

Article

# On Supervised Classification of Feature Vectors with Independent and Non-Identically Distributed Elements

Farzad Shahrivari  and Nikola Zlatanov \*

Electrical and Computer Systems Engineering, Monash University, Alliance Ln, Clayton, VIC 3168, Australia; Farzad.shahrivari@monash.edu

\* Correspondence: Nikola.zlatanov@monash.edu

**Abstract:** In this paper, we investigate the problem of classifying feature vectors with mutually independent but non-identically distributed elements that take values from a finite alphabet set. First, we show the importance of this problem. Next, we propose a classifier and derive an analytical upper bound on its error probability. We show that the error probability moves to zero as the length of the feature vectors grows, even when there is only one training feature vector per label available. Thereby, we show that for this important problem at least one asymptotically optimal classifier exists. Finally, we provide numerical examples where we show that the performance of the proposed classifier outperforms conventional classification algorithms when the number of training data is small and the length of the feature vectors is sufficiently high.

**Keywords:** supervised classification; independent and non-identically distributed features; analytical error probability



**Citation:** Shahrivari, F.; Zlatanov, N. On Supervised Classification of Feature Vectors with Independent and Non-Identically Distributed Elements. *Entropy* **2021**, *23*, 1045. <https://doi.org/10.3390/e23081045>

Academic Editors: Lizhong Zheng and Chao Tian

Received: 26 July 2021

Accepted: 10 August 2021

Published: 13 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

### 1.1. Background

Supervised classification is a machine learning technique that maps an input feature vector to an output label based on a set of correctly labeled training data. There is no single learning algorithm that works best on all supervised learning problems, as shown by the no free lunch theorem in [1]. As a result, there are many algorithms proposed in the literature whose performance depends on the underlying problem and the amount of training data available. The most widely used algorithms in the literature are decision trees [2,3], Support Vector Machines (SVM) [4,5], Rule-Based Systems [6], naive Bayes classifiers [7], k-nearest neighbors (KNN) [8], logistic regressions, and neural networks [9,10].

### 1.2. Motivation

In the following, we discuss the motivation for this work.

#### 1.2.1. Lack of Tight Upper Bounds on the Performance of Classifiers

In general, there are no tight upper bounds on the performance of the classifiers used in practice. Many of the previous works only provide experimental performance results. However, this approach has drawbacks. For example, one has to rely on the trial-and-error approach in order to develop a good classifier for a given problem, which impacts the reliability. Next, the algorithms whose performance has been verified only experimentally may work for a given problem, but may fail to work when applied to a similar problem. Finally, experimental results do not provide intuition into the underlying problem, whereas the analytical results provide the understanding of the underlying problem and the corresponding solutions.

Motivated by this, in the paper, we aim to investigate classifiers with analytical upper bounds on their performance.

### 1.2.2. Independent and Non-Identically Distributed Features

In general, we can categorize the statistical properties of the feature vectors, which are the input to the classifier, into three types. To this end, let  $Y^n(X) = [Y_1(X), Y_2(X), \dots, Y_n(X)]$  denote the input feature vector to the supervised classifier, where  $n$  is the length of the feature vector and  $X$  is the label to which the feature vector  $Y^n(X)$  belongs. Then, we can distinguish the following three types of feature vectors depending on the statistics of the elements in the feature vector  $Y^n(X)$ .

The first type of feature vector is when the elements of  $Y^n(X)$  are independent and identically distributed (i.i.d.). This is the simplest features model, but also the least applicable in practice. This model is identical to hypothesis testing, which has been well investigated in the literature [11–13]. As a result, tight upper bounds on the performance of supervised learning algorithms for this type of feature vector are available in the hypothesis testing literature. For instance, the authors in [11] showed that the posterior entropy and the maximum a posterior error probability decay to zero with the length of the feature vector at the identical exponential rate, where the maximum achievable exponent is the minimum Chernoff information. In [12], the authors determine the requirements for the length of the vector  $Y^n(X)$  and the number of labels  $m$  in order to achieve vanishing exponential error probability in testing  $m$  hypothesis that minimizes the rejection zone. In [13], the authors provide an upper bound and a lower-bound on the error probability of Bayesian  $m$ -ary hypothesis testing in terms of conditional entropy.

The second type of feature vectors is when the elements of  $Y^n(X)$  are mutually dependent and non-identically distributed (d.non-i.d.). This type of features model is the most general model and the most applicable in practice. However, it is also the most difficult to tackle analytically. As a result, supervised learning algorithms proposed for this features model lack analytical tight upper bounds on their performance [14–23]. This is because there are not any frameworks that produce closed-form results when deriving statistics of vectors with d.non-i.d. elements when the underlying distributions are unknown. Then how can we investigate analytically classifiers for practical scenarios when the feature vectors have d.non-i.d. elements? A possible approach leads us to the third type of feature vectors, explained in the following.

The third type of feature vectors is when the elements of  $Y^n(X)$  are mutually independent but non-identically distributed (i.non-i.d.). This features model is much simpler than the d.non-i.d. features model and, more importantly, it is analytically tractable, as we show in this paper. Furthermore, this features model is applicable in practice. Specifically, there exists a class of algorithms, known as Independent Component Analysis (ICA), that transform vectors with d.non-i.d. elements into vectors with i.non-i.d. elements with a zero or a negligible loss of information [24–28]. The origins of ICA can be traced back to Barlow [29], who argued that a good representation of binary data can be achieved by an invertible transformation that transform vectors with d.non-i.d. elements into vectors with i.non-i.d. elements. Finding such a transformation with no prior information about the distribution of the data has been considered an open problem until recently [28]. Specifically, the authors in [28] show that this hard problem can be accurately solved with a branch and bound search tree algorithm, or tightly approximated with a series of linear problems. Thereby, the authors in [28] provide the first efficient set of solutions to Barlow's problem. So far, the complexity of the fastest such algorithm is  $\mathcal{O}(n \times 2^n)$  [28]. Nevertheless, since there exist such invertible transformations (i.e., no loss of information) which can transform vectors with d.non-i.d. elements into vectors with i.non-i.d. elements, we can tackle the features model comprised of d.non-i.d. elements by first transforming it (without loss of information) into the features model comprised of i.non-i.d. elements and then tackling the i.non-i.d. features model.

Motivated by this, in this paper, we investigate supervised classification of feature vectors with i.non-i.d. elements.

### 1.2.3. Small Training Set

The main factor that impacts the accuracy of supervised classification is the amount of training data. In fact, most supervised algorithms are able to learn only if there is a very large set of training data available [30]. The main reason for this is the curse of dimensionality [31,32], which states that “the higher the dimensionality of the feature vectors, the more training data are needed for the supervised classifier” [33]. For example, supervised classification methods such as random forest [34,35] and KNN [36] suffer from the curse of dimensionality. However, having large training data sets is not always possible in practice. As a result, designing a supervised classification algorithm that exhibits good performance even when the training data set is extremely small is important.

Motivated by this, in this paper, we investigate supervised classifiers for the case when  $t$  training feature vectors per label are available, where  $t = 1, 2, \dots$

### 1.3. Contributions

In this paper, we propose an algorithm for supervised classification of feature vectors with i.non-i.d. elements when the number of training feature vectors per label is  $t$ , where  $t = 1, 2, \dots$ . Next, we derive an upper bound on the error probability of the proposed classifier for uniformly distributed labels and prove that the error probability exponentially decays to zero when the length of the feature vector,  $n$ , grows, even when only one training vector per label is available, i.e., when  $t = 1$ . Hence, the proposed classification algorithm provides an asymptotically optimal performance even when the number of training vectors per label is extremely small. We compare the performance of the proposed classifier with the naive Bayes classifier and to the KNN algorithm. Our numerical results show that the proposed classifier significantly outperforms the naive Bayes classifier and the KNN algorithm when the number of training feature vectors per label is small and the length of the feature vectors  $n$  is sufficiently high.

The proposed algorithm is a form of the nearest neighbor classification algorithm, where the nearest neighbor is searched in the domain of empirical distributions. As a result, we refer to the algorithm as *the nearest empirical distribution*. The nearest empirical distribution algorithm is not new and, to the best of our knowledge, it was first proposed in [37] for the case when the elements of  $Y^n(X)$  are i.i.d., i.e., for the equivalent problem of hypothesis testing. However, in this paper, we propose the nearest empirical distribution algorithm for the case when the elements of  $Y^n(X)$  are i.non-i.d., which is much more complex than the problem of hypothesis testing where the elements of  $Y^n(X)$  are i.i.d.

To the best of our knowledge, this is the first paper that investigates the important problem of classifying feature vectors with i.non-i.d. elements and provides an upper bound on its error probability. The novelty of this paper is not with the classifier itself, but rather in showing the importance of the problem of classifying feature vectors with i.non-i.d elements and in showing analytically that at least one classifier with an asymptotically optimal error probability exists when at least one training feature vectors per label is available.

The remainder of this paper is structured as follows. In Section 2, we formulate the considered classification problem. In Section 3, we provide our classifier and derive an upper bound on its error probability. In Section 4, we provide numerical examples of the performance on the proposed classifier. Finally, Section 5 concludes the paper.

## 2. Problem Formulation

The machine learning model is comprised of a label  $X$ , a feature vector  $Y^n(X) = [Y_1(X), Y_2(X), \dots, Y_n(X)]$  of length  $n$  mapped to the label  $X$ , and a learned label  $\hat{X}$ , as shown in Figure 1. In this paper, we adopt the information-theoretic style of notations and thereby random variables are denoted by capital letters and their realizations are denoted with small letters. The feature vector  $Y^n(X)$  is the input to the machine learning algorithm whose aim is to detect the label  $X$  from the observed feature vector  $Y^n(X)$ .

The performance of the machine learning algorithm is measured by the error probability  $\mathbb{P}_e = \Pr\{X \neq \hat{X}\}$ .

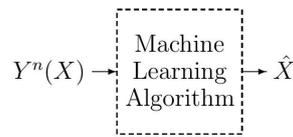


Figure 1. A typical structural modelling of the classification learning problem.

