
| SSub    | 18.15      | <b>1.05</b> | 1.33         | 13.35      | 8.71   | 7.09       | 16.19  | 10.79      | 6.35       |
| $J = 3$ | (0.27)     | (0.06)      | (0.08)       | (0.28)     | (0.20) | (0.22)     | (0.44) | (0.28)     | (0.19)     |

## 6 CONCLUDING REMARKS

In this article, we present classifiers that are capable of perfectly classifying an observation if the one-dimensional marginals of the underlying distributions are different. We study the theoretical properties of the classifiers in the HDLSS and *ultra-high-dimensional* setting. The proposed classifiers are robust in nature. They yield *perfect classification* even when the competing distributions are heavy-tailed. Their tuning free discriminants make the methods fast and easy to implement. Analysis of several simulated and real data sets show the promising performance of the classifiers.

Suppose the underlying distributions are such that their one dimensional marginals are identical and the discriminatory information comes from the joint distributions of groups of variables. Under such circumstances, the proposed measures of separation become 0 and the classifiers may fail to yield *perfect classification*. Developing fast and robust classifiers that harness the distance between joint distributions of components could be a future direction of research.

## Acknowledgments

We thank the reviewers for their careful reading of an earlier version of the article and for providing us with helpful comments. We would also like to thank Purushottam Kar and Soham Sarkar for their valuable inputs which improved the article.

## Bibliography

- Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. Springer.
- Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H., and Marron, J. (2018). A survey of high dimension low sample size asymptotics. *Australian & New Zealand Journal of Statistics*, 60(1):4–19.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171.
- Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144.
- Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*. Kendrick Press.
- Chan, Y.-B. and Hall, P. (2009a). Robust nearest-neighbor methods for classifying high-dimensional data. *The Annals of Statistics*, 37(6A):3186–3203.
- Chan, Y.-B. and Hall, P. (2009b). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika*, 96(2):469–478.
- Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.
- Dau, H. A., Keogh, E., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., Yanping, Hu, B., Begum, N., Bagnall, A., Mueen, A., and Batista, G. (2018). The UCR time series classification archive. [https://www.cs.ucr.edu/~eamonn/time\\_series\\_data\\_2018/](https://www.cs.ucr.edu/~eamonn/time_series_data_2018/).
- Deegalla, S. and Bostrom, H. (2006). Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*, pages 245–250. IEEE.
- Dutta, S. and Ghosh, A. K. (2016). On some transformations of high dimension, low sample size data for nearest neighbor classification. *Machine Learning*, 102(1):57–83.
- Francois, D., Wertz, V., and Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics: New York.
- Globerson, A. and Roweis, S. (2005). Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, 18:451–458.
- Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer, New York.
- Li, Z. and Zhang, Y. (2020). On a projective ensemble approach to two sample test for equality of distributions. In *International Conference on Machine Learning*, pages 6020–6027. PMLR.
- Thrampoulidis, C. (2020). Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Neural Information Processing Systems (NeurIPS 2020)*.
- Tomašev, N., Radovanović, M., Mladenović, D., and Ivanović, M. (2014). Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. *International Journal of Machine Learning and Cybernetics*, 5(3):445–458.
- Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley & Sons.
- Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.
- Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press.
- Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2).

---

# Supplementary Material: On Some Fast And Robust Classifiers For High Dimension, Low Sample Size Data

---

## A MATHEMATICAL DETAILS AND PROOFS

We will use the following definitions in our proofs presented below.

1.  $a_n = o(b_n)$  as  $n \rightarrow \infty$  implies that for every  $\epsilon > 0$  there exists an  $N \in \mathbb{N}$  such that  $|a_n/b_n| < \epsilon$  for all  $n \geq N$ .
2.  $a_n = O(b_n)$  as  $n \rightarrow \infty$  implies that there exist  $M > 0$  and  $N \in \mathbb{N}$  such that  $|a_n/b_n| < M$  for all  $n \geq N$ .

**Lemma A.1** Suppose  $\mathbf{U} \sim \mathbf{F}_j$  and  $\mathbf{V} \sim \mathbf{F}_{j'}$  for  $j, j' \in \{1, 2\}$  and  $\mathbf{U}, \mathbf{V}$  are independent. If assumptions (i)-(iii) are satisfied, then

$$\left| h(\mathbf{U}, \mathbf{V}) - \frac{1}{2\pi} \sin^{-1} \left( \frac{\nu_{jj'}}{[(\sigma_j^2 + \nu_{jj})(\sigma_{j'}^2 + \nu_{j'j'})]^{\frac{1}{2}}} \right) \right| \xrightarrow{P} 0 \text{ as } p \rightarrow \infty.$$

**Proof of Lemma A.1** We have assumed in (ii) that the limiting constants  $\nu_{jj'}$ , and  $\sigma_j^2$  exist for  $j, j' \in \{1, 2\}$ . Fix  $\epsilon > 0$ . Now, observe that

$$\begin{aligned} \mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \nu_{jj'} \right| > \epsilon \right] &= \mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} + \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} - \nu_{jj'} \right| > \epsilon \right] \\ &\leq \mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} \right| > \frac{\epsilon}{2} \right] + \mathbb{I} \left[ \left| \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} - \nu_{jj'} \right| > \frac{\epsilon}{2} \right] \text{ [using the union bound].} \end{aligned}$$

Since  $\lim_{p \rightarrow \infty} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} = \nu_{jj'}$ , there exists  $p_0 \in \mathbb{N}$  such that  $\mathbb{I} \left[ \left| \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} - \nu_{jj'} \right| > \frac{\epsilon}{2} \right] = 0$  for all  $p \geq p_0$ . So, we get

$$\mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \nu_{jj'} \right| > \epsilon \right] \leq \mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} \right| > \frac{\epsilon}{2} \right] \text{ for all } p \geq p_0.$$

Observe that

$$\begin{aligned} &\mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} \right| > \frac{\epsilon}{2} \right] \tag{A.1} \\ &= \mathbb{P} \left[ \left| \frac{1}{p} \sum_{k=1}^p U_k V_k - \frac{1}{p} \sum_{k=1}^p \mathbb{E}[U_k] \mathbb{E}[V_k] \right| > \frac{\epsilon}{2} \right] \\ &\leq \frac{4}{\epsilon^2} \text{Var} \left[ \frac{1}{p} \sum_{k=1}^p U_k V_k \right] \text{ [using Chebyshev's inequality]} \\ &= \frac{4}{\epsilon^2 p^2} \sum_{k=1}^p \text{Var}[U_k V_k] + \frac{8}{\epsilon^2 p^2} \sum_{1 \leq k < k' \leq p} \text{Cov}(U_k V_k, U_{k'} V_{k'}) \\ &\leq \frac{4}{\epsilon^2 p^2} \sum_{k=1}^p \mathbb{E}[U_k^2 V_k^2] + \frac{8}{\epsilon^2 p^2} \sum_{1 \leq k < k' \leq p} \text{Corr}(U_k V_k, U_{k'} V_{k'}) \sqrt{\mathbb{E}[U_k^2 V_k^2] \mathbb{E}[U_{k'}^2 V_{k'}^2]} \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{4C}{\epsilon^2 p} + \frac{8C}{\epsilon^2 p^2} \sum_{1 \leq k < k' \leq p} \text{Corr}(U_k V_k, U_{k'} V_{k'}) \quad [\text{for some } C < \infty \text{ (due to (i))}] \\
 &= o(1) \text{ as } p \rightarrow \infty \text{ [using (iii)]}.
 \end{aligned} \tag{A.2}$$

Therefore,  $\mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \nu_{jj'} \right| > \epsilon \right] \leq \mathbb{P} \left[ \left| \frac{1}{p} \mathbf{U}^\top \mathbf{V} - \frac{1}{p} \boldsymbol{\mu}_j^\top \boldsymbol{\mu}_{j'} \right| > \frac{\epsilon}{2} \right] = o(1)$  for  $\mathbf{U} \sim \mathbf{F}_j$  and  $\mathbf{V} \sim \mathbf{F}_{j'}$  with  $j, j' \in \{1, 2\}$  as  $p \rightarrow \infty$ .

Following similar arguments, one can also prove that (as  $p \rightarrow \infty$ ),

$$\begin{aligned}
 &\mathbb{P} \left[ \left| \frac{1}{p} \|\mathbf{U}\|^2 - \frac{1}{p} \mathbb{E}[\|\mathbf{U}\|^2] \right| > \epsilon \right] \leq o(1) \\
 \Rightarrow &\mathbb{P} \left[ \left| \frac{1}{p} \|\mathbf{U}\|^2 - \frac{1}{p} \{ \|\boldsymbol{\mu}_j\|^2 + \text{tr}(\Sigma_j) \} \right| > \epsilon \right] \leq o(1) \\
 \Rightarrow &\mathbb{P} \left[ \left| \frac{1}{p} \|\mathbf{U}\|^2 - \{ \nu_{jj} + \sigma_j^2 \} \right| > \epsilon \right] \leq o(1) \quad \left[ \lim_{p \rightarrow \infty} \|\boldsymbol{\mu}_j\|^2/p = \nu_{jj} \text{ and } \lim_{p \rightarrow \infty} \text{tr}(\Sigma_j)/p = \sigma_j^2 \right].
 \end{aligned}$$

Using the continuous mapping theorem (repeatedly), we obtain

$$\sin(2\pi h(\mathbf{U}, \mathbf{V})) = \frac{1 + \mathbf{U}^\top \mathbf{V}}{\sqrt{(1 + \|\mathbf{U}\|^2)(1 + \|\mathbf{V}\|^2)}} = \frac{\frac{1}{p} + \frac{\mathbf{U}^\top \mathbf{V}}{p}}{\sqrt{\left(\frac{1}{p} + \frac{\|\mathbf{U}\|^2}{p}\right) \left(\frac{1}{p} + \frac{\|\mathbf{V}\|^2}{p}\right)}} \xrightarrow{\mathbb{P}} \frac{\nu_{jj'}}{\sqrt{(\sigma_j^2 + \nu_{jj})(\sigma_{j'}^2 + \nu_{j'j'})}}$$

as  $p \rightarrow \infty$ . Consequently, we have  $h(\mathbf{U}, \mathbf{V}) \xrightarrow{\mathbb{P}} \frac{1}{2\pi} \sin^{-1} \left\{ \frac{\nu_{jj'}}{\sqrt{(\sigma_j^2 + \nu_{jj})(\sigma_{j'}^2 + \nu_{j'j'})}} \right\}$  as  $p \rightarrow \infty$ .