We adopt the modeling in [38–40] and represent the dependency between the label  $X$  and the feature vector  $Y^n(X)$  via a joint probability distribution  $p_{X,Y^n}(x, y^n)$ . Now, in order to gain a better understanding of the problem, we include the joint probability distribution  $p_{X,Y^n}(x, y^n)$  into the model in Figure 1. To this end, since  $p_{X,Y^n}(x, y^n) = p_{Y^n|X}(y^n|x)p_X(x)$  holds, instead of  $p_{X,Y^n}(x, y^n)$ , we can include the conditional probability distribution  $p_{Y^n|X}(y^n|x)$  and the probability distribution  $p_X(x)$  into the model in Figure 1, and thereby obtain the model in Figure 2.

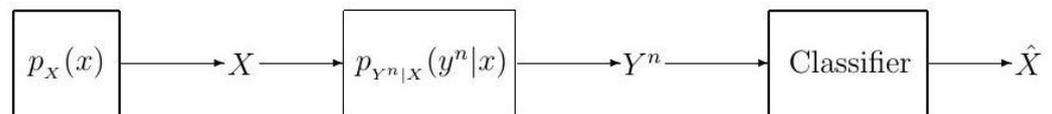


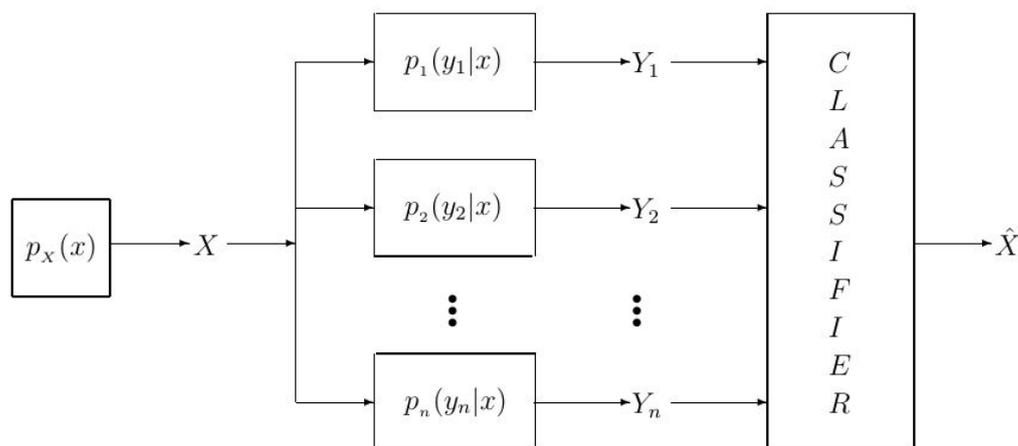
Figure 2. An alternative modeling of the classification learning problem.

Now, the classification learning model in Figure 2 is a system comprised of a label generating source  $X$  according to the distribution  $p_X(x)$ , a feature vector generator modelled by the conditional probability distribution  $p_{Y^n|X}(y^n|x)$ , a feature vector  $Y^n$ , a classifier that aims to detect  $X$  from the observed feature vector  $Y^n$ , and the detected label  $\hat{X}$ . Note that the system model in Figure 2 can be seen equivalently as a communication system comprised of a source  $X$ , a channel with input  $X$  and output  $Y^n$ , and a decoder (i.e., detector) that aims to detect  $X$  from  $Y^n$ . The notation used in this paper, letter  $X$  for labels and letter  $Y$  for features, is based on the notation used in information theory for modelling communication systems. In the classification model shown in Figure 2, we assume that the label  $X$  can take values from the set  $\mathcal{X}$ , according to  $p_X(x) = 1/|\mathcal{X}|$ , where  $|\cdot|$  denotes the cardinality of a set. Next, we assume that the  $i$ -th element of the feature vector  $Y^n$ ,  $Y_i$ , for  $i = 1, 2, \dots, n$ , takes values from the set  $\mathcal{Y} = \{y_1, y_2, \dots, y_{|\mathcal{Y}|}\}$ , according to the conditional probability distribution  $p_{Y_i|X}(y_i|x)$ .

Moreover, we assume that the elements of the feature vector  $Y^n$  are i.i.d. As a result, the feature vector  $Y^n$  takes values from the set  $\mathcal{Y}^n$  according to the conditional probability distribution  $p_{Y^n|X}(y^n|x)$  given by

$$p_{Y^n|X}(y^n|x) = p_{Y_1, Y_2, \dots, Y_n|X}(y_1, y_2, \dots, y_n|x) \stackrel{(a)}{=} \prod_{i=1}^n p_{Y_i|X}(y_i|x) \stackrel{(b)}{=} \prod_{i=1}^n p_i(y_i|x), \quad (1)$$

where (a) comes from the fact that elements in the feature vector  $Y^n$  are mutually independent and (b) is for the sake of notational simplicity, where  $p_i$  is used instead of  $p_{Y_i|X}$ . As a result of (1), the considered classification model in Figure 2 can be represented equivalently as in Figure 3.



**Figure 3.** An alternative modelling of the classification learning problem when the elements of  $Y^n(X)$  are mutually independent but non-identically distributed (i.non-i.d.).

Next, we assume that  $p_i(y_i|x), \forall i$ , and thereby  $p_{Y^n|X}(y^n|x)$  are unknown to the classifier. Instead, the classifier knows  $\mathcal{X}, \mathcal{Y}$ , and for each  $x_i \in \mathcal{X}$ , where  $i = 1, 2, \dots, |\mathcal{X}|$ , it has access to a finite set of  $t$  correctly labelled input–output pairs  $(x_i, \hat{y}_{i_1}^n), (x_i, \hat{y}_{i_2}^n), \dots, (x_i, \hat{y}_{i_t}^n)$ , denoted by  $\mathcal{T}_i$ , referred to as the training set for label  $x_i$ .

Finally, we assume that the following holds

$$\sum_{l=1}^n p_l(y|x_i) \neq \sum_{l=1}^n p_l(y|x_j), \text{ for } y \in \mathcal{Y} \text{ and } i \neq j. \tag{2}$$

The condition in (2) means that the distribution of the feature vectors  $Y^n(X)$  for label  $X = i$  is not a perturbation of distribution of the feature vectors  $Y^n(X)$  for label  $X = j$ . As a result, the proposed classifier only applies to the subset of data vectors with i.non-i.d. elements that satisfy (2).

For the classification system model defined above and illustrated in Figure 3, we wish to propose a classifier that exhibits an asymptotically optimal error probability  $\mathbb{P}_e = \Pr\{X \neq \hat{X}\}$  with respect to the length of  $Y^n, n$ , for any  $t \geq 1$ , i.e., for any  $t \geq 1, \mathbb{P}_e \rightarrow 0$  as  $n \rightarrow \infty$ . Moreover, we wish to obtain an analytical upper bound on the error probability of the proposed classifier for a given  $t$  and  $n$ .

### 3. The Proposed Classifier and Its Performance

In this section, we propose our classifier, derive an analytical upper bound on its error probability, and prove that the classifier exhibits an asymptotically optimal performance when the length of the feature vector  $Y^n, n$ , satisfies  $n \rightarrow \infty$ . This is conducted in the following.

For a given vector  $\mathbf{v}^n = (v_1, v_2, \dots, v_n)$ , let the Minkowski distance  $r$  be defined as

$$\|\mathbf{v}\|_r = \left( \sum_{i=1}^n v_i^r \right)^{(1/r)}. \tag{3}$$

Moreover, for a given feature vector  $y^k = (y_1, y_2, \dots, y_k)$ , let  $\mathcal{I}[y^k = y]$  be a function defined as

$$\mathcal{I}[y^k = y] = \sum_{i=1}^k \mathcal{Z}[y_i = y], \tag{4}$$

where  $\mathcal{Z}[y_i = y]$  is an indicator function assuming the value 1 if  $y_i = y$  and 0 otherwise. Hence,  $\mathcal{I}[y^k = y]$  counts the number of elements in  $Y^k$  that have the value  $y$ .

### 3.1. The Proposed Classifier

Let  $\hat{y}_i^{nt}$  be a vector obtained by concatenating all training feature vectors for the input label  $x_i$  as

$$\hat{y}_i^{nt} = (\hat{y}_{i_1}^n, \hat{y}_{i_2}^n, \dots, \hat{y}_{i_t}^n). \tag{5}$$

Let  $P_{\hat{y}_i^{nt}}$  be the empirical probability distribution of the concatenated training feature vector for label  $x_i$ ,  $\hat{y}_i^{nt}$ , given by

$$P_{\hat{y}_i^{nt}} = \left[ \frac{\mathcal{I}[\hat{y}_i^{nt} = y_1]}{nt}, \frac{\mathcal{I}[\hat{y}_i^{nt} = y_2]}{nt}, \dots, \frac{\mathcal{I}[\hat{y}_i^{nt} = y_{|\mathcal{Y}|}]}{nt} \right]. \tag{6}$$

Let  $y^n$  be the observed feature vector at the classifier whose label it wants to detect and let  $P_{y^n}$  denote the empirical probability distribution of  $y^n$ , given by

$$P_{y^n} = \left[ \frac{\mathcal{I}[y^n = y_1]}{n}, \frac{\mathcal{I}[y^n = y_2]}{n}, \dots, \frac{\mathcal{I}[y^n = y_{|\mathcal{Y}|}]}{n} \right]. \tag{7}$$

Using the above notations, we propose the following classifier.

**Proposition 1.** For the considered system model, we propose a classifier with the following classification rule

$$\hat{x} = x_i, \text{ where } i = \arg \min_i \|P_{y^n} - P_{\hat{y}_i^{nt}}\|_r, \tag{8}$$

where  $r \geq 1$  and ties are resolved by assigning the label among the ties uniformly at random. (For example, if  $\|P_{y^n} - P_{\hat{y}_i^{nt}}\|_r = \|P_{y^n} - P_{\hat{y}_j^{nt}}\|_r$  holds for,  $i \neq j$ , we set  $\hat{x} = x_i$  or  $\hat{x} = x_j$  uniformly at random).

As seen from (8), the proposed classifier assigns the label  $x_i$  if the empirical probability distribution of the concatenated training feature vector mapped to label  $x_i$ ,  $P_{\hat{y}_i^{nt}}$  is the closest, in terms of Minkowski distance  $r$ , to the empirical probability distribution of the observed feature vector  $P_{y^n}$ . In that sense, the proposed classifier can be considered as the nearest empirical distribution classifier.