Hence, the proof.  $\square$

Define  $\tau_{ii} = \frac{1}{2\pi} \sin^{-1} \left\{ \frac{\nu_{ii}}{(\sigma_i^2 + \nu_{ii})} \right\}$  for  $i = 1, 2$  and  $\tau_{12} = \frac{1}{2\pi} \sin^{-1} \left\{ \frac{\nu_{12}}{\sqrt{(\sigma_1^2 + \nu_{11})(\sigma_2^2 + \nu_{22})}} \right\}$ . Lemma 2.1 suggests that  $h(\mathbf{U}, \mathbf{V}) \xrightarrow{\mathbb{P}} \tau_{jj'}$  as  $p \rightarrow \infty$ , where  $\mathbf{U} \sim \mathbf{F}_j$ ,  $\mathbf{V} \sim \mathbf{F}_{j'}$  for  $j, j' \in \{1, 2\}$  and  $\mathbf{U}, \mathbf{V}$  are independent.

**Corollary A.2** For  $j, j' \in \{1, 2\}$ , if assumptions (i)-(iii) are satisfied, then

- (a)  $|T_{jj'} - \tau_{jj'}| \xrightarrow{\mathbb{P}} 0$  as  $p \rightarrow \infty$ , and
- (b) if  $\mathbf{Z} \sim \mathbf{F}_{j'}$ , then  $|T_j(\mathbf{Z}) - \tau_{jj'}| \xrightarrow{\mathbb{P}} 0$  as  $p \rightarrow \infty$ .

### Proof of Corollary A.2

- (a) Fix  $\epsilon > 0$ . It follows from Lemma 2.1 that

$$\begin{aligned}
 \mathbb{P}[|T_{11} - \tau_{11}| > \epsilon] &= \mathbb{P} \left[ \left| \frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} \{h(\mathbf{X}_i, \mathbf{X}_j) - \tau_{11}\} \right| > \epsilon \right] \\
 &\leq \mathbb{P} \left[ \frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} |h(\mathbf{X}_i, \mathbf{X}_j) - \tau_{11}| > \epsilon \right] \\
 &\leq \sum_{1 \leq i \neq j \leq n_1} \mathbb{P}[|h(\mathbf{X}_i, \mathbf{X}_j) - \tau_{11}| > \epsilon] \\
 &= n_1(n_1 - 1)o(1) = o(1) \text{ as } p \rightarrow \infty \text{ [} n_1 \text{ is fixed]}.
 \end{aligned} \tag{A.3}$$

Therefore,  $|T_{11} - \tau_{11}| \xrightarrow{\mathbb{P}} 0$  as  $p \rightarrow \infty$ . Similarly,  $|T_{12} - \tau_{12}|$  and  $|T_{22} - \tau_{22}|$  also converge in probability to 0 as  $p \rightarrow \infty$ .

- (b) Fix  $\epsilon > 0$ . Let  $\mathbf{U} \in \chi_i$  (i.e.,  $\mathbf{U} \sim \mathbf{F}_j$ ) and  $\mathbf{Z} \sim \mathbf{F}_{j'}$  for  $j, j' \in \{1, 2\}$ . Since  $n_j$  is fixed for  $j \in \{1, 2\}$ , using Lemma 2.1, we have

$$\begin{aligned}
 P[|T_j(\mathbf{Z}) - \tau_{jj'}| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_{j'}] &= P\left[\left|\left\{\frac{1}{n_j} \sum_{\mathbf{U} \in \chi_j} \{h(\mathbf{U}, \mathbf{Z}) - E[h(\mathbf{U}, \mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_{j'}]\}\right\}\right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_{j'}\right] \\
 &\leq P\left[\frac{1}{n_j} \sum_{\mathbf{U} \in \chi_j} |h(\mathbf{U}, \mathbf{Z}) - E[h(\mathbf{U}, \mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_{j'}]| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_{j'}\right] \\
 &\leq \sum_{\mathbf{U} \in \chi_j} P[|h(\mathbf{U}, \mathbf{Z}) - E[h(\mathbf{U}, \mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_{j'}]| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_{j'}] \\
 &\leq n_j o(1) = o(1) \text{ as } p \rightarrow \infty [n_j \text{ is fixed}].
 \end{aligned} \tag{A.4}$$

Hence, the proof.  $\square$

Recall the definition of  $\tau_0$  given as follows:

$$\begin{aligned}
 \tau_0 &= \frac{1}{2\pi} \sin^{-1} \left\{ \frac{\nu_{11}}{(\sigma_1^2 + \nu_{11})} \right\} + \frac{1}{2\pi} \sin^{-1} \left\{ \frac{\nu_{22}}{(\sigma_2^2 + \nu_{22})} \right\} - \frac{1}{\pi} \sin^{-1} \left\{ \frac{\nu_{12}}{\sqrt{(\sigma_1^2 + \nu_{11})(\sigma_2^2 + \nu_{22})}} \right\} \\
 \text{i.e., } \tau_0 &= \tau_{11} + \tau_{22} - 2\tau_{12}.
 \end{aligned}$$

If  $\nu_{11} = \nu_{12} = \nu_{22} = 0$ , then  $\tau_0 = 0$ . Also, if  $\nu_{11} = \nu_{12} = \nu_{22}$  and  $\sigma_1^2 = \sigma_2^2$ , then  $\tau_0 = 0$ .

#### Proof of Lemma 2.1



- (a) First of all, we have  $|L(\mathbf{Z}) - \tau| \leq |L(\mathbf{Z}) - \tau_0| + |\tau - \tau_0|$  using triangle inequality for all  $\mathbf{z} \in \mathbb{R}^p$ .

Now, observe that  $L(\mathbf{Z}) = L_2(\mathbf{Z}) - L_1(\mathbf{Z}) = \{T_{22} - 2T_2(\mathbf{Z})\} - \{T_{11} - 2T_1(\mathbf{Z})\}$ . If  $\mathbf{Z} \sim \mathbf{F}_1$ , then it follows from Corollary A.2 that

$$L(\mathbf{Z}) \xrightarrow{P} \{\tau_{22} - 2\tau_{12}\} - \{\tau_{11} - 2\tau_{11}\} = \tau_{11} + \tau_{22} - 2\tau_{12} = \tau_0 \text{ as } p \rightarrow \infty.$$


It follows from Lemma A.1 that  $h(\mathbf{X}_1, \mathbf{X}_2) \xrightarrow{P} \tau_{11}$ ,  $h(\mathbf{X}_1, \mathbf{Y}_1) \xrightarrow{P} \tau_{12}$  and  $h(\mathbf{Y}_1, \mathbf{Y}_2) \xrightarrow{P} \tau_{22}$  as  $p \rightarrow \infty$ . Since,  $h$  is a bounded function, using the Dominated Convergence Theorem, we have  $E[h(\mathbf{X}_1, \mathbf{X}_2)] \rightarrow \tau_{11}$ ,  $E[h(\mathbf{X}_1, \mathbf{Y}_1)] \rightarrow \tau_{12}$  and  $E[h(\mathbf{Y}_1, \mathbf{Y}_2)] \rightarrow \tau_{22}$  as  $p \rightarrow \infty$ . Therefore,  $\tau = E[h(\mathbf{X}_1, \mathbf{X}_2)] + E[h(\mathbf{X}_1, \mathbf{X}_2)] - 2E[h(\mathbf{X}_1, \mathbf{X}_2)] \rightarrow \tau_{11} + \tau_{22} - 2\tau_{12} = \tau_0$  as  $p \rightarrow \infty$ . Thus,  $|L(\mathbf{Z}) - \tau| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ .



- (b) The arguments of the proof for this part are similar to part (a), and we skip it.

Hence, the proof.  $\square$

#### Proof of Theorem 2.2

Recall that the prior probability of an observation  $\mathbf{Z}$  belonging to the first class is given by  $\pi_1$  ( $0 < \pi_1 < 1$ ). The misclassification probability of  $\delta_0$  is as follows: 

$$\begin{aligned}
 P[\delta_0(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] &= \pi_1 P[\delta_0(\mathbf{Z}) = 2 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\delta_0(\mathbf{Z}) = 1 \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 &= \pi_1 P[L_2(\mathbf{Z}) \leq L_1(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[L_2(\mathbf{Z}) > L_1(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_2].
 \end{aligned} \tag{A.5}$$

We have assumed that either (a)  $\nu_{11}, \nu_{12}, \nu_{22}$  are unequal, or (b)  $\nu_{11} = \nu_{12} = \nu_{22} \neq 0$ , and  $\sigma_1^2 = \sigma_2^2$  holds. As a consequence,  $\tau_0$  is strictly positive. Fix  $0 < \epsilon < \tau_0$ . Now, we have

$$P[L_2(\mathbf{Z}) \leq L_1(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] \leq P[L_2(\mathbf{Z}) - L_1(\mathbf{Z}) \leq \tau_0 - \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1]$$



$$\begin{aligned}
 &\leq \mathbb{P}[L_2(\mathbf{Z}) - L_1(\mathbf{Z}) - \tau_0 \leq -\epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 &\leq \mathbb{P}[|L_2(\mathbf{Z}) - L_1(\mathbf{Z}) - \tau_0| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 &= o(1) \text{ as } p \rightarrow \infty \text{ [using Corollary A.1(a)]}.
 \end{aligned} \tag{A.6}$$

Similarly,

$$\begin{aligned}
 \mathbb{P}[L_2(\mathbf{Z}) > L_1(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_2] &\leq \mathbb{P}[L_2(\mathbf{Z}) - L_1(\mathbf{Z}) > -\tau_0 + \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 &\leq \mathbb{P}[L_2(\mathbf{Z}) - L_1(\mathbf{Z}) + \tau_0 > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 &\leq \mathbb{P}[|L_2(\mathbf{Z}) - L_1(\mathbf{Z}) + \tau_0| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 &= o(1) \text{ as } p \rightarrow \infty \text{ [using Corollary A.1(b)]}.
 \end{aligned} \tag{A.7}$$

Combining (A.5), (A.6) and (A.7), we get  $\mathbb{P}[\delta_0(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] = o(1)$  as  $p \rightarrow \infty$ .  $\square$

**Lemma A.3** For  $j, j' \in \{1, 2\}$ , if A1 is satisfied, then

- (a)  $|\bar{T}_{jj'} - \bar{\tau}_p(j, j')| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ , and  
 (b) if  $\mathbf{Z} \sim \mathbf{F}_j$ , then  $|\bar{T}_{jj'}(\mathbf{Z}) - \bar{\tau}_p(j, j')| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ .

**Proof of Lemma A.3**

- (a) Recall the definitions of  $\bar{T}_{11}$  and  $\bar{\tau}_p(1, 1)$  given in (2.4) and (2.5), respectively. Fix  $\epsilon > 0$ . We have

$$\begin{aligned}
 &\mathbb{P}[|\bar{T}_{11} - \bar{\tau}_p(1, 1)| > \epsilon] \\
 &= \mathbb{P}\left[\left|\frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} \bar{h}_p(\mathbf{X}_i, \mathbf{X}_j) - \mathbb{E}[\bar{h}_p(\mathbf{X}_1, \mathbf{X}_2)]\right| > \epsilon\right] \\
 &= \mathbb{P}\left[\left|\frac{1}{p} \sum_{k=1}^p \frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} h(X_{ik}, X_{jk}) - \frac{1}{p} \sum_{k=1}^p \mathbb{E}[h(X_{1k}, X_{2k})]\right| > \epsilon\right] \text{ (using the definition of } \bar{h}_p) \\
 &= \mathbb{P}\left[\left|\frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} \frac{1}{p} \sum_{k=1}^p h(X_{ik}, X_{jk}) - \frac{1}{p} \sum_{k=1}^p \mathbb{E}[h(X_{1k}, X_{2k})]\right| > \epsilon\right] \\
 &\leq \mathbb{P}\left[\frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} \left|\frac{1}{p} \sum_{k=1}^p h(X_{ik}, X_{jk}) - \frac{1}{p} \sum_{k=1}^p \mathbb{E}[h(X_{1k}, X_{2k})]\right| > \epsilon\right] \text{ (using triangle inequality)} \\
 &\leq \sum_{1 \leq i \neq j \leq n_1} \sum_{k=1}^p \mathbb{P}\left[\left|\frac{1}{p} \sum_{k=1}^p \{h(X_{ik}, X_{jk}) - \mathbb{E}[h(X_{1k}, X_{2k})]\}\right| > \epsilon\right] \text{ (using the union bound)} \\
 &\leq \sum_{1 \leq i \neq j \leq n_1} \sum_{k=1}^p \frac{1}{\epsilon^2} \text{Var}\left[\frac{1}{p} \sum_{k=1}^p h(X_{ik}, X_{jk})\right] \text{ (using Chebyshev's inequality)}.
 \end{aligned} \tag{A.8}$$

Now, we will show that  $\text{Var}[\sum_{k=1}^p h(X_{ik}, X_{jk})/p]$  converges to 0 for all  $i \neq j$  as  $p \rightarrow \infty$ .

Fix  $1 \leq i, j \leq n_1$  with  $i \neq j$ . Observe that

$$\text{Var}\left[\frac{1}{p} \sum_{k=1}^p h(X_{ik}, X_{jk})\right] = \frac{1}{p^2} \sum_{k=1}^p \text{Var}[h(X_{ik}, X_{jk})] + \frac{2}{p^2} \sum_{1 \leq k < k' \leq p} \text{Cov}(h(X_{ik}, X_{jk}), h(X_{ik'}, X_{jk'})). \tag{A.9}$$

Since  $0 \leq h \leq 1$ , we have  $\text{Var}[h(X_{ik}, X_{jk})] \leq 1$  for all  $1 \leq k \leq p$ . Using the inequality  $\text{Cov}(X, Y) \leq \text{Corr}(X, Y)\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$  and the boundedness of  $h$ , we get

$$\text{Cov}(h(X_{ik}, X_{jk}), h(X_{ik'}, X_{jk'})) \leq \text{Corr}(h(X_{ik}, X_{jk}), h(X_{ik'}, X_{jk'})) \text{ for all } 1 \leq k < k' \leq p.$$

Since A1 is satisfied, from (A.9) we obtain

$$\text{Var} \left[ \frac{1}{p} \sum_{k=1}^p h(X_{ik}, X_{jk}) \right] \leq \frac{1}{p} + \frac{2}{p^2} \sum_{1 \leq k < k' \leq p} \text{Corr}(h(X_{ik}, X_{jk}), h(X_{ik'}, X_{jk'})) = o(1) \text{ as } p \rightarrow \infty.$$

It now follows from (A.8) that  $|\bar{T}_{11} - \bar{\tau}_p(1, 1)| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ . Following similar arguments, one can show that if A1 is satisfied, then both  $|\bar{T}_{12} - \bar{\tau}_p(1, 2)|$  and  $|\bar{T}_{22} - \bar{\tau}_p(2, 2)|$  converge in probability to 0 as  $p \rightarrow \infty$ .

(b) Fix  $\epsilon > 0$ , and recall the definitions of  $\bar{T}_1(\mathbf{Z})$  and  $\bar{\tau}_p(1, 1)$ . We have

$$\begin{aligned} & \mathbb{P} \left[ |\bar{T}_1(\mathbf{Z}) - \bar{\tau}_p(1, 1)| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \\ &= \mathbb{P} \left[ \left| \frac{1}{p} \sum_{k=1}^p T_{1k}(Z_k) - \frac{1}{p} \sum_{k=1}^p \mathbb{E}[h(X_{1k}, X_{2k})] \right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \\ &= \mathbb{P} \left[ \left| \frac{1}{p} \sum_{k=1}^p \frac{1}{n_1} \sum_{i=1}^{n_1} \{h(X_{ik}, Z_k) - \mathbb{E}[h(X_{1k}, Z_k) \mid \mathbf{Z} \sim \mathbf{F}_1]\} \right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \\ &= \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{p} \sum_{k=1}^p \{h(X_{ik}, Z_k) - \mathbb{E}[h(X_{1k}, Z_k) \mid \mathbf{Z} \sim \mathbf{F}_1]\} \right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \\ &\leq \mathbb{P} \left[ \frac{1}{n_1} \sum_{i=1}^{n_1} \left| \frac{1}{p} \sum_{k=1}^p \{h(X_{ik}, Z_k) - \mathbb{E}[h(X_{1k}, Z_k) \mid \mathbf{Z} \sim \mathbf{F}_1]\} \right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \text{ [using triangle inequality]} \\ &\leq \sum_{i=1}^{n_1} \mathbb{P} \left[ \left| \frac{1}{p} \sum_{k=1}^p \{h(X_{ik}, Z_k) - \mathbb{E}[h(X_{1k}, Z_k) \mid \mathbf{Z} \sim \mathbf{F}_1]\} \right| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \text{ [using the union bound]} \\ &\leq \sum_{i=1}^{n_1} \frac{1}{\epsilon^2} \text{Var} \left[ \frac{1}{p} \sum_{k=1}^p h(X_{ik}, Z_k) \mid \mathbf{Z} \sim \mathbf{F}_1 \right] \text{ [using Chebyshev's inequality]} \\ &= \sum_{i=1}^{n_1} \frac{1}{\epsilon^2} \text{Var} \left[ \frac{1}{p} \sum_{k=1}^p h(X_{ik}, X'_k) \right], \end{aligned} \tag{A.10}$$

where  $\mathbf{X}' = (X'_1, \dots, X'_p)^\top \sim \mathbf{F}_1$  and it is independent of  $\mathbf{X}_1$ . Using the boundedness of  $h$  and assumption A1, we have shown in part (a) of Lemma 3.1 that  $\text{Var} \left[ \frac{1}{p} \sum_{k=1}^p h(X_{ik}, X'_k) \right] = o(1)$  as  $p \rightarrow \infty$ . Since  $n_1$  is fixed,  $\sum_{i=1}^{n_1} \text{Var} \left[ \frac{1}{p} \sum_{k=1}^p h(X_{ik}, X'_k) \right] = o(1)$  as  $p \rightarrow \infty$ . Therefore, it follows from (A.10) that  $|\bar{T}_1(\mathbf{Z}) - \bar{\tau}_p(1, 1)|$  converges in probability to 0 as  $p \rightarrow \infty$  (when  $\mathbf{Z} \sim \mathbf{F}_1$ ).

Following similar arguments, one can prove that  $\mathbb{P} \left[ |\bar{T}_2(\mathbf{Z}) - \bar{\tau}_p(1, 2)| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1 \right]$ ,  $\mathbb{P} \left[ |\bar{T}_1(\mathbf{Z}) - \bar{\tau}_p(1, 2)| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2 \right]$  and  $\mathbb{P} \left[ |\bar{T}_2(\mathbf{Z}) - \bar{\tau}_p(2, 2)| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2 \right]$  also converge to 0 as  $p \rightarrow \infty$ .

Hence, the proof.  $\square$

### Proof of Lemma 3.1

Recall that  $\bar{L}_1(\mathbf{Z}) = \bar{T}_{11} - 2\bar{T}_1(\mathbf{Z})$ ,  $\bar{L}_2(\mathbf{Z}) = \bar{T}_{22} - 2\bar{T}_2(\mathbf{Z})$  and

$$\begin{aligned} \bar{\theta}(\mathbf{Z}) &= \frac{1}{2} \bar{T}(\bar{L}_2(\mathbf{Z}) - \bar{L}_1(\mathbf{Z})) + \frac{1}{2} (\bar{T}_{22} - \bar{T}_{11})(\bar{L}_2(\mathbf{Z}) + \bar{L}_1(\mathbf{Z}) + 2\bar{T}_{12}) \\ &= \frac{1}{2} \{(\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}) \times (\bar{L}_2(\mathbf{Z}) - \bar{L}_1(\mathbf{Z}))\} \\ &\quad + \frac{1}{2} \{(\bar{T}_{22} - \bar{T}_{11}) \times (\bar{T}_{22} - 2\bar{T}_2(\mathbf{Z}) + \bar{T}_{11} - 2\bar{T}_1(\mathbf{Z}) + 2\bar{T}_{12})\}. \end{aligned} \tag{A.11}$$

Let us denote  $\bar{L}_2(\mathbf{Z}) - \bar{L}_1(\mathbf{Z})$  by  $\bar{L}(\mathbf{Z})$  and  $\bar{T}_{22} - 2\bar{T}_2(\mathbf{Z}) + \bar{T}_{11} - 2\bar{T}_1(\mathbf{Z}) + 2\bar{T}_{12}$  by  $\bar{S}(\mathbf{Z})$ .

$$\text{We can write } \bar{\theta}(\mathbf{Z}) = \frac{1}{2} \{(\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}) \times \bar{L}(\mathbf{Z})\} + \frac{1}{2} \{(\bar{T}_{22} - \bar{T}_{11}) \times \bar{S}(\mathbf{Z})\}. \tag{A.12}$$

(a) Fix  $\epsilon > 0$ . Now,

$$\begin{aligned}
 & \mathbb{P}[|\bar{L}(\mathbf{Z}) - \bar{\tau}_p| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] = \mathbb{P}[|\bar{L}_2(\mathbf{Z}) - \bar{L}_1(\mathbf{Z}) - \bar{\tau}_p| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 & = \mathbb{P}[|\{\bar{T}_{22} - 2\bar{T}_2(\mathbf{Z}) - \bar{T}_{11} + 2\bar{T}_1(\mathbf{Z})\} - \{\bar{\tau}_p(1, 1) - 2\bar{\tau}_p(1, 2) + \bar{\tau}_p(2, 2)\}| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 & \leq \mathbb{P}[|\{\bar{T}_{22} - 2\bar{T}_2(\mathbf{Z}) - \bar{T}_{11} + 2\bar{T}_1(\mathbf{Z})\} - \{2\bar{\tau}_p(1, 1) - \bar{\tau}_p(1, 1) - 2\bar{\tau}_p(1, 2) + \bar{\tau}_p(2, 2)\}| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 & \leq \mathbb{P}\left[|\bar{T}_{11} - \bar{\tau}_p(1, 1)| > \frac{\epsilon}{4}\right] + \mathbb{P}\left[|\bar{T}_{22} - \bar{\tau}_p(2, 2)| > \frac{\epsilon}{4}\right] \\
 & + \mathbb{P}\left[2|\bar{T}_2(\mathbf{Z}) - \bar{\tau}_p(1, 2)| > \frac{\epsilon}{4} \mid \mathbf{Z} \sim \mathbf{F}_1\right] + \mathbb{P}\left[2|\bar{T}_1(\mathbf{Z}) - \bar{\tau}_p(1, 1)| > \frac{\epsilon}{4} \mid \mathbf{Z} \sim \mathbf{F}_1\right] \\
 & = o(1) \text{ as } p \rightarrow \infty \text{ [using Lemma A.3]}.
 \end{aligned} \tag{A.13}$$

Therefore, if  $\mathbf{Z} \sim \mathbf{F}_1$ , then  $|\bar{L}(\mathbf{Z}) - \bar{\tau}_p| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ . Next, we use the continuous mapping theorem and Lemma A.3 to obtain that if  $\mathbf{Z} \sim \mathbf{F}_1$ , then

$$\begin{aligned}
 & |\{\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}\} - \bar{\tau}_p| \xrightarrow{P} 0, \\
 & |\{\bar{T}_{22} - \bar{T}_{11}\} - \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\}| \xrightarrow{P} 0 \text{ and} \\
 & |\bar{S}(\mathbf{Z}) - \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\}| \xrightarrow{P} 0 \text{ as } p \rightarrow \infty.
 \end{aligned}$$

Using the continuous mapping theorem once again, we conclude from (A.12) that if  $\mathbf{Z} \sim \mathbf{F}_1$ , then

$$\left| \bar{\theta}(\mathbf{Z}) - \left\{ \frac{1}{2}\bar{\tau}_p^2 + \frac{1}{2}(\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1))^2 \right\} \right| \xrightarrow{P} 0 \Rightarrow |\bar{\theta}(\mathbf{Z}) - \bar{\psi}_p| \xrightarrow{P} 0 \text{ as } p \rightarrow \infty. \tag{A.14}$$



(b) The arguments of the proof for this part are similar to part (a), and we skip it.

□

### Proof of Theorem 3.2

(a) The misclassification probability of the classifier  $\delta_1$  can be written as

$$\begin{aligned}
 \mathbb{P}[\delta_1(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] & = \mathbb{P}[\delta_1(\mathbf{Z}) = 2, \mathbf{Z} \sim \mathbf{F}_1] + \mathbb{P}[\delta_1(\mathbf{Z}) = 1, \mathbf{Z} \sim \mathbf{F}_2] \\
 & = \pi_1 \mathbb{P}[\delta_1(\mathbf{Z}) = 2 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 \mathbb{P}[\delta_1(\mathbf{Z}) = 1 \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 & = \pi_1 \mathbb{P}[\bar{L}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 \mathbb{P}[\bar{L}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2].
 \end{aligned} \tag{A.15}$$

Since A2 is satisfied (i.e.,  $\liminf_p \bar{\tau}_p > 0$ ), we can choose  $\epsilon > 0$  such that  $\epsilon < \bar{\tau}_p$  for all  $p \geq p_0$  for some  $p_0 \in \mathbb{N}$ . Therefore, we have

$$\begin{aligned}
 \mathbb{P}[\bar{L}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] & \leq \mathbb{P}[\bar{L}(\mathbf{Z}) \leq \bar{\tau}_p - \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \\
 & \leq \mathbb{P}[\bar{L}(\mathbf{Z}) - \bar{\tau}_p \leq -\epsilon \mid \mathbf{Z} \sim \mathbf{F}_1] \leq \mathbb{P}[|\bar{L}(\mathbf{Z}) - \bar{\tau}_p| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_1]
 \end{aligned}$$

for all  $p \geq p_0$ . Now, it follows from part (a) of Lemma 3.1 that  $\mathbb{P}[\bar{L}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] = o(1)$  as  $p \rightarrow \infty$ . Similarly,

$$\begin{aligned}
 \mathbb{P}[\bar{L}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] & \leq \mathbb{P}[\bar{L}(\mathbf{Z}) > -\bar{\tau}_p + \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] \\
 & \leq \mathbb{P}[\bar{L}(\mathbf{Z}) + \bar{\tau}_p > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] \leq \mathbb{P}[|\bar{L}(\mathbf{Z}) + \bar{\tau}_p| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2]
 \end{aligned}$$

for all  $p \geq p_0$ . Since  $\mathbb{P}[|\bar{L}(\mathbf{Z}) + \bar{\tau}_p| > \epsilon \mid \mathbf{Z} \sim \mathbf{F}_2] = o(1)$  as  $p \rightarrow \infty$  (using part (b) of Lemma 3.1),  $\mathbb{P}[\bar{L}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] = o(1)$  as  $p \rightarrow \infty$ . Consequently, it follows from (A.15) that  $\mathbb{P}[\delta_1(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] = \pi_1 o(1) + \pi_2 o(1) = o(1)$  as  $p \rightarrow \infty$ .

(b) Firstly, observe that

$$\liminf_p \bar{\tau}_p > 0 \Rightarrow \liminf_p \frac{1}{2}\bar{\tau}_p^2 > 0 \Rightarrow \liminf_p \frac{1}{2}\{\bar{\tau}_p^2 + (\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1))^2\} = \liminf_p \bar{\psi}_p > 0.$$

Thus, if A2 is satisfied, then  $\liminf_p \bar{\psi}_p > 0$ . Now, let us consider the misclassification probability of  $\delta_2$ .

$$\begin{aligned} P[\delta_2(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] &= P[\delta_2(\mathbf{Z}) = 2, \mathbf{Z} \sim \mathbf{F}_1] + P[\delta_2(\mathbf{Z}) = 1, \mathbf{Z} \sim \mathbf{F}_2] \\ &= \pi_1 P[\delta_2(\mathbf{Z}) = 2 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\delta_2(\mathbf{Z}) = 1 \mid \mathbf{Z} \sim \mathbf{F}_2] \\ &= \pi_1 P[\bar{\theta}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\bar{\theta}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2]. \end{aligned} \quad (\text{A.16})$$

The arguments for the rest of the proof are similar to part (a), and we skip it.  $\square$

### Proof of Theorem 3.3

In assumption A3, we have assumed that there exists a  $p_0$  such that  $\bar{\tau}_p(1, 2)$  lies between  $\bar{\tau}_p(1, 1)$  and  $\bar{\tau}_p(2, 2)$  for all  $p \geq p_0$ . Without loss of generality, let us assume that  $\bar{\tau}_p(1, 1) < \bar{\tau}_p(2, 2)$ . As a result,

$$\bar{\tau}_p < \bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1) \text{ for all } p \geq p_0. \quad (\text{A.17})$$

Recall that

$$\begin{aligned} \Delta_1 &= P[\delta_1(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] = \pi_1 P[\bar{L}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\bar{L}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2], \text{ and} \\ \Delta_2 &= P[\delta_2(\mathbf{Z}) \neq \text{true label of } \mathbf{Z}] = \pi_1 P[\bar{\theta}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\bar{\theta}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2]. \end{aligned}$$

It follows from (A.17) that

$$\begin{aligned} P[\bar{\theta}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] &= P[\bar{\tau}_p \bar{L}(\mathbf{Z}) + \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\} \bar{S}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] \\ &\leq P[\bar{\tau}_p \{\bar{L}(\mathbf{Z}) + \bar{S}(\mathbf{Z})\} \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] \text{ for all } p \geq p_0. \end{aligned}$$

Consequently, for all  $p \geq p_0$ , we have the following:

$$\begin{aligned} &P[\bar{\theta}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] \\ &\leq P[\bar{L}(\mathbf{Z}) + \bar{S}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] \quad (\text{since } \bar{\tau}_p > 0) \\ &= P[\bar{L}(\mathbf{Z}) \leq -\bar{S}(\mathbf{Z}) \mid \mathbf{Z} \sim \mathbf{F}_1] \\ &= P[\bar{L}(\mathbf{Z}) \leq -\bar{S}(\mathbf{Z}), \bar{S}(\mathbf{Z}) \geq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + P[\bar{L}(\mathbf{Z}) \leq -\bar{S}(\mathbf{Z}), \bar{S}(\mathbf{Z}) < 0 \mid \mathbf{Z} \sim \mathbf{F}_1] \\ &\leq P[\bar{L}(\mathbf{Z}) \leq 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + P[\bar{S}(\mathbf{Z}) < 0 \mid \mathbf{Z} \sim \mathbf{F}_1]. \end{aligned} \quad (\text{A.18})$$

Similarly, one can show that

$$P[\bar{\theta}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] \leq P[\bar{L}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] + P[\bar{S}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] \text{ for all } p \geq p_0. \quad (\text{A.19})$$

Adding the two inequalities in (A.18) and (A.19), we obtain

$$\Delta_2 \leq \Delta_1 + \pi_1 P[\bar{S}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_1] + \pi_2 P[\bar{S}(\mathbf{Z}) > 0 \mid \mathbf{Z} \sim \mathbf{F}_2] \text{ for all } p \geq p_0. \quad (\text{A.20})$$