### 3.2. Upper Bound on the Error Probability

The following theorem establishes an upper bound on the error probability of the proposed classifier.

**Theorem 1.** Let  $\bar{P}_j$ , for  $j = 1, 2, \dots, |\mathcal{X}|$ , be a vector defined as

$$\bar{P}_j = [\bar{p}(y_1|x_j), \bar{p}(y_2|x_j), \dots, \bar{p}(y_{|\mathcal{Y}|}|x_j)], \tag{9}$$

where  $\bar{p}(y|x_j)$  is given by

$$\bar{p}(y|x_j) = \frac{1}{n} \sum_{k=1}^n p_k(y|x_j). \tag{10}$$

Then, for a given  $r \geq 1$ , the error probability of the proposed classifier is upper bounded by

$$\mathbb{P}_e \leq 2|\mathcal{Y}|e^{-2n\epsilon^2} + 2|\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2}, \tag{11}$$

where  $\epsilon$  is given by

$$\epsilon = \min_{\substack{i,j \\ i \neq j}} \frac{\|P_{\hat{y}_i^{nt}} - \bar{P}_j\|_r}{(2 + t^{-1/3})|\mathcal{Y}|^{1/r}}. \tag{12}$$

**Proof of Theorem 1.** Without loss of generality we assume that  $x_1$  is the input to  $p_{Y^n|X}(y^n|x)$  and  $y^n$  is observed.

Let  $\mathcal{A}_k^\epsilon$ , for  $1 \leq k \leq |\mathcal{Y}|$ , be a set defined as

$$\mathcal{A}_k^\epsilon = \left\{ y^n : \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1) \right| \leq \epsilon \right\}. \tag{13}$$

Furthermore, let  $\mathcal{B}_k^\epsilon$ , for  $1 \leq k \leq |\mathcal{Y}|$ , be a set defined as

$$\mathcal{B}_k^\epsilon = \left\{ \hat{y}^{nt} : \left| \frac{\mathcal{I}[\hat{y}^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right| \leq \frac{\epsilon}{\sqrt[3]{t}} \right\}. \tag{14}$$

Let  $\mathcal{A}^\epsilon = \bigcap_{k=1}^{|\mathcal{Y}|} \mathcal{A}_k^\epsilon$  and  $\mathcal{B}^\epsilon = \bigcap_{k=1}^{|\mathcal{Y}|} \mathcal{B}_k^\epsilon$ . Now, for any  $y^n \in \mathcal{A}^\epsilon$ , we have

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} \stackrel{(a)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \epsilon^r \right)^{1/r}, \tag{15}$$

where (a) follows from (13). Moreover, for  $\hat{y}_1^{nt} \in \mathcal{B}^\epsilon$ , we have

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} \stackrel{(a)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left( \frac{\epsilon}{\sqrt[3]{t}} \right)^r \right)^{1/r}, \tag{16}$$

where (a) follows from (14). Next, we have the following upper bound

$$\begin{aligned} & \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \\ &= \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1) - \left( \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right) \right|^r \right)^{1/r} \\ &\stackrel{(a)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} + \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r}, \end{aligned} \tag{17}$$

where (a) follows from the Minkowski inequality. Combining (15)–(17), we obtain

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \leq |\mathcal{Y}|^{1/r} \epsilon + |\mathcal{Y}|^{1/r} \frac{\epsilon}{\sqrt[3]{t}}. \tag{18}$$

Hence, the Minkowski distance between the empirical probability distribution of the observed vector  $y^n$  and the empirical probability distribution of the concatenated training

vector for label  $x_1$  is upper bounded by the right hand side of (18). We now derive a lower bound for  $\hat{y}_i^{nt}$ , where  $i \neq 1$ . For any  $x_i$ , such that  $i \neq 1$ , we have

$$\begin{aligned} & \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} + \left( \sum_{k=1}^{|\mathcal{Y}|} \epsilon^r \right)^{1/r} \\ & \stackrel{(a)}{\geq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} + \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} \\ & \stackrel{(b)}{\geq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r}, \end{aligned} \tag{19}$$

where (a) follows from (15) and (b) is again due to the Minkowski inequality. The expression in (19), can be written equivalently as

$$\begin{aligned} & \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \\ & \geq \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} - |\mathcal{Y}|^{1/r} \epsilon, \end{aligned} \tag{20}$$

where  $i \neq 1$ . Now, using the definitions of  $P_{\hat{y}_i^{nt}}$  and  $\bar{P}_1$  given by (6) and (9), respectively, into (20) we can replace the expression in the right-hand side of (20) by  $\|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r$ , and thereby for any  $i \neq 1$  we have

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \geq \|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon. \tag{21}$$

The expression in (21) represents a lower bound on the Minkowski  $r$  distance between the empirical probability distribution of the observed vector  $y^n$  and the empirical probability distribution of the concatenated training vector for any label  $x_i$ , where  $i \neq 1$ .

Using the bounds in (18) and (21), we now relate the left-hand sides of (18) and (21). As long as the following inequality holds for each  $i \neq 1$ ,

$$|\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right) < \|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon, \tag{22}$$

which is equivalent to the following for  $i \neq 1$

$$\epsilon < \frac{\|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r}{(2 + t^{-1/3})|\mathcal{Y}|^{1/r}}, \tag{23}$$

$$\begin{aligned} & \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \stackrel{(a)}{\leq} |\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right) \\ & \stackrel{(b)}{<} \|P_{\hat{y}_1^{nt}} - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon \\ & \stackrel{(c)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r}, \end{aligned} \tag{24}$$

where (a), (b), and (c) follow from (18), (22), and (21), respectively. Thereby, from (24), we have the following for  $i \neq 1$

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nT} \right|^r \right)^{1/r} < \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nT} \right|^r \right)^{1/r}. \tag{25}$$

Note that the right- and left-hand sides of (25) can be replaced by the Minkowski distance of the vectors

$$v_1 = \left[ \frac{\mathcal{I}[y^n = y_1]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_1]}{nt}, \dots, \frac{\mathcal{I}[y^n = y_{|\mathcal{Y}|}]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_{|\mathcal{Y}|}]}{nt} \right], \tag{26}$$

and

$$v_2 = \left[ \frac{\mathcal{I}[y^n = y_1]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_1]}{nt}, \dots, \frac{\mathcal{I}[y^n = y_{|\mathcal{Y}|}]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_{|\mathcal{Y}|}]}{nt} \right], \tag{27}$$

respectively. Now, (26) and (27) can be replaced by  $P_{y^n} - P_{\hat{y}_1^{nt}}$  and  $P_{y^n} - P_{\hat{y}_i^{nt}}$ , respectively, by the definitions of  $P_{y^n}$  and  $P_{\hat{y}_i^{nt}}$  given by (7) and (6), respectively. Therefore, (25) can be written equivalently as

$$\|P_{y^n} - P_{\hat{y}_1^{nt}}\|_r < \|P_{y^n} - P_{\hat{y}_i^{nt}}\|_r. \tag{28}$$

Now, let us highlight what we have obtained. We obtained that there is an  $\epsilon$  for which if (23) holds for  $i \neq 1$ , and for that  $\epsilon$  there are sets  $\mathcal{A}^\epsilon$  and  $\mathcal{B}^\epsilon$  for which  $y^n \in \mathcal{A}^\epsilon$  and  $\hat{y}_1^{nt} \in \mathcal{B}^\epsilon$  then (28) holds for  $i \neq 1$ , and thereby our classifier will detect that  $x_1$  is the correct label. Using this, we can upper bound the error probability as

$$\begin{aligned} \mathbb{P}_e &= 1 - \Pr\{\hat{x}_1 = x_1\} \\ &\leq 1 - \Pr\{(y^n \in \mathcal{A}^\epsilon) \cap (\hat{y}_1^{nt} \in \mathcal{B}^\epsilon) | \epsilon \in \mathcal{S}\}, \end{aligned} \tag{29}$$

where  $\mathcal{S}$  is a set defined as

$$\mathcal{S} = \left\{ \epsilon : \epsilon \leq \min_{i \neq 1} \frac{\|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r}{(2 + t^{-1/3})|\mathcal{Y}|^{1/r}} \right\}. \tag{30}$$

In the following, we derive the expression in (29). The right-hand side of (29) can be upper bounded as

$$\begin{aligned} 1 - \Pr\{(y^n \in \mathcal{A}^\epsilon) \cap (\hat{y}_1^{nt} \in \mathcal{B}^\epsilon) | \epsilon \in \mathcal{S}\} &= \Pr\{(y^n \notin \mathcal{A}^\epsilon) \cup (\hat{y}_1^{nt} \notin \mathcal{B}^\epsilon) | \epsilon \in \mathcal{S}\} \\ &\stackrel{(a)}{\leq} \Pr\{y^n \notin \mathcal{A}^\epsilon | \epsilon \in \mathcal{S}\} + \Pr\{\hat{y}_1^{nt} \notin \mathcal{B}^\epsilon | \epsilon \in \mathcal{S}\}, \end{aligned} \tag{31}$$

where (a) follows from Boole’s inequality. Now, note that we have the following upper bound for the first expression in the right-hand side of (31)

$$\begin{aligned}
 \Pr\{y^n \notin \mathcal{A}^\epsilon \mid \epsilon \in \mathcal{S}\} &= \Pr\left\{y^n \notin \bigcap_{k=1}^{|\mathcal{Y}|} \mathcal{A}_k^\epsilon \mid \epsilon \in \mathcal{S}\right\} \\
 &= \Pr\left\{y^n \in \bigcup_{k=1}^{|\mathcal{Y}|} \overline{\mathcal{A}_k^\epsilon} \mid \epsilon \in \mathcal{S}\right\} \\
 &\stackrel{(a)}{\leq} \sum_{k=1}^{|\mathcal{Y}|} \Pr\{y^n \in \overline{\mathcal{A}_k^\epsilon} \mid \epsilon \in \mathcal{S}\} \\
 &= \sum_{k=1}^{|\mathcal{Y}|} \Pr\left\{\left|\frac{\mathcal{I}[y^n = y_k]}{n} - \bar{p}(y_k|x_1)\right| > \epsilon \mid \epsilon \in \mathcal{S}\right\} \\
 &= \sum_{k=1}^{|\mathcal{Y}|} \Pr\left\{\left|\sum_{j=1}^n \frac{\mathcal{Z}[y_j = y_k]}{n} - \bar{p}(y_k|x_1)\right| > \epsilon \mid \epsilon \in \mathcal{S}\right\}, \tag{32}
 \end{aligned}$$

where  $\overline{\mathcal{A}_k^\epsilon}$  is the complement of  $\mathcal{A}_k^\epsilon$  and (a) follows from Boole’s inequality. Note that  $\mathcal{Z}[y_1 = y_k], \mathcal{Z}[y_2 = y_k], \dots, \mathcal{Z}[y_n = y_k]$  in (32) are  $n$  independent Bernoulli random variables with probabilities of success  $p_1(y_k|x_1), p_2(y_k|x_1), \dots, p_n(y_k|x_1)$ , respectively. Let  $\mathcal{W}[y_k]$  be a binomial random variable with parameters  $(n, \bar{p}(y_k|x_1))$ . We proceed the proof by introducing the following well-known Hoeffding’s Theorem from [41].  $\square$