Now, it follows from part (a) of Lemma 3.1 that for  $\mathbf{Z} \sim \mathbf{F}_1$ ,  $|\bar{S}(\mathbf{Z}) - \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\}| \xrightarrow{P} 0$  as  $p \rightarrow \infty$ . Therefore, for any  $\epsilon_1 > 0$  and  $\epsilon_2 > 0$ , there exists a  $\tilde{p}_1(\epsilon_1, \epsilon_2)$  such that for all  $p \geq \tilde{p}_1(\epsilon_1, \epsilon_2)$

$$\begin{aligned} &P[|\bar{S}(\mathbf{Z}) - \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\}| > \epsilon_1 \mid \mathbf{Z} \sim \mathbf{F}_1] < \epsilon_2 \\ &\Rightarrow P[\bar{S}(\mathbf{Z}) - \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\} < -\epsilon_1 \mid \mathbf{Z} \sim \mathbf{F}_1] < \epsilon_2 \\ &\Rightarrow P[\bar{S}(\mathbf{Z}) < \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\} - \epsilon_1 \mid \mathbf{Z} \sim \mathbf{F}_1] < \epsilon_2. \end{aligned}$$

We have already assumed that  $\bar{\tau}_p(2, 2) > \bar{\tau}_p(1, 1)$  for all  $p \geq p_0$ . Define  $\lambda_0 = \liminf_p \{\bar{\tau}_p(2, 2) - \bar{\tau}_p(1, 1)\}$ . It follows from (A.17) that  $\lambda_0 \geq \liminf_p \bar{\tau}_p$ . Consequently, using assumption A2, we have  $\lambda_0 > 0$ . Therefore, it follows from the above inequality that for any  $0 < \epsilon_1 < \lambda_0$ ,

$$P[\bar{S}(\mathbf{Z}) < 0 \mid \mathbf{Z} \sim \mathbf{F}_1] < \epsilon_2 \text{ for all } p \geq \max\{\tilde{p}_1(\epsilon_1, \epsilon_2), p_0\}.$$

Following similar arguments, one can show that for any  $0 < \epsilon < \lambda_0$ , we have

$$P[\bar{S}(\mathbf{Z}) > 0 | \mathbf{Z} \sim F_2] < \epsilon_2 \text{ for all } p \geq \max\{\tilde{p}_1(\epsilon_1, \epsilon_2), p_0\}.$$

Now, it follows from (A.20) that for any  $0 < \epsilon_1 < \lambda_0$ ,

$$\begin{aligned} \Delta_2 &\leq \Delta_1 + \epsilon_2 \text{ for all } p \geq \max\{\tilde{p}_2(\epsilon_1, \epsilon_2), p_0\}, \\ \Rightarrow \Delta_2 &\leq \Delta_1 \text{ for all } p \geq p'_0 = \max\{\tilde{p}_2(\epsilon_1, \epsilon_2), p_0\} \text{ (since } \epsilon_2 > 0 \text{ is arbitrary).} \end{aligned}$$



This completes the proof. □

Let us define the following statistics:

$$\begin{aligned} T_{11k} &= \frac{1}{n_1(n_1 - 1)} \sum_{1 \leq i \neq j \leq n_1} h(X_{ik}, X_{jk}), \quad T_{12k} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_{ik}, Y_{jk}) \text{ and} \\ T_{22k} &= \frac{1}{n_2(n_2 - 1)} \sum_{1 \leq i \neq j \leq n_2} h(Y_{ik}, Y_{jk}) \text{ for } 1 \leq k \leq p_n. \end{aligned} \quad (\text{A.21})$$

Also, for  $\mathbf{z} = (z_1, \dots, z_{p_n})^\top \in \mathbb{R}^{p_n}$ , we define

$$\begin{aligned} T_{1k}(z_k) &= \frac{1}{n_1} \sum_{i=1}^{n_1} h(X_{ik}, z_k), \quad T_{2k}(z_k) = \frac{1}{n_2} \sum_{j=1}^{n_2} h(Y_{jk}, z_k), \quad L_{1k}(z_k) = T_{11k} - 2T_{1k}(z_k) \text{ and} \\ L_{2k}(z_k) &= T_{22k} - 2T_{2k}(z_k) \text{ for } 1 \leq k \leq p_n. \end{aligned} \quad (\text{A.22})$$

Observe that the estimators of  $\bar{\tau}_{11}$ ,  $\bar{\tau}_{12}$  and  $\bar{\tau}_{22}$  defined in (2.4) can be expressed as follows:

$$\begin{aligned} \bar{T}_{11} &= \frac{1}{n_1(n_1 - 1)p_n} \sum_{k=1}^{p_n} \sum_{1 \leq i \neq j \leq n_1} h(X_{ik}, X_{jk}), \quad \bar{T}_{12} = \frac{1}{n_1 n_2 p_n} \sum_{k=1}^{p_n} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_{ik}, Y_{jk}) \text{ and} \\ \bar{T}_{22} &= \frac{1}{n_2(n_2 - 1)p_n} \sum_{k=1}^{p_n} \sum_{1 \leq i \neq j \leq n_2} h(Y_{ik}, Y_{jk}), \\ \text{i.e., } \bar{T}_{11} &= \frac{1}{p_n} \sum_{k=1}^{p_n} T_{11k}, \quad \bar{T}_{12} = \frac{1}{p_n} \sum_{k=1}^{p_n} T_{12k} \text{ and } \bar{T}_{22} = \frac{1}{p_n} \sum_{k=1}^{p_n} T_{22k}. \end{aligned}$$

Similarly, for  $\mathbf{z} \in \mathbb{R}^{p_n}$ , we can write

$$\bar{T}_1(\mathbf{z}) = \frac{1}{p_n} \sum_{k=1}^{p_n} T_{1k}(z_k) \text{ and } \bar{T}_2(\mathbf{z}) = \frac{1}{p_n} \sum_{k=1}^{p_n} T_{2k}(z_k).$$

Recall the definitions of  $\bar{L}_1(\mathbf{z})$ ,  $\bar{L}_2(\mathbf{z})$  and  $\bar{\theta}(\mathbf{z})$  given in (A.11). We now derive upper bounds on the rates of convergence of these random variables.

First, we present the bounded differences inequality that will be used to derive concentration bounds.

Given vectors  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$  and an index  $l \in \{1, \dots, n\}$ , we define a new vector  $\mathbf{x}^{\setminus l} \in \mathbb{R}^n$  as follows:

$$\mathbf{x}^{\setminus l} = \begin{cases} x_j, & \text{if } j \neq l, \\ x'_l, & \text{if } j = l. \end{cases} \quad (\text{A.23})$$

With this notation, we say that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies the bounded difference inequality with parameters  $(M_1, \dots, M_n)^\top$  if

$$|f(\mathbf{x}) - f(\mathbf{x}^{\setminus l})| \leq M_l \text{ for each } l = 1, \dots, n \text{ and for all } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^n.$$



**Lemma A.4** (*Wainwright, 2019, page 37*) Suppose that  $f$  satisfies the bounded difference property (A.23) with parameters  $(M_1, \dots, M_n)^\top$  and that the random vector  $\mathbf{U} = (U_1, \dots, U_n)^\top$  has independent components. Then,

$$\mathbb{P}[|f(\mathbf{U}) - \mathbb{E}[f(\mathbf{U})]| > \epsilon] \leq 2e^{-\frac{2\epsilon^2}{\sum_{l=1}^n M_l^2}} \text{ for all } \epsilon > 0.$$

Using Lemma A.4, we first derive the rates of convergence of  $\bar{T}_{jj'}$  and  $\bar{T}_i(\mathbf{z})$  for  $j, j' \in \{1, 2\}$  and  $\mathbf{z} \in \mathbb{R}^{p_n}$ .

**Lemma A.5** Fix  $0 < \gamma < 1/2$ . There exist positive constants  $a_{jj'}, b_j$  for  $j, j' \in \{1, 2\}$  such that

- (a)  $\mathbb{P}[|\bar{T}_{jj'} - \bar{\tau}_p(j, j')| > n^{-\gamma}] \leq O(p_n e^{-a_{jj'} n^{1-2\gamma}})$  and  
 (b)  $\mathbb{P}[|\bar{T}_i(\mathbf{z}) - \mathbb{E}[\bar{T}_i(\mathbf{z})]| > n^{-\gamma}] \leq O(p_n e^{-b_i n^{1-2\gamma}})$  for all  $\mathbf{z} \in \mathbb{R}^{p_n}$ .

### Proof of Lemma A.5

- (a) Fix  $k \in \{1, \dots, p_n\}$ . Recall the definitions of  $T_{11k}, T_{22k}$  and  $T_{12k}$  in (A.21) and note that the first two random variables are one sample U-statistics with kernel of order 2, while the third random variable is a two sample U-statistic with kernel of order (1,1).

The random vector  $\mathcal{X}_k = (X_{1k}, \dots, X_{n_1 k})^\top$  has independent components. Observe that the random variable  $T_{11k}$  is a function of  $\mathcal{X}_k$ , say  $f(\mathcal{X}_k)$ . Since  $|h| < 1$ , for any given co-ordinate  $l \in \{1, \dots, n_1\}$ , we have

$$|f(\mathcal{X}_k) - f(\mathcal{X}_k^{\setminus l})| \leq \frac{2}{n_1(n_1 - 1)} \sum_{j \neq l} |h(X_{jk}, X_{lk}) - h(X_{jk}, X'_{lk})| \leq 2(n_1 - 1) \frac{2}{n_1(n_1 - 1)} = \frac{4}{n_1}.$$

So, the bounded difference property holds with parameter  $M_l = 4/n_1$  in each coordinate. We conclude from Lemma A.4 that

$$\mathbb{P}[|T_{11k} - \mathbb{E}[T_{11k}]| > n^{-\gamma}] \leq 2e^{-\frac{n_1 n^{-2\gamma}}{8}}. \quad (\text{A.24})$$

Since  $\lim_{n \rightarrow \infty} n_1/n = \pi_1 < 1$ , there exist constants  $a_{11} > 0$  and  $N \in \mathbb{N}$  such that

$$\mathbb{P}[|T_{11k} - \mathbb{E}[T_{11k}]| \geq n^{-\gamma}] \leq 2e^{-a_{11} n^{1-2\gamma}} \text{ for all } n \geq N. \quad (\text{A.25})$$

Clearly, (A.25) is true for all  $1 \leq k \leq p_n$ . So, we have

$$\begin{aligned} \mathbb{P}[|T_{11k} - \mathbb{E}[T_{11k}]| \geq n^{-\gamma}] &\leq O\left(e^{-a_{11} n^{1-2\gamma}}\right) \text{ for all } 1 \leq k \leq p_n \\ \Rightarrow \sum_{k=1}^{p_n} \mathbb{P}[|T_{11k} - \mathbb{E}[T_{11k}]| \geq n^{-\gamma}] &\leq O\left(p_n e^{-a_{11} n^{1-2\gamma}}\right) \\ \Rightarrow \mathbb{P}\left[\frac{1}{p_n} \sum_{k=1}^{p_n} |T_{11k} - \mathbb{E}[T_{11k}]| \geq n^{-\gamma}\right] &\leq O\left(p_n e^{-a_{11} n^{1-2\gamma}}\right) \\ \Rightarrow \mathbb{P}\left[\left|\frac{1}{p_n} \sum_{k=1}^{p_n} (T_{11k} - \mathbb{E}[T_{11k}])\right| \geq n^{-\gamma}\right] &\leq O\left(p_n e^{-a_{11} n^{1-2\gamma}}\right) \\ \Rightarrow \mathbb{P}[|\bar{T}_{11} - \bar{\tau}_p(1, 1)| \geq n^{-\gamma}] &\leq O\left(p_n e^{-a_{11} n^{1-2\gamma}}\right) \left[\sum_{k=1}^{p_n} \mathbb{E}[T_{11k}]/p_n = \bar{\tau}_{p_n}(1, 1)\right]. \end{aligned} \quad (\text{A.26})$$

Following similar arguments, it can be shown that there exist positive constants  $a_{12}$  and  $a_{22}$  such that

$$\mathbb{P}[|\bar{T}_{12} - \bar{\tau}_p(1, 2)| > n^{-\gamma}] \leq O(p_n e^{-a_{12} n^{1-2\gamma}}) \text{ and } \mathbb{P}[|\bar{T}_{22} - \bar{\tau}_p(2, 2)| > n^{-\gamma}] \leq O(p_n e^{-a_{22} n^{1-2\gamma}}). \quad (\text{A.27})$$

(b) Recall the definition of  $\bar{T}_1(\mathbf{z})$  from (A.22) and observe that for each  $\mathbf{z} \in \mathbb{R}^{p_n}$ , we have the following:

$$\begin{aligned}
 & \mathbb{P} [|\bar{T}_1(\mathbf{z}) - E[\bar{T}_1(\mathbf{z})]| > n^{-\gamma}] \\
 &= \mathbb{P} \left[ \left| \frac{1}{p_n} \sum_{k=1}^{p_n} T_{1k}(z_k) - \frac{1}{p_n} \sum_{k=1}^{p_n} E[T_{1k}(z_k)] \right| > n^{-\gamma} \right] \\
 &\leq \mathbb{P} \left[ \frac{1}{p_n} \sum_{k=1}^{p_n} |T_{1k}(z_k) - E[T_{1k}(z_k)]| > n^{-\gamma} \right] \\
 &\leq \sum_{k=1}^{p_n} \mathbb{P} [|T_{1k}(z_k) - E[T_{1k}(z_k)]| > n^{-\gamma}] \\
 &\leq \sum_{k=1}^{p_n} \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} h(X_{ik}, z_k) - \frac{1}{n_1} \sum_{i=1}^{n_1} E[h(X_{ik}, z_k)] \right| > n^{-\gamma} \right] \\
 &= \sum_{k=1}^{p_n} \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{h(X_{ik}, z_k) - E[h(X_{ik}, z_k)]\} \right| > n^{-\gamma} \right]. \tag{A.28}
 \end{aligned}$$

Here,  $\sum_{i=1}^{n_1} h(X_{ik}, z_k)/n_1$  is an average of independently distributed random variables for each  $\mathbf{z} \in \mathbb{R}^{p_n}$ . Using Hoeffding's inequality, we obtain the following:

$$\begin{aligned}
 & \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{h(X_{ik}, z_k) - E[h(X_{ik}, z_k)]\} \right| > n^{-\gamma} \right] \leq 2e^{-2n_1 n^{-2\gamma}} \text{ for all } 1 \leq k \leq p_n \\
 &\Rightarrow \sum_{k=1}^{p_n} \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{h(X_{ik}, z_k) - E[h(X_{ik}, z_k)]\} \right| > n^{-\gamma} \right] \leq 2p_n e^{-2n_1 n^{-2\gamma}} \\
 &\Rightarrow \sum_{k=1}^{p_n} \mathbb{P} \left[ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{h(X_{ik}, z_k) - E[h(X_{1k}, z_k)]\} \right| > n^{-\gamma} \right] = O(p_n e^{-b_1 n^{1-2\gamma}}) \text{ for some } b_1 > 0. \tag{A.29}
 \end{aligned}$$

Combining (A.28) and (A.29), for every  $\mathbf{z} \in \mathbb{R}^{p_n}$ , we obtain

$$\mathbb{P} [|\bar{T}_1(\mathbf{z}) - E[\bar{T}_1(\mathbf{z})]| > n^{-\gamma}] \leq O(p_n e^{-b_1 n^{1-2\gamma}}) \text{ for some } b_1 > 0.$$

Similarly, one can show that there exists a constant  $b_2 > 0$  such that

$$\mathbb{P} [|\bar{T}_2(\mathbf{z}) - E[\bar{T}_2(\mathbf{z})]| > n^{-\gamma}] \leq O(p_n e^{-b_2 n^{1-2\gamma}}).$$

Hence, the proof.  $\square$

**Lemma A.6** Suppose  $\mathbb{P}[|X_n - a_0| > \epsilon] = O(p_n e^{-M_1 n \epsilon^2})$  and  $\mathbb{P}[|Y_n - b_0| > \epsilon] = O(p_n e^{-M_2 n \epsilon^2})$  for all  $\epsilon > 0$  where  $\max\{|a_0|, |b_0|\} > 0$  and  $M_1, M_2$  are positive constants. Then, there exists a positive constant  $M_3$  such that  $\mathbb{P}[|X_n Y_n - a_0 b_0| > \epsilon] = O(p_n e^{-M_3 n \epsilon^2})$  for all  $\epsilon > 0$ .

**Proof:** Define  $c_0 = \max\{|a_0|, |b_0|\}$ . Using triangle inequality, we get

$$\begin{aligned}
 & |X_n Y_n - a_0 b_0| \leq |X_n Y_n - b_0 X_n - a_0 Y_n + a_0 b_0| + |b_0| |X_n - a_0| + |a_0| |Y_n - b_0| \\
 &\Rightarrow |X_n Y_n - a_0 b_0| \leq |X_n - a_0| |Y_n - b_0| + |b_0| |X_n - a_0| + |a_0| |Y_n - b_0| \\
 &\Rightarrow |X_n Y_n - a_0 b_0| \leq |X_n - a_0| |Y_n - b_0| + c_0 (|X_n - a_0| + |Y_n - b_0|).
 \end{aligned}$$

Therefore,  $|X_n - a_0| \leq \epsilon$  and  $|Y_n - b_0| \leq \epsilon$  implies that  $|X_n Y_n - a_0 b_0| \leq \epsilon^2 + 2c_0 \epsilon$ . We choose  $M$  such that  $M > 2 + \epsilon/c_0$ . Therefore,  $\epsilon^2 + 2c_0 \epsilon \leq M c_0 \epsilon$ . Now,

$$\mathbb{P}[|X_n - a_0| \leq \epsilon, |Y_n - b_0| \leq \epsilon] \leq \mathbb{P}[|X_n Y_n - a_0 b_0| \leq \epsilon^2 + 2c_0 \epsilon]$$

$$\begin{aligned}
 &\Rightarrow \mathbb{P}[|X_n - a_0| \leq \epsilon, |Y_n - b_0| \leq \epsilon] \leq \mathbb{P}[|X_n Y_n - a_0 b_0| \leq M c_0 \epsilon] \\
 &\Rightarrow \mathbb{P}[|X_n Y_n - a_0 b_0| > M c_0 \epsilon] \leq \mathbb{P}[|X_n - a_0| > \epsilon] + \mathbb{P}[|Y_n - b_0| > \epsilon] \\
 &\Rightarrow \mathbb{P}[|X_n Y_n - a_0 b_0| > M c_0 \epsilon] \leq O(p_n e^{-M_1 n \epsilon^2}) + O(p_n e^{-M_2 n \epsilon^2}) \\
 &\Rightarrow \mathbb{P}[|X_n Y_n - a_0 b_0| > M c_0 \epsilon] \leq O(p_n e^{-\min\{M_1, M_2\} n \epsilon^2}) \\
 &\Rightarrow \mathbb{P}[|X_n Y_n - a_0 b_0| > \epsilon] \leq O(p_n e^{-\frac{\min\{M_1, M_2\}}{M c_0} n \epsilon^2}).
 \end{aligned}$$

Therefore,  $\mathbb{P}[|X_n - a_0| \leq \epsilon, |Y_n - b_0| \leq \epsilon] \leq O(p_n e^{-\frac{\min\{M_1, M_2\}}{M c_0} n \epsilon^2})$  for all  $\epsilon > 0$  with  $M > 2 + c_0/\epsilon$ . Hence, the proof.  $\square$

### Proof of Lemma 3.4

(a) For a  $\mathbf{z} \in \mathbb{R}^{p_n}$ , recall the definitions of  $\bar{L}(\mathbf{z})$  and  $\bar{L}_0(\mathbf{z})$  given in Section 3.2. For any  $0 < \gamma < 1/2$ , we have

$$\begin{aligned}
 &\mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] \\
 &= \mathbb{P}[|\bar{L}_2(\mathbf{z}) - \bar{L}_1(\mathbf{z}) - \bar{L}_2^0(\mathbf{z}) + \bar{L}_1^0(\mathbf{z})| > n^{-\gamma}] \\
 &= \mathbb{P}[|\bar{T}_{22} - 2\bar{T}_2(\mathbf{z}) - \bar{T}_{11} + 2\bar{T}_1(\mathbf{z}) - \bar{\tau}_{p_n}(2, 2) + 2E[\bar{h}_{p_n}(\mathbf{Y}_1, \mathbf{z})] - \bar{\tau}_{p_n}(1, 1) + 2E[\bar{h}_{p_n}(\mathbf{X}_1, \mathbf{z})]| > n^{-\gamma}] \\
 &\leq \mathbb{P}\left[|\bar{T}_{11} - \bar{\tau}_{p_n}(1, 1)| > \frac{n^{-\gamma}}{4}\right] + \mathbb{P}\left[|\bar{T}_{22} - \bar{\tau}_{p_n}(2, 2)| > \frac{n^{-\gamma}}{4}\right] \\
 &+ \mathbb{P}\left[|\bar{T}_1(\mathbf{z}) - E[\bar{h}_{p_n}(\mathbf{X}_1, \mathbf{z})]| > \frac{n^{-\gamma}}{2}\right] + \mathbb{P}\left[|\bar{T}_2(\mathbf{z}) - E[\bar{h}_{p_n}(\mathbf{Y}_1, \mathbf{z})]| > \frac{n^{-\gamma}}{2}\right] \\
 &= P_1 + P_2 + P_3 + P_4.
 \end{aligned} \tag{A.30}$$