**Theorem 2** (Hoeffding [41]). *Assume that  $Z_1, Z_2, \dots,$  and  $Z_n$  are  $n$  independent Bernoulli random variables with probabilities of success  $p_1, p_2, \dots,$  and  $p_n$ , respectively. Next, let  $Z$  be defined as  $Z = Z_1 + Z_2 + \dots + Z_n$  and, let  $\bar{p}$  be defined as  $\bar{p} = (p_1 + p_2 + \dots + p_n)/n$ . Let  $W$  be a binomial random variable with parameters  $(n, \bar{p})$ . Then, for a given  $a$  and  $b$ , where  $0 \leq a \leq n\bar{p} \leq b \leq n$  holds, we have*

$$\Pr\{a \leq W \leq b\} \leq \Pr\{a \leq Z \leq b\}. \tag{33}$$

In other words, the probability distribution of  $W$  is more dispersed around its mean  $n\bar{p}$  than is the probability distribution of  $Z$ . Except in the trivial case when  $a = b = 0$ , the bound in (33) holds with equality if and only if  $p_1 = \dots = p_n = \bar{p}$ .

**Proof of Theorem 2.** Please refer to [41].  $\square$

Setting  $a = n(\bar{p} - \delta)$  and  $b = n(\bar{p} + \delta)$  in (33), we obtain

$$\Pr\{n(\bar{p} - \delta) \leq W \leq n(\bar{p} + \delta)\} \leq \Pr\{n(\bar{p} - \delta) \leq Z \leq n(\bar{p} + \delta)\}. \tag{34}$$

Using (34), we have the following upper bound

$$\begin{aligned}
 \Pr\left\{\left|\frac{Z}{n} - \bar{p}\right| > \delta\right\} &= 1 - \Pr\{n(\bar{p} - \delta) \leq Z \leq n(\bar{p} + \delta)\} \\
 &\stackrel{(a)}{\leq} 1 - \Pr\{n(\bar{p} - \delta) \leq W \leq n(\bar{p} + \delta)\} \\
 &= \Pr\left\{\left|\frac{W}{n} - \bar{p}\right| > \delta\right\}, \tag{35}
 \end{aligned}$$

where (a) follows from (34).

We now turn to the proof of Theorem 1. According to Theorem 2, the probability distribution of  $\mathcal{W}[y_k]$  is more dispersed around its mean  $n\bar{p}(y_k|x_1)$  than is the probability

distribution of  $\sum_{1 \leq j \leq n} \mathcal{Z}[y_j = y_k]$ . Therefore, we can upper bound the probability in the last line of (32) as

$$\Pr \left\{ \left| \sum_{j=1}^n \frac{\mathcal{Z}[y_j = y_k]}{n} - \bar{p}(y_k|x_1) \right| > \epsilon \mid \epsilon \in \mathcal{S} \right\} \stackrel{(a)}{\leq} \Pr \left\{ \left| \frac{\mathcal{W}[y_k]}{n} - \bar{p}(y_k|x_1) \right| > \epsilon \mid \epsilon \in \mathcal{S} \right\}, \tag{36}$$

where  $\epsilon \in \mathcal{S}$  is defined in (30) and (a) follows from (35). Now, let us introduce another well-known Hoeffding’s Theorem from [42].

**Theorem 3** (Hoeffding’s inequality [42]). *Let  $W_1, W_2, \dots, W_n$  be  $n$  independent random variables such that for each  $1 \leq i \leq n$ , we have  $\Pr\{W_i \in [a_i, b_i]\} = 1$ . Then for  $S_n$ , defined as  $S_n = \sum_{i=1}^n W_i$ , we have*

$$\Pr \{ S_n - \mathbb{E}[S_n] \geq \delta \} \leq \exp \left( - \frac{2\delta^2}{\sum_{i=1}^n (b_i - a_i)^2} \right), \tag{37}$$

where  $\mathbb{E}[S_n]$  is the expectation of  $S_n$ .

**Proof of Theorem 3.** Please refer to [42].  $\square$

Back to (36), by using the result of (37) for  $a_i = 0$  and  $b_i = 1$  since the binomial random variable  $\mathcal{W}[y_k]$  can take values 0 or 1, respectively, we have

$$\Pr \left\{ \left| \sum_{j=1}^n \frac{\mathcal{Z}[y_j = y_k]}{n} - \bar{p}(y_k|x_1) \right| > \epsilon \mid \epsilon \in \mathcal{S} \right\} \leq 2 \exp \left( - \frac{2n^2\epsilon^2}{\sum_{1 \leq i \leq n} (1 - 0)^2} \right) \leq 2e^{-2n\epsilon^2}, \tag{38}$$

where  $\epsilon \in \mathcal{S}$  is defined in (30). Inserting (38) into (32), we obtain the following upper bound

$$\Pr \{ y^n \notin \mathcal{A}^\epsilon \mid \epsilon \in \mathcal{S} \} \leq 2|\mathcal{Y}|e^{-2n\epsilon^2}. \tag{39}$$

Similarly, we have the following result for the second expression in the right-hand side of (31)

$$\begin{aligned} \Pr \{ \hat{y}_1^{nt} \notin \mathcal{B}^\epsilon \mid \epsilon \in \mathcal{S} \} &= \Pr \left\{ \hat{y}_1^{nt} \notin \bigcap_{k=1}^{|\mathcal{Y}|} \mathcal{B}_k^\epsilon \mid \epsilon \in \mathcal{S} \right\} \\ &= \Pr \left\{ \hat{y}_1^{nt} \in \bigcup_{k=1}^{|\mathcal{Y}|} \overline{\mathcal{B}_k^\epsilon} \mid \epsilon \in \mathcal{S} \right\} \\ &\stackrel{(a)}{\leq} \sum_{k=1}^{|\mathcal{Y}|} \Pr \{ \hat{y}_1^{nt} \in \overline{\mathcal{B}_k^\epsilon} \mid \epsilon \in \mathcal{S} \} \\ &= \sum_{k=1}^{|\mathcal{Y}|} \Pr \left\{ \left| \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right| > \frac{\epsilon}{\sqrt[3]{t}} \mid \epsilon \in \mathcal{S} \right\} \\ &= \sum_{k=1}^{|\mathcal{Y}|} \Pr \left\{ \left| \sum_{j=1}^{nt} \frac{\mathcal{Z}[y_j = y_k]}{nt} - \bar{p}(y_k|x_1) \right| > \frac{\epsilon}{\sqrt[3]{t}} \mid \epsilon \in \mathcal{S} \right\}, \tag{40} \end{aligned}$$

where again (a) follows from Boole’s inequality. Note that due to (5), for any integer number  $l$  such that  $0 \leq l \leq t - 1$  the random variables  $\mathcal{Z}[y_{nl+1} = y_k], \mathcal{Z}[y_{nl+2} = y_k], \dots$ , and  $\mathcal{Z}[y_{nl+n} = y_k]$  in (40) are  $n$  independent Bernoulli random variables with the probabilities of success  $p_1(y_k|x_1), p_2(y_k|x_1), \dots$ , and  $p_n(y_k|x_1)$ , respectively ( $y_{nl+1}, y_{nl+2}, \dots, y_{nl+n}$  are elements of  $\hat{y}_{l+1}^n$ ). In addition, note that

$$\begin{aligned} \bar{p}(y_k|x_1) &= \frac{1}{n} \sum_{j=1}^n p_j(y_k|x_1) \\ &= \frac{1}{nt} \left( \sum_{l=0}^{t-1} \sum_{j=1}^n p_j(y_k|x_1) \right). \end{aligned} \tag{41}$$

Notice that for each  $0 \leq l \leq t - 1$ ,  $p_1(y_k|x_1) + p_2(y_k|x_1) + \dots + p_n(y_k|x_1)$  is the summation of the probabilities of success of the random variables  $\mathcal{Z}[y_{nl+1} = y_k]$ ,  $\mathcal{Z}[y_{nl+2} = y_k]$ ,  $\dots$ , and  $\mathcal{Z}[y_{nl+n} = y_k]$ . Thereby, the last expression on the right-hand side of (41) is the average probability of success of random variables  $\mathcal{Z}[y_j = y_k]$  for  $1 \leq j \leq nt$ . Now, let  $\mathcal{W}[y_k]$  be a binomial random variable with parameters  $(nt, \bar{p}(y_k|x_1))$ . Once again, according to Theorem 2, the probability distribution of  $\mathcal{W}[y_k]$  is more dispersed around its mean  $nt\bar{p}(y_k|x_1)$  than is the probability distribution of  $\sum_{1 \leq j \leq nt} \mathcal{Z}[y_j = y_k]$ . Therefore, the probability in the last line of (40) can be upper bounded as

$$\begin{aligned} \Pr \left\{ \left| \sum_{j=1}^{nt} \frac{\mathcal{Z}[y_j = y_k]}{nt} - \bar{p}(y_k|x_1) \right| > \frac{\epsilon}{\sqrt[3]{t}} \mid \epsilon \in \mathcal{S} \right\} &\stackrel{(a)}{\leq} \Pr \left\{ \left| \frac{\mathcal{W}[y_k]}{nt} - \bar{p}(y_k|x_1) \right| > \frac{\epsilon}{\sqrt[3]{t}} \mid \epsilon \in \mathcal{S} \right\} \\ &\stackrel{(b)}{\leq} 2 \exp \left( - \frac{2(nt)^2(t^{-1/3}\epsilon)^2}{\sum_{1 \leq i \leq nt} (1-0)^2} \right) \\ &\leq 2e^{-2nt(t^{-2/3}\epsilon^2)} \\ &= 2e^{-2nt^{1/3}\epsilon^2}, \end{aligned} \tag{42}$$

where  $\epsilon \in \mathcal{S}$ , defined in (30), (a) follows from (35) (in which  $n$  is replaced by  $nt$ ), and (b) is the result of (37) for  $a_i = 0$  and  $b_i = 1$  since the binomial random variable  $\mathcal{W}[y_k]$  can take values 0 or 1, respectively. Inserting (42) into (40), we have the following upper bound

$$\Pr \left\{ \hat{y}_1^{nt} \notin \mathcal{B}^\epsilon \mid \epsilon \in \mathcal{S} \right\} \leq 2|\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2}. \tag{43}$$

Inserting (39) and (43) into (31), and then inserting (31) into (29), we obtain the following upper bound for the error probability

$$\mathbb{P}_e \leq 2|\mathcal{Y}|e^{-2n\epsilon^2} + 2|\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2}, \tag{44}$$

where

$$\epsilon = \min_{\substack{i,j \\ i \neq j}} \frac{\|P_{\hat{y}_i^{nt}} - \bar{P}_j\|_r}{(2 + t^{-1/3})|\mathcal{Y}|^{1/r}}, \tag{45}$$

which is the optimal value of  $\epsilon$  that exhibits the tightest upper bound for the error probability  $\mathbb{P}_e$  given by (44). This completes the proof of Theorem 1.

The following corollary provides a simplified upper bound on the error probability when  $t \rightarrow \infty$ .

**Corollary 1.** *When the number of training vectors per label reaches infinity, i.e., when  $t \rightarrow \infty$ , which is equivalently to the case when the probability distribution  $p(y^n|x)$  is known at the classifier, the error probability of the proposed classifier is upper bounded as*

$$\mathbb{P}_e \leq 2|\mathcal{Y}|e^{-2n\epsilon^2}, \tag{46}$$

where  $\epsilon$  is given by

$$\epsilon = \min_{\substack{i,j \\ i \neq j}} \frac{\|\bar{P}_i - \bar{P}_j\|_r}{2|\mathcal{Y}|^{1/r}}. \quad (47)$$

**Proof.** The proof is straightforward.  $\square$

As can be seen from (8) and (11), the performance of the proposed classifier depends on  $r$ . We cannot derive the optimal value of  $r$  that minimizes the error probability since we do not have the exact expression of the error probability, we only have its upper bound. On the other hand, in practice, the optimal  $r$  with respect to the upper bound on the error probability also cannot be derived since the upper bound depends on  $\bar{P}_j$ , which would be unknown in practice due to  $p_{Y^n|X}(y^n|x)$  being unknown. As a result, for our numerical examples, we consider the Euclidean distance ( $r = 2$ ), which is one of the most widely used distance metrics in practice.

The following corollary establishes the asymptotic optimality of the proposed classifier with respect to  $n$ .

**Corollary 2.** *The proposed classifier has an error probability that satisfies  $\mathbb{P}_e \rightarrow 0$  as  $n \rightarrow \infty$  if  $|\mathcal{Y}| \leq \mathcal{O}(n^m)$ ,  $m$  is fixed, and  $r > 2m$ . Here,  $n^m$  indicates the dimension of our space, i.e., maximum number of alphabets each element in the feature vector  $y^n$  can take. Thereby, the proposed classifier is asymptotically optimal.*

**Proof.** For the proof, please see Appendix A.  $\square$

#### 4. Simulation Results

In this section, we provide simulation results of the performance of the proposed classifier for  $r = 2$  and compare it to benchmark schemes. The benchmark schemes that we adopt for comparison are the naive Bayes classifier and the KNN algorithm. We cannot adopt a classifier based on a neural network since neural networks require a very large training set, which we assume is not available. For the naive Bayes classifier, the probability distribution  $p_{Y^n|X}(y^n|x)$  is estimated from the training vectors as follows. Let again  $\hat{y}_i^{nt}$  be a vector obtained by concatenating all training feature vectors for the input label  $x_i$  as in (5). Then, the estimated probability distribution of  $p(y_j = y|x_i)$ , denoted by  $\hat{p}(y_j = y|x_i)$ , is found as

$$\hat{p}(y_j = y|x_i) = \frac{\mathcal{I}[\hat{y}_i^{nt} = y]}{nt}, \quad (48)$$

and the naive Bayes classifier decides according to

$$\hat{x} = \arg \max_{x_i} \prod_{k=1}^n \hat{p}(y_k|x_i). \quad (49)$$

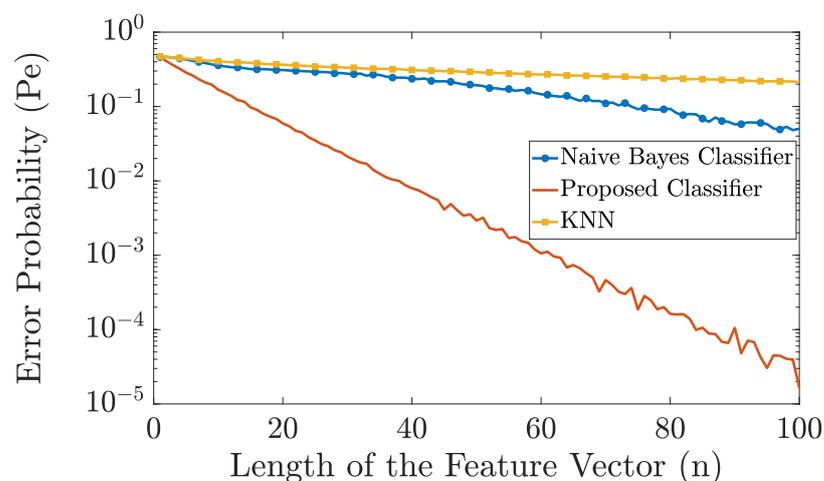
The main problem of the naive Bayes classifier occurs when an alphabet  $y_j \in \mathcal{Y}$  is not present in the training feature vectors. In that case,  $\hat{p}(y_j|x_i)$  in (48) is  $\hat{p}(y_j|x_i) = 0$ ,  $\forall x_i \in \mathcal{X}$  and, as a result, the right hand side of (49) is zero since at least one of the elements in the product in (49) is zero. In this case, the naive Bayes classifier fails to provide an accurate classification of the labels. In what follows, we see that this issue of the naive Bayes classifier appears frequently when we have a small number of training feature vectors. On the other hand, the KNN classifier works as follows. For the observed feature vector  $y^n$ , the KNN classifier looks for the  $k$  nearest feature vectors to  $y^n$ , among all training feature vectors  $\hat{y}_{r_s}^n$ , for all  $1 \leq r \leq |\mathcal{X}|$  and  $1 \leq s \leq T$ . Then by considering a set of  $K$  input–output pairs  $(x_k, \hat{y}_{k_l}^n)$ , for  $k \in \{1, 2, \dots, |\mathcal{X}|\}$  and  $l \in \{1, 2, \dots, |T|\}$ , the KNN classifier decides a label which is the most frequent among  $x_k$ -s. The optimum value of  $k$  for  $t = 1$  is  $k = 1$ .

In the following, we provide numerical examples where we illustrate the performance of the proposed classifier when  $p_{Y^n|X}(y^n|x)$  is artificially generated.

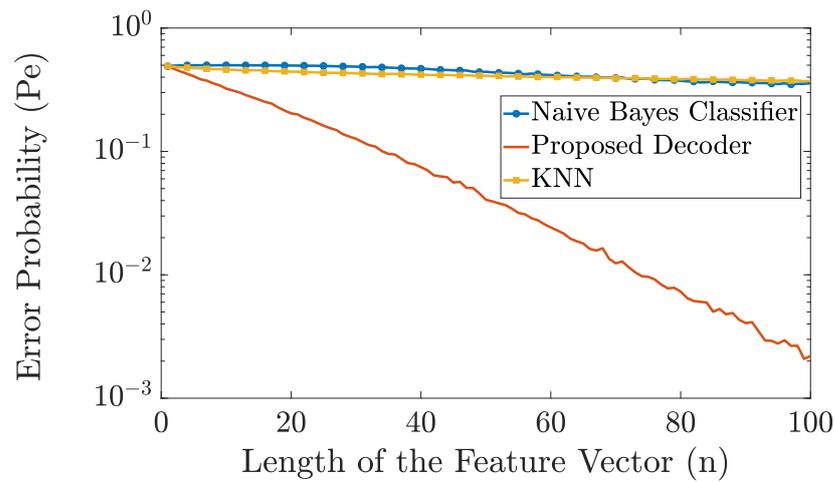
#### 4.1. The I.I.D. Case with One Training Sample per Label

In the following examples, we assume that the classifiers have access to only one training feature vector for each label, the elements of the feature vectors are generated i.i.d., and the alphabet size of the feature vector,  $|\mathcal{Y}|$ , is fixed.

In Figures 4 and 5, we compare the error probability of the proposed classifier with the naive Bayes classifier and the KNN algorithm for the case when  $|\mathcal{Y}| = 6$  and  $|\mathcal{Y}| = 20$ , respectively. In both examples, we have two different labels, i.e.,  $|\mathcal{X}| = 2$ . As a result, we have two different probability distributions  $p_{Y^n|X_1}(y^n|x_1)$  and  $p_{Y^n|X_2}(y^n|x_2)$ . The probability distributions  $p_{Y^n|X_1}(y^n|x_1)$  and  $p_{Y^n|X_2}(y^n|x_2)$  are randomly generated as follows. We first generate two random vectors of length 6 and length 20 for Figures 4 and 5, respectively, where the elements of these vectors are drawn independently from a uniform probability distribution. Then we normalize these vectors such that the sum of their elements is equal to one. These two normalized randomly generated vectors then represent the two probability distributions  $p_{Y_i|X_1}(y_i|x_1) = p_{Y_i|X_1}(y_i|x_1)$  and  $p_{Y_i|X_2}(y_i|x_2) = p_{Y_i|X_2}(y_i|x_2)$ ,  $\forall i$ . Then,  $p_{Y^n|X_k}(y^n|x_k)$  is obtained as  $p_{Y^n|X_k}(y^n|x_k) = \prod_{i=1}^n p_{Y_i|X_k}(y_i|x_k)$ , for  $k = 1, 2$ . The simulation is carried out as follows. For each  $n$ , we generate one training vector for each label, using the aforementioned probability distributions. Then, as test samples, we generate 1000 feature vectors for each label and pass these feature vectors through our proposed classifier, the naive Bayes classifier, and the KNN algorithm, and compute the errors. The length of the feature vector  $n$  is varied from  $n = 1$  to  $n = 100$ . We repeat the simulation 5000 times and then plot the error probability. Figures 4 and 5 show that the proposed classifier outperforms both the naive Bayes classification and KNN. The main reason for this performance gain is because when only one training vector per label is available, the proposed classifier is more resilient to errors than the naive Bayes classifier, whereas the KNN algorithm has very poor performance because of the “curse of dimensionality”. Specifically, the naive Bayes classifier cannot perform an accurate classification for small  $n$  compared to  $|\mathcal{Y}|$  since the chance that an alphabet will not be present in one of the training feature vectors is close to 1. On the other hand, the KNN algorithm cannot perform an accurate classification for large  $n$  since the dimension of the input feature vector becomes much larger than the training data and the “curse of dimensionality” occurs.

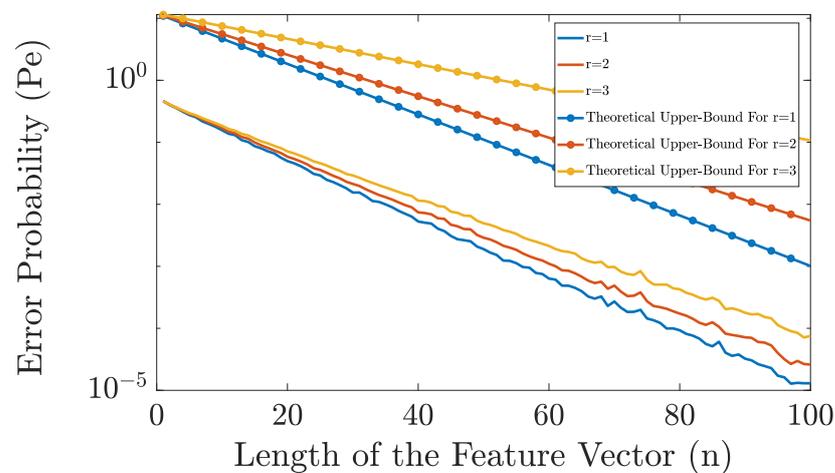


**Figure 4.** Comparison in error probability between the naive Bayes classifier, KNN, and the proposed classifier.



**Figure 5.** Comparison in error probability between the naive Bayes classifier, KNN, and the proposed classifier.

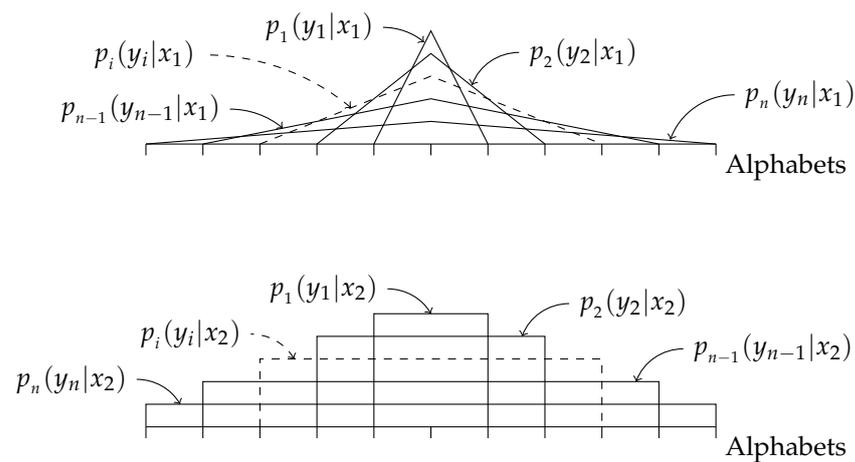
In Figure 6, we compare the performance of the proposed classifier for different values of  $r$  when  $|\mathcal{Y}| = 6$  with the derived upper bounds. As can be seen, for this example, the derived theoretical upper bounds have similar slope as the exact error probabilities. Moreover, we can see that for this example, the optimal  $r$  is  $r = 1$ . However, this is not always the case and it depends on  $p_{Y^n|X_k}(y^n|x_k)$ ,  $|\mathcal{Y}|$ , and  $|\mathcal{X}|$ .



**Figure 6.** Comparison in error probability of the proposed classifier for different values of  $r$  when  $|\mathcal{Y}| = 6$ . The related theoretical upper bounds for each value of  $r$  are also given.

4.2. The Overlapping I.Non-I.D. Case with One Training Sample per Label

In this example, we consider the i.non-i.d. case where the probability distributions  $p_i(y_i|x_k)$  are overlapping for all  $i$ , as shown in Figure 7. The small orthogonal lines on the x-axis in Figure 7 represent alphabets, i.e., the elements in  $\mathcal{Y}$ , and the probability of occurrence of an alphabet  $y_i$  is equal to the intersection between the corresponding orthogonal line to the represented probability distribution  $p_i(y_i|x_k)$  for  $k = 1, 2$ . By “overlapping”, we mean the following. Let  $\mathcal{Y}_v$  and  $\mathcal{Y}_u$  denote the set of outputs generated by  $p_v(y_v|x_k)$  and  $p_u(y_u|x_k)$ , respectively. If for any  $v$  and  $u$ ,  $\mathcal{Y}_v \cap \mathcal{Y}_u \neq \emptyset$  holds, we say that the output alphabets are overlapping.



**Figure 7.** Illustration of the probability distributions  $p_i(y_i|x_1)$  (upper figure) and  $p_i(y_i|x_2)$  (lower figure), for  $i = 1, 2, \dots, n$ .

To demonstrate the performance of our proposed classifier in the overlapping case, we assume that we have two different labels,  $\mathcal{X} = \{x_1, x_2\}$ , where the corresponding conditional probability distributions  $p_i(y_i|x_1)$  and  $p_i(y_i|x_2)$  are obtained as follows. For a given  $n$ , let  $\mathcal{Y} = \{-n, -n + 1, \dots, 0, \dots, n - 1, n\}$  be the set of all alphabets. Note that the size of  $\mathcal{Y}$  grows with  $n$ . Moreover, let  $\mathbf{u}_i$  and  $\mathbf{v}_i$  ( $1 \leq i \leq n$ ) be vectors of length  $2n + 1$ , given by

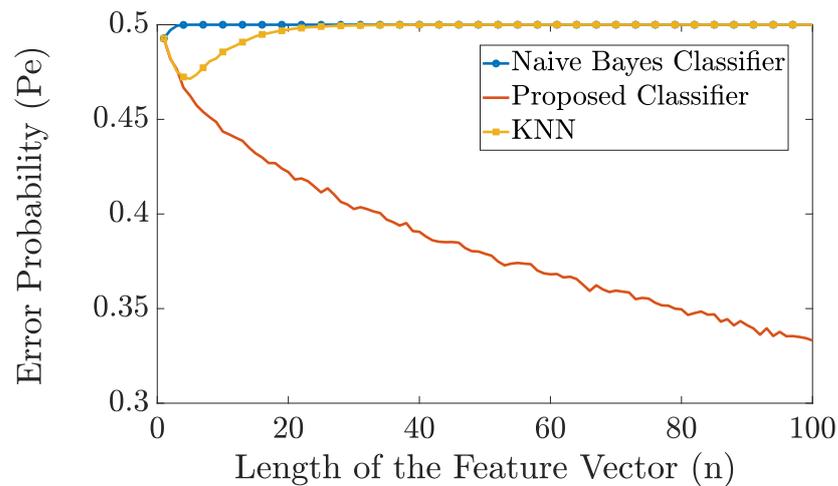
$$\mathbf{u}_i = \left[ 0, \dots, 0, \frac{1}{i(i+1)}, \frac{2}{i(i+1)}, \dots, \frac{i}{i(i+1)}, \frac{i+1}{i(i+1)}, \frac{i}{i(i+1)}, \dots, \frac{1}{i(i+1)}, 0, \dots, 0 \right], \quad (50)$$

$$\mathbf{v}_i = \left[ 0, \dots, 0, \frac{1}{i(i+1)}, \frac{1}{i(i+1)}, \dots, \frac{1}{i(i+1)}, \frac{1}{i(i+1)}, 0, \dots, 0 \right]. \quad (51)$$

The number of zeros in each side of the vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$  is  $(n - i)$ . To generate a feature vector from label  $x_1(x_2)$ , we generate the vector  $\mathbf{y}^n = (y_1, y_2, \dots, y_n)$ , where  $y_k$  takes values from the set  $\mathcal{Y}$ , with a probability distribution  $p_i(y_i|x_1) = \mathbf{u}_i(1 + 2(n + y_i))$  ( $p_i(y_i|x_2) = \mathbf{v}_i(1 + 2(n + y_i))$ ).

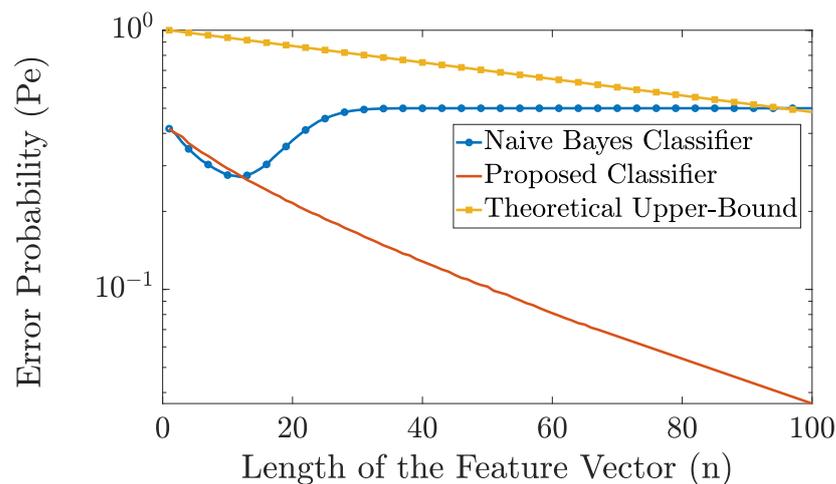
The simulation is carried out as follows. For each  $n$ , we generate one training feature vector for each label. Then, we generate 1000 feature vectors for each label and pass them through our proposed classifier, the naive Bayes classifier, and the KNN algorithm and calculate the error probability. We change the length of the feature vector from  $n = 1$  to  $n = 100$  and repeat the simulation 1000 times and then plot the error probability.

As shown in Figure 8, there is a huge difference between the performance of the two benchmark classifiers and the proposed classifier. The error probability of the naive Bayes classifier is almost 0.5 for all shown values of  $n$  as it is susceptible to the problem of unseen alphabets in the training vectors. The error probability of the KNN classifier is also almost 0.5 for  $n > 20$  as it is susceptible to the “curse of dimensionality”. However, the error probability of our proposed classifier continuously decays as  $n$  increases.



**Figure 8.** Comparison in error probability between the naive Bayes classifier, KNN, and the proposed classifier ( $T = 1$ ).

In Figure 9, we run the same experiments as in Figure 8 but with  $T = 100$ , i.e., 100 training feature vectors per label. As can be seen from Figure 9, the performance of the proposed classifier is better than the naive Bayes classifier, for  $n > 15$ . Since  $|\mathcal{Y}| = 2n + 1$ , for small values of  $n$ , the naive Bayes classifier has access to many training samples and, thereby, its performance is very close to the case when the probability distribution  $p_{\mathcal{Y}^n|x}(y^n|x)$  is known, i.e., to the maximum-likelihood classifier, and hence it has the optimal performance. As  $n$  increases, the number of alphabets rises, i.e.,  $|\mathcal{Y}|$  rises, and due to the aforementioned issue of the naive Bayes classifier with unseen alphabets, our proposed classifier performs much better classification than the naive Bayes classifier. Furthermore, note that the error probability of our proposed classifier decays exponentially as  $n$  increases which is not the case with the naive Bayes classifier. Moreover, Figure 9 also shows the theoretical upper bound on the error probability we derived in (11).



**Figure 9.** Comparison in error probability between the naive Bayes classifier and the proposed classifier ( $T = 100$ ).

4.3. The Non-Overlapping I.Non-I.D. Case with One Training Sample for Each Label

In this example, we consider the i.non-i.d. case where the probability distributions  $p_j(y_j|x_i)$  are non-overlapping for all  $j$  as shown in Figure 10, where we defined “overlapping” in Section 4.2. Hence, we test the other extreme in terms of possible distribution of the elements in the feature vectors  $Y^n$ .

To demonstrate the performance of our proposed classifier in the non-overlapping case, we assume that we have two different labels  $\mathcal{X} = \{x_1, x_2\}$ , the corresponding conditional probability distributions  $p_i(y_i|x_1)$  and  $p_i(y_i|x_2)$  are obtained as follows. For a given  $n$ , let  $\mathcal{Y} = \{1, 2, 3, \dots, (n + 1)^2 - 1\}$  be the set of all alphabets of the element in the feature vectors. Note again that the size of  $\mathcal{Y}$  grows with  $n$ . in addition, let  $\mathbf{u}_i$  and  $\mathbf{v}_i$  for  $(1 \leq i \leq n)$ , be vectors of length  $(n + 1)^2 - 1$ , given by

$$\mathbf{u}_i = \left[ 0, \dots, 0, \frac{1}{i(i+1)}, \frac{2}{i(i+1)}, \dots, \frac{i}{i(i+1)}, \frac{i+1}{i(i+1)}, \frac{i}{i(i+1)}, \dots, \frac{1}{i(i+1)}, 0, \dots, 0 \right], \tag{52}$$

$$\mathbf{v}_i = \left[ 0, \dots, 0, \frac{1}{i(i+1)}, \frac{1}{i(i+1)}, \dots, \frac{1}{i(i+1)}, \frac{1}{i(i+1)}, 0, \dots, 0 \right]. \tag{53}$$

The number of zeros in the left-hand sides of  $\mathbf{u}_i$  and  $\mathbf{v}_i$  is  $i^2 - 1$ . To generate a feature vector from the label  $x_1(x_2)$ , we generate the vector  $y^n = (y_1, y_2, \dots, y_n)$ , where  $y_k$  take values from the set  $\mathcal{Y}$ , with probability distribution  $p_i(y_i|x_1) = \mathbf{u}_i(y_i)$  ( $p_i(y_i|x_2) = \mathbf{v}_i(y_i)$ ).

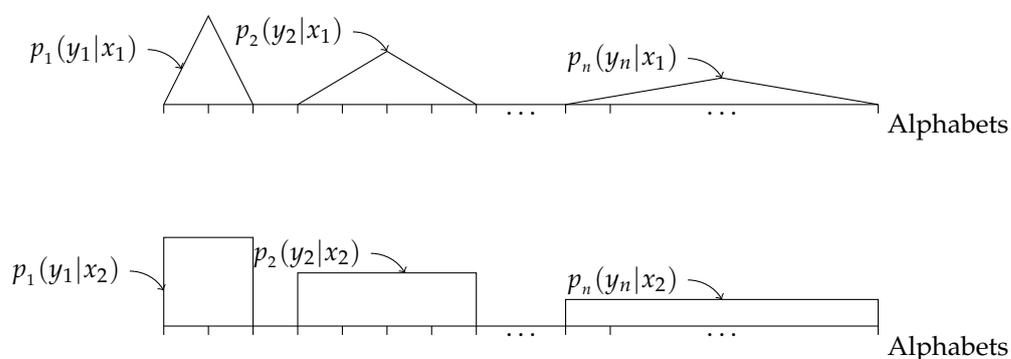
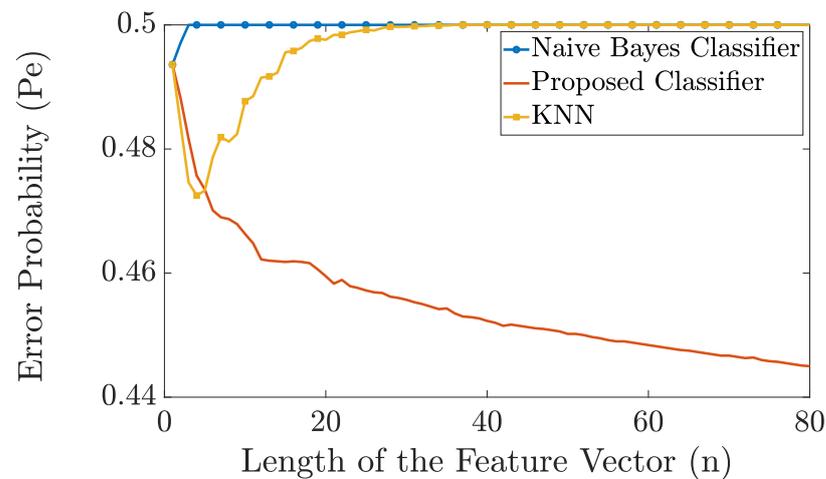


Figure 10. Illustration of the probability distributions  $p_i(y_i|x_1)$  (upper figure) and  $p_i(y_i|x_2)$  (lower figure), for  $i = 1, 2, \dots, n$ .

The simulation is carried out as follows. For each  $n$ , we generate one training feature vector for each label. Then we generate 250 feature vectors for each label and pass it through our proposed classifier, the naive Bayes classifier and KNN and calculate the error probabilities. We change the length of the vector from 1 to 80 and repeat the simulation 250 times and then plot the error probability. As shown in Figure 11, there is a huge difference between the performance of the proposed classifier and the two benchmark classifiers. The error probability of the naive Bayes classifier is almost 0.5 for all shown values of  $n$  as it is susceptible to the issue with unseen alphabets in the training feature vector. The error probability of the KNN classifier is almost 0.5 for all shown values of  $n > 30$  as it becomes susceptible to the “curse of dimensionality”. However, the error probability of our proposed classifier still decays continuously as  $n$  increases.

Note that, in our numerical examples, we compared our algorithm with the benchmark schemes on two extreme cases of i.non-i.d. vectors, referred to as “overlapping” and “non-overlapping”. Any other i.non-i.d. vector can be represented as a combination of the “overlapping” and “non-overlapping” vectors. Since our algorithm works better than the benchmark schemes for small  $t$  on both these cases, it will work better than the benchmark schemes on any combination between “overlapping” and “non-overlapping” vectors, i.e., for any other i.non-i.d. vectors.



**Figure 11.** Comparison in error probability between the naive Bayes classifier and the proposed classifier ( $T = 1$ ).

## 5. Conclusions

In this paper, we proposed a supervised classification algorithm that assigns labels to input feature vectors with independent but non-identically distributed elements, a statistical property found in practice. We proved that the proposed classifier is asymptotically optimal since the error probability moves to zero as the length of the input feature vectors grows. We showed that this asymptotic optimality is achievable even when one training feature vector per label is available. In the numerical examples, we compared the proposed classifier with the naive Bayes classifier and the KNN algorithm. Our numerical results show that the proposed classifier outperforms the benchmark classifiers when the number of training data is small and the length of the input feature vectors is sufficiently large.

**Author Contributions:** Methodology, F.S. and N.Z.; software, F.S.; formal analysis, F.S. and N.Z.; investigation, F.S.; supervision, N.Z.; writing—original draft preparation, F.S.; writing—review and editing, N.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable to this article as no new data were created or analyzed in this study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Proof of Corollary 2

The proof is almost identical to the proof of Theorem 1; however, here we derive a looser upper-bound on the error-probability than that in (11), which is independent of  $P_{y_i^{nt}}$ .

Without loss of generality we assume that  $x_1$  is the input to  $p_{Y^n|X}(y^n|x)$  and  $y^n$  is observed at the classifier.

Let  $\mathcal{B}_{k,l}^\epsilon$ , for  $1 \leq k \leq |\mathcal{Y}|$  and  $1 \leq l \leq |\mathcal{X}|$ , be a set defined as

$$\mathcal{B}_{k,l}^\epsilon = \left\{ \hat{y}^{nt} : \left| \frac{\mathcal{I}[\hat{y}^{nt} = y_k]}{nt} - \bar{p}(y_k|x_l) \right| \leq \frac{\epsilon}{\sqrt[3]{t}} \right\}. \quad (\text{A1})$$

Let  $\mathcal{B}_i^\epsilon = \bigcap_{k=1}^{|\mathcal{Y}|} \mathcal{B}_{k,l}^\epsilon$ . For  $\hat{y}_1^{nt} \in \mathcal{B}_1^\epsilon$ , we have

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} \stackrel{(a)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left( \frac{\epsilon}{\sqrt[3]{t}} \right)^r \right)^{1/r}, \tag{A2}$$

Using the same derivation as (18), for any  $y^n \in \mathcal{A}^\epsilon$  and for  $\hat{y}_1^{nt} \in \mathcal{B}_1^\epsilon$ , we have:

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nT} \right|^r \right)^{1/r} \leq |\mathcal{Y}|^{1/r} \epsilon + |\mathcal{Y}|^{1/r} \frac{\epsilon}{\sqrt[3]{t}}. \tag{A3}$$

On the other hand, the same as the derivation in (21), for each  $i \neq 1$ , we have:

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \geq \|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon. \tag{A4}$$

Now, for any  $\hat{y}_i^{nt} \in \mathcal{B}_i^\epsilon$ , we have

$$\begin{aligned} & \|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r + \left( \sum_{k=1}^{|\mathcal{Y}|} \left( \frac{\epsilon}{\sqrt[3]{t}} \right)^r \right)^{1/r} \\ & \stackrel{(a)}{\geq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} - \bar{p}(y_k|x_1) \right|^r \right)^{1/r} + \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} - \bar{p}(y_k|x_i) \right|^r \right)^{1/r} \\ & \stackrel{(b)}{\geq} \left( \sum_{k=1}^{|\mathcal{Y}|} |\bar{p}(y_k|x_1) - \bar{p}(y_k|x_i)|^r \right)^{1/r}, \end{aligned} \tag{A5}$$

where (a) follows from (A1) and (b) is again due to the Minkowski inequality. The expression in (A5), can be written equivalently as

$$\|P_{\hat{y}_i^{nt}} - \bar{P}_1\|_r \geq \|\bar{P}_i - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \frac{\epsilon}{\sqrt[3]{t}}. \tag{A6}$$

where  $i \neq 1$ . Using the bounds in (A6) and (A4), for any  $i \neq 1$  we have

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \geq \|\bar{P}_i - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right). \tag{A7}$$

Using the bounds in (A3) and (A7), we now relate the left-hand sides of (A3) and (A7) as follows. As long as the following inequality holds for each  $i \neq 1$ ,

$$|\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{T}} \right) < \|\bar{P}_i - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right), \tag{A8}$$

which is equivalent to the following for  $i \neq 1$

$$\epsilon < \frac{\|\bar{P}_i - \bar{P}_1\|_r}{2(1 + t^{-1/3})|\mathcal{Y}|^{1/r}}, \tag{A9}$$

we have the following for  $i \neq 1$

$$\begin{aligned} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} \right|^r \right)^{1/r} &\stackrel{(a)}{\leq} |\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right) \\ &\stackrel{(b)}{<} \|\bar{P}_i - \bar{P}_1\|_r - |\mathcal{Y}|^{1/r} \epsilon \left( 1 + \frac{1}{\sqrt[3]{t}} \right) \\ &\stackrel{(c)}{\leq} \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r}, \end{aligned} \tag{A10}$$

where (a), (b), and (c) follow from (A3), (A8), and (A7), respectively. Thereby, from (A10), we have the following for  $i \neq 1$

$$\left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_1^{nt} = y_k]}{nt} \right|^r \right)^{1/r} \leq \left( \sum_{k=1}^{|\mathcal{Y}|} \left| \frac{\mathcal{I}[y^n = y_k]}{n} - \frac{\mathcal{I}[\hat{y}_i^{nt} = y_k]}{nt} \right|^r \right)^{1/r}, \tag{A11}$$

or equivalently as

$$\|P_{y^n} - P_{\hat{y}_1^{nt}}\|_r < \|P_{y^n} - P_{\hat{y}_i^{nt}}\|_r. \tag{A12}$$

Once again, we obtained that if there is an  $\epsilon$  for which (A9) holds for  $i \neq 1$  and for that  $\epsilon$  there are sets  $\mathcal{A}^\epsilon$  and  $\mathcal{B}_i^\epsilon$  for which  $y^n \in \mathcal{A}^\epsilon$  and  $\hat{y}_j^{nt} \in \mathcal{B}_i^\epsilon$  for all  $1 \leq l \leq |\mathcal{X}|$ , then (A12) holds for  $i \neq 1$ , and thereby our classifier will detect that  $x_1$  is the correct label. Using this, we can upper-bound the error probability as

$$\begin{aligned} \mathbb{P}_e &= 1 - \Pr\{\hat{x}_1 = x_1\} \\ &\leq 1 - \Pr\left\{ (y^n \in \mathcal{A}^\epsilon) \cap \left( \bigcap_{j=1}^{|\mathcal{X}|} \hat{y}_j^{nt} \in \mathcal{B}_i^\epsilon \right) \middle| \epsilon \in \mathcal{S} \right\}, \end{aligned} \tag{A13}$$

where  $\mathcal{S}$  is a set defined as

$$\mathcal{S} = \left\{ \epsilon : \epsilon \leq \min_{i \neq 1} \frac{\|\bar{P}_i - \bar{P}_1\|_r}{(2 + t^{-1/3})|\mathcal{Y}|^{1/r}} \right\}. \tag{A14}$$

The right-hand side of (A13) can be upper-bounded as

$$\begin{aligned} 1 - \Pr\left\{ (y^n \in \mathcal{A}^\epsilon) \cap \left( \bigcap_{l=1}^{|\mathcal{X}|} \hat{y}_l^{nt} \in \mathcal{B}_i^\epsilon \right) \middle| \epsilon \in \mathcal{S} \right\} &= \Pr\left\{ (y^n \notin \mathcal{A}^\epsilon) \cup \left( \bigcup_{l=1}^{|\mathcal{X}|} \hat{y}_l^{nt} \notin \mathcal{B}_i^\epsilon \right) \middle| \epsilon \in \mathcal{S} \right\} \\ &\stackrel{(a)}{\leq} \Pr\{y^n \notin \mathcal{A}^\epsilon | \epsilon \in \mathcal{S}\} \\ &\quad + \sum_{l=1}^{|\mathcal{X}|} \Pr\{\hat{y}_l^{nt} \notin \mathcal{B}_i^\epsilon | \epsilon \in \mathcal{S}\}, \end{aligned} \tag{A15}$$

Using the same derivation as (39), we have:

$$\Pr\{y^n \notin \mathcal{A}^\epsilon | \epsilon \in \mathcal{S}\} \leq 2|\mathcal{Y}|e^{-2n\epsilon^2}. \tag{A16}$$

Similarly, we have the following result for the second expression in the right-hand side of (A15), which is the same as the derivation in (43)

$$\Pr\{\hat{y}_l^{nt} \notin \mathcal{B}_i^\epsilon | \epsilon \in \mathcal{S}\} \leq 2|\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2}. \tag{A17}$$

Inserting (A16) and (A17) into (A15), and then inserting (A15) into (A13), we obtain the following upper-bound for the error probability

$$\mathbb{P}_e \leq 2|\mathcal{Y}|e^{-2n\epsilon^2} + 2|\mathcal{X}||\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2}, \quad (\text{A18})$$

where

$$\epsilon = \min_{\substack{i,j \\ i \neq j}} \frac{\|\bar{\mathbb{P}}_i - \bar{\mathbb{P}}_j\|_r}{2(1+t^{-1/3})|\mathcal{Y}|^{1/r}}, \quad (\text{A19})$$

Now, if  $|\mathcal{Y}| \leq n^m$ , (A18) can be written as

$$\begin{aligned} \mathbb{P}_e &\leq 2|\mathcal{Y}|e^{-2n\epsilon^2} + 2|\mathcal{X}||\mathcal{Y}|e^{-2nt^{1/3}\epsilon^2} \\ &\leq 2n^m \exp\left(-2n \min_{\substack{i,j \\ i \neq j}} \frac{\|\bar{\mathbb{P}}_i - \bar{\mathbb{P}}_j\|_r^2}{2(1+t^{-1/3})2n^{2m/r}}\right) \\ &\quad + 2|\mathcal{X}|n^m \exp\left(-2nt^{1/3} \min_{\substack{i,j \\ i \neq j}} \frac{\|\bar{\mathbb{P}}_i - \bar{\mathbb{P}}_j\|_r^2}{2(1+t^{-1/3})2n^{2m/r}}\right) \\ &\leq \mathcal{O}\left(n^m \exp\left(-n^{1-\frac{2m}{r}}\right)\right). \end{aligned} \quad (\text{A20})$$

According to (A20), for a fixed  $r > 2m$ , the right-hand side of (A20) moves to zero as  $n \rightarrow \infty$  and, thereby, the classifier is asymptotically optimal.

## References

- Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, MA, USA, 2014.
- Quinlan, J.R. *Learning Efficient Classification Procedures and Their Application to Chess End Games*; Springer: Berlin/Heidelberg, Germany, 1983; pp. 453–482.
- Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth and Brooks: Monterey, CA, USA, 1984.
- Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, San Jose, CA, USA, 27–29 July 1992; pp. 144–152.
- Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
- Lallich, S.; Teytaud, O.; Prudhomme, E. *Association Rule Interestingness: Measure and Statistical Validation*; Studies in Computational Intelligence; Springer: Berlin, Germany, 2007; Volume 43, pp. 251–275.
- Langley, P.; Iba, W.; Thompson, K. An analysis of bayesian classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA, 