We have already proved in part (a) of Lemma A.5 that  $P_1 \leq O(p_n e^{-a_{11}^* n^{1-2\gamma}})$  and  $P_2 \leq O(p_n e^{-a_{22}^* n^{1-2\gamma}})$  for some positive constants  $a_{11}^*$  and  $a_{22}^*$ . Now, let us consider the term  $P_3$ . Observe that

$$P_3 = \mathbb{P}\left[|\bar{T}_2(\mathbf{z}) - E[\bar{h}_{p_n}(\mathbf{X}_1, \mathbf{z})]| > \frac{n^{-\gamma}}{2}\right] = \mathbb{P}\left[|\bar{T}_1(\mathbf{z}) - E[\bar{T}_1(\mathbf{z})]| > \frac{n^{-\gamma}}{2}\right]$$

It is shown in part (b) of Lemma A.5 that

$$\mathbb{P}\left[|\bar{T}_1(\mathbf{z}) - E[\bar{T}_1(\mathbf{z})]| > \frac{n^{-\gamma}}{2}\right] \leq O(p_n e^{-b_1^* n^{1-2\gamma}}) \text{ for some positive constant } b_1^*.$$

Therefore,  $P_3 \leq O(p_n e^{-b_1^* n^{1-2\gamma}})$ . Similarly,  $P_4 \leq O(p_n e^{-b_2^* n^{1-2\gamma}})$  for some positive constant  $b_2^*$ . It follows from (A.30) that

$$\begin{aligned}
 &\mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] \\
 &\leq O(p_n e^{-a_{11}^* n^{1-2\gamma}}) + O(p_n e^{-a_{22}^* n^{1-2\gamma}}) + O(p_n e^{-b_1^* n^{1-2\gamma}}) + O(p_n e^{-b_2^* n^{1-2\gamma}}) \\
 &= O(p_n e^{-B_0^* n^{1-2\gamma}}), \text{ where } B_0^* = \min\{a_{11}^*, a_{22}^*, b_1^*, b_2^*\}.
 \end{aligned}$$

Recall that there exist  $M > 0$  and  $N \in \mathbb{N}$  such that

$$p_n \leq e^{M n^\beta} \Rightarrow p_n e^{-B_0^* n^{1-2\gamma}} \leq e^{-\{B_0^* n^{1-2\gamma} - M n^\beta\}} \Rightarrow p_n e^{-B_0^* n^{1-2\gamma}} \leq e^{-B_0 \{n^{1-2\gamma} - n^\beta\}}$$

for all  $n \geq N$ , where  $B_0 = \min\{B_0^*, M\}$ . Therefore,  $\mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] \leq O(e^{-B_0 \{n^{1-2\gamma} - n^\beta\}})$ .

- (b) Now, we derive a rate of convergence for the random variable  $\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})$  for  $\mathbf{z} \in \mathbb{R}^{p_n}$ . As defined in (A.11), we have

$$\bar{\theta}(\mathbf{z}) = \frac{1}{2} \{ (\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}) \times \bar{L}(\mathbf{z}) \} + \frac{1}{2} \{ (\bar{T}_{22} - \bar{T}_{11}) \times \bar{S}(\mathbf{z}) \},$$

where  $\bar{L}(\mathbf{z}) = \bar{T}_{22} - 2\bar{T}_2(\mathbf{z}) - \bar{T}_{11} + 2\bar{T}_1(\mathbf{z})$  and  $\bar{S}(\mathbf{z}) = \bar{T}_{22} - 2\bar{T}_2(\mathbf{z}) + \bar{T}_{11} - 2\bar{T}_1(\mathbf{z}) + 2\bar{T}_{12}$ . Further,  $\bar{\theta}^0(\mathbf{z})$  is defined as

$$\begin{aligned} \bar{\theta}^0(\mathbf{z}) &= \frac{\bar{\tau}_{p_n}}{2} \{ \bar{\tau}_{p_n}(2, 2) - 2E[\bar{h}_{p_n}(\mathbf{Y}_1, \mathbf{z}) - \bar{\tau}_{p_n}(1, 1) + 2E[\bar{h}_{p_n}(\mathbf{X}_1, \mathbf{z})]] \} \\ &\quad + \frac{1}{2} (\bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1)) \{ \bar{\tau}_{p_n}(2, 2) - 2E[\bar{h}_{p_n}(\mathbf{Y}_1, \mathbf{z})] + \bar{\tau}_{p_n}(1, 1) - 2E[\bar{h}_{p_n}(\mathbf{X}_1, \mathbf{z})] + 2\bar{\tau}_{p_n}(1, 2) \} \\ \Rightarrow \bar{\theta}^0(\mathbf{z}) &= \frac{\bar{\tau}_{p_n}}{2} E[\bar{L}(\mathbf{z})] + \frac{1}{2} (\bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1)) E[\bar{S}(\mathbf{z})] \end{aligned}$$

Note that  $E[\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}] = \bar{\tau}_{p_n}$  and  $E[\bar{T}_{22} - \bar{T}_{11}] = \bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1)$ . It follows from part (a) of Lemma A.5 that there exist positive constants  $c_1$  and  $c_2$  such that

$$\begin{aligned} P[|\{\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}\} - \bar{\tau}_{p_n}| > n^{-\gamma}] &\leq O(p_n e^{-c_1 n^{1-2\gamma}}) \text{ and} \\ P[|\{\bar{T}_{22} - \bar{T}_{11}\} - \{\bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1)\}| > n^{-\gamma}] &\leq O(p_n e^{-c_2 n^{1-2\gamma}}). \end{aligned} \quad (\text{A.31})$$

Part (b) of Lemma A.5 suggests that there exist positive constants  $c_3$  and  $c_4$  such that

$$\begin{aligned} P[|\bar{L}(\mathbf{z}) - E[\bar{L}(\mathbf{z})]| > n^{-\gamma}] &\leq O(p_n e^{-c_3 n^{1-2\gamma}}) \text{ and} \\ P[|\bar{S}(\mathbf{z}) - E[\bar{S}(\mathbf{z})]| > n^{-\gamma}] &\leq O(p_n e^{-c_4 n^{1-2\gamma}}) \text{ for all } \mathbf{z} \in \mathbb{R}^{p_n}. \end{aligned} \quad (\text{A.32})$$

Now, for  $\mathbf{z} \in \mathbb{R}^{p_n}$ , we have

$$\begin{aligned} P[|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})| > n^{-\gamma}] &\leq P \left[ \left| \frac{1}{2} \{ (\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}) \times \bar{L}(\mathbf{z}) \} - \frac{\bar{\tau}_{p_n}}{2} E[\bar{L}(\mathbf{z})] \right| > \frac{n^{-\gamma}}{2} \right] \\ &\quad + P \left[ \left| \frac{1}{2} \{ (\bar{T}_{22} - \bar{T}_{11}) \times \bar{S}(\mathbf{z}) \} - \frac{1}{2} \{ \bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1) \} E[\bar{S}(\mathbf{z})] \right| > \frac{n^{-\gamma}}{2} \right]. \end{aligned} \quad (\text{A.33})$$

Combining (A.31) and (A.32) with Lemma A.6, we conclude that there exists a constant  $c_{10}$  such that

$$P \left[ \left| \frac{1}{2} \{ (\bar{T}_{11} - 2\bar{T}_{12} + \bar{T}_{22}) \times \bar{L}(\mathbf{z}) \} - \frac{\bar{\tau}_{p_n}}{2} E[\bar{L}(\mathbf{z})] \right| > \frac{n^{-\gamma}}{2} \right] \leq O(p_n e^{-c_{10} n^{1-2\gamma}}). \quad (\text{A.34})$$

Similarly, there exists a constant  $c_{11} > 0$  such that

$$P \left[ \left| \frac{1}{2} \{ (\bar{T}_{22} - \bar{T}_{11}) \times \bar{S}(\mathbf{z}) \} - \frac{1}{2} \{ \bar{\tau}_{p_n}(2, 2) - \bar{\tau}_{p_n}(1, 1) \} E[\bar{S}(\mathbf{z})] \right| > \frac{n^{-\gamma}}{2} \right] \leq O(p_n e^{-c_{11} n^{1-2\gamma}}). \quad (\text{A.35})$$

Define  $B_1^* = \min\{c_{10}, c_{11}\}$ . Now, it follows from (A.33), (A.34) and (A.35) that

$$P[|\bar{\theta}(\mathbf{z}) - \bar{\theta}^0(\mathbf{z})| > n^{-\gamma}] \leq O(p_n e^{-B_1^* n^{1-2\gamma}}) \text{ for all } \mathbf{z} \in \mathbb{R}^{p_n}.$$

Since there exist  $M > 0$  and  $N \in \mathbb{N}$  such that

$$p_n \leq e^{Mn^\beta} \Rightarrow p_n e^{-B_1^* n^{1-2\gamma}} \leq e^{-B_1 \{n^{1-2\gamma} - n^\beta\}} \text{ for all } n \geq N,$$

where  $B_1 = \min\{B_1^*, M\}$ . Therefore,  $P[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] \leq O(e^{-B_1 \{n^{1-2\gamma} - n^\beta\}})$  for all  $\mathbf{z} \in \mathbb{R}^{p_n}$ .

Hence, the proof.  $\square$

### Proof of Theorem 3.5

Let  $l_{\mathbf{Z}}$  denote the true class label of  $\mathbf{Z}$  with  $P[l_{\mathbf{Z}} = j] = \pi_j$  where  $\pi_1 + \pi_2 = 1$ . Therefore,  $\mathbf{Z} \mid l_{\mathbf{Z}} = 1 \sim \mathbf{F}_1$  and  $\mathbf{Z} \mid l_{\mathbf{Z}} = 2 \sim \mathbf{F}_2$ . The unconditional distribution of  $\mathbf{Z}$  is defined as  $\mathbf{H}(\mathbf{z}) = \pi_1 \mathbf{F}_1(\mathbf{z}) + \pi_2 \mathbf{F}_2(\mathbf{z})$  for  $\mathbf{z} \in \mathbb{R}^{p_n}$ .

- (a) Recall that the misclassification probabilities of  $\delta_1$  and  $\delta_1^0$  are defined as  $\Delta_1 = P[\delta_1(\mathbf{Z}) \neq l_{\mathbf{Z}}]$  and  $\Delta_1^0 = P[\delta_1^0(\mathbf{Z}) \neq l_{\mathbf{Z}}]$ , respectively. Now,

$$\begin{aligned}
 & \Delta_1 - \Delta_1^0 \\
 &= P[\delta_1(\mathbf{Z}) \neq l_{\mathbf{Z}}] - P[\delta_1^0(\mathbf{Z}) \neq l_{\mathbf{Z}}] \\
 &= \int \{P[\delta_1(\mathbf{z}) \neq l_{\mathbf{z}}] - P[\delta_1^0(\mathbf{z}) \neq l_{\mathbf{z}}]\} d\mathbf{H}(\mathbf{z}) \\
 &= \int \{P[\delta_1^0(\mathbf{z}) = l_{\mathbf{z}}] - P[\delta_1(\mathbf{z}) = l_{\mathbf{z}}]\} d\mathbf{H}(\mathbf{z}) \\
 &= \int \{I[\delta_1^0(\mathbf{z}) = 1]P[l_{\mathbf{z}} = 1] + I[\delta_1^0(\mathbf{z}) = 0]P[l_{\mathbf{z}} = 0] - P[\delta_1(\mathbf{z}) = 1]P[l_{\mathbf{z}} = 1] + P[\delta_1(\mathbf{z}) = 0]P[l_{\mathbf{z}} = 0]\} d\mathbf{H}(\mathbf{z}) \\
 &= \int \{(I[\delta_1^0(\mathbf{z}) = 1] - P[\delta_1(\mathbf{z}) = 1])P[l_{\mathbf{z}} = 1] + (I[\delta_1^0(\mathbf{z}) = 0] - P[\delta_1(\mathbf{z}) = 0])P[l_{\mathbf{z}} = 0]\} d\mathbf{H}(\mathbf{z}) \\
 &= \int (I[\delta_1^0(\mathbf{z}) = 1] - E[I[\delta_1(\mathbf{z}) = 1]])(2P[l_{\mathbf{z}} = 1] - 1) d\mathbf{H}(\mathbf{z}) \\
 &\leq \int |E[I[\delta_1^0(\mathbf{z}) = 1] - I[\delta_1(\mathbf{z}) = 1]]| |2P[l_{\mathbf{z}} = 1] - 1| d\mathbf{H}(\mathbf{z}) \\
 &= \int E[|I[\delta_1^0(\mathbf{z}) = 1] - I[\delta_1(\mathbf{z}) = 1]|] d\mathbf{H}(\mathbf{z}) \\
 &= \int E[I[\delta_1^0(\mathbf{z}) \neq \delta_1(\mathbf{z})]] d\mathbf{H}(\mathbf{z}) \\
 &= \int P[\delta_1^0(\mathbf{z}) \neq \delta_1(\mathbf{z})] d\mathbf{H}(\mathbf{z}) \\
 &= \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0] d\mathbf{H}(\mathbf{z}) + \int P[\bar{L}(\mathbf{z}) > 0, \bar{L}^0(\mathbf{z}) \leq 0] d\mathbf{H}(\mathbf{z}) \\
 &= P_1 + P_2.
 \end{aligned} \tag{A.36}$$

Consider the first term. For any  $\gamma > 0$ , we have the following:

$$\begin{aligned}
 P_1 &= \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0] d\mathbf{H}(\mathbf{z}) \\
 &= \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0, |\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &\quad + \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0, |\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &\leq \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0, |\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) + \int P[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &= P_{11}(\gamma) + \int P[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}).
 \end{aligned} \tag{A.37}$$

Note that

$$\begin{aligned}
 P_{11}(\gamma) &= \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0, |\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &= \int P[\bar{L}(\mathbf{z}) \leq 0, \bar{L}^0(\mathbf{z}) > 0, -\bar{L}(\mathbf{z}) + \bar{L}^0(\mathbf{z}) \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z})
 \end{aligned}$$



$$\begin{aligned}
 &\leq \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq n^{-\gamma}, \bar{L}^0(\mathbf{z}) > 0, \bar{L}(\mathbf{z}) \leq 0] d\mathbf{H}(\mathbf{z}) \\
 &\leq \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq n^{-\gamma}, \bar{L}^0(\mathbf{z}) > 0] d\mathbf{H}(\mathbf{z}) = \mathbb{P}[0 < \bar{L}^0(\mathbf{Z}) \leq n^{-\gamma}].
 \end{aligned} \tag{A.38}$$

Combining (A.37) and (A.38), we have

$$P_1 \leq \mathbb{P}[0 < \bar{L}^0(\mathbf{Z}) \leq n^{-\gamma}] + \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}). \tag{A.39}$$

Following similar arguments, we can write  $P_2$  as

$$\begin{aligned}
 P_2 &= \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq 0, \bar{L}(\mathbf{z}) > 0] d\mathbf{H}(\mathbf{z}) \\
 &\leq \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq 0, \bar{L}(\mathbf{z}) > 0, |\bar{L}^0(\mathbf{z}) - \bar{L}(\mathbf{z})| \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) + \int \mathbb{P}[|\bar{L}^0(\mathbf{z}) - \bar{L}(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &= \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq 0, \bar{L}(\mathbf{z}) > 0, |\bar{L}^0(\mathbf{z}) - \bar{L}(\mathbf{z})| \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) + \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &\leq \int \mathbb{P}[\bar{L}^0(\mathbf{z}) \leq 0, \bar{L}(\mathbf{z}) > 0, -\bar{L}^0(\mathbf{z}) + \bar{L}(\mathbf{z}) \leq n^{-\gamma}] d\mathbf{H}(\mathbf{z}) + \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &\leq \int \mathbb{P}[-n^{-\gamma} < \bar{L}^0(\mathbf{z}) \leq 0] d\mathbf{H}(\mathbf{z}) + \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \\
 &= \mathbb{P}[-n^{-\gamma} < \bar{L}^0(\mathbf{Z}) \leq 0] + \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}).
 \end{aligned} \tag{A.40}$$

Combining (A.36), (A.39) and (A.40), we obtain

$$\Delta_1 - \Delta_1^0 \leq \mathbb{P}[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}] + 2 \int \mathbb{P}[|\bar{L}(\mathbf{z}) - \bar{L}^0(\mathbf{z})| > n^{-\gamma}] d\mathbf{H}(\mathbf{z}) \text{ for all } \gamma > 0.$$

Using part (a) of Lemma 3.4, it now follows that

$$\Delta_1 - \Delta_1^0 \leq \mathbb{P}[|\bar{L}^0(\mathbf{Z})| < n^{-\gamma}] + O\left(e^{-B_0\{n^{1-2\gamma} - n^\beta\}}\right) \text{ for all } 0 < \gamma < (1 - \beta)/2.$$



(b) The arguments for this part of the proof are similar to part (a), and we skip it.

□

## B TABLES AND ADDITIONAL MATERIAL

Table 1: Estimates of  $\bar{\tau}_p(1, 1)$ ,  $\bar{\tau}_p(1, 2)$  and  $\bar{\tau}_p(2, 2)$  in the simulated examples (along with the standard errors in parentheses) based on 100 replications.

| Example | $\bar{T}_{11}$     | $\bar{T}_{12}$     | $\bar{T}_{22}$     | $\bar{T}_{12} \geq \min\{\bar{T}_{11}, \bar{T}_{22}\}$ |
|---------|--------------------|--------------------|--------------------|--|
| 1       | 0.1562<br>(0.0019) | 0.1446<br>(0.0020) | 0.1273<br>(0.0022) | True   |
| 2       | 0.0909<br>(0.0014) | 0.0984<br>(0.0010) | 0.1109<br>(0.0015) | True   |
| 3       | 0.0857<br>(0.0018) | 0.0821<br>(0.0016) | 0.1018<br>(0.0027) | False  |
| 4       | 0.0857<br>(0.0018) | 0.0748<br>(0.0016) | 0.0545<br>(0.0016) | True   |
| 5       | 0.2077<br>(0.0005) | 0.2106<br>(0.0004) | 0.2136<br>(0.0004) | True   |

### B.1 Details on Implementation of Popular Classifiers

- GLMNET: The R-package `glmnet` was used for the implementation of GLMNET. The tuning parameter  $\alpha$  in the elastic-net penalty term was kept fixed at the default value 1. The weight  $\lambda$  of the penalty term was chosen by cross-validation using the function `cv.glmnet` with default values of its arguments.
- 1NN: The `knn1` function from the R-package `class` was used for implementation of the usual 1-nearest neighbor classifier.
- NN-RAND: The function `classify` from the package `RandPro` was used with default values of the arguments.
- NNET: We used `nnet` from the package `nnet` to fit a single-hidden-layer neural network with default parameters. The number of units in the hidden layer were allowed to vary in the set  $\{1, 3, 5, 10\}$ , and the minimum misclassification rate was reported as NNET.
- SVM: The R package `e1071` was used for implementing SVM with linear and RBF kernel. For the RBF kernel, i.e.,  $K_\theta(\mathbf{x}, \mathbf{y}) = \exp\{-\theta\|\mathbf{x} - \mathbf{y}\|^2\}$ , we considered the default value of the tuning parameter  $\theta$ , i.e.,  $\theta = 1/p$ .

Table 3: Average time (in seconds) taken by the classifiers to classify 200 test observations in Example 1

| $p$  | $\delta_0$ | $\delta_1$ | $\delta_2$ | GLM<br>NET | 1NN    | NN<br>RAND | NNET*  |        |         |          | SVM<br>LIN | SVM<br>RBF |
|------|------------|------------|------------|------------|--------|------------|--------|--------|---------|----------|------------|------------|
|      |            |            |            |            |        |            | 1      | 3      | 5       | 10       |            |            |
| 50   | 0.0149     | 0.0189     | 0.0188     | 0.0940     | 0.0008 | 2.7834     | 0.0090 | 0.0162 | 0.0328  | 0.1110   | 0.0052     | 0.0060     |
| 100  | 0.0156     | 0.0236     | 0.0238     | 0.0978     | 0.0024 | 3.4872     | 0.0130 | 0.0454 | 0.1070  | 0.4012   | 0.0104     | 0.0102     |
| 250  | 0.0185     | 0.0390     | 0.0389     | 0.1050     | 0.0048 | 4.6608     | 0.0382 | 0.2232 | 0.5982  | 4.2194   | 0.0224     | 0.0240     |
| 500  | 0.0209     | 0.0551     | 0.0549     | 0.1132     | 0.0070 | 5.3308     | 0.1104 | 0.8512 | 3.9240  | 19.7896  | 0.0398     | 0.0402     |
| 1000 | 0.0263     | 0.0807     | 0.0808     | 0.1530     | 0.0120 | 6.7963     | 0.3883 | 6.3370 | 19.1236 | 100.7417 | 0.0713     | 0.0797     |

\* 1,3,5,10 represent the numbers of units in the single-hidden-layer of the neural network.

### B.2 Codes

The R-codes for implementation of the proposed classifiers are available [here](#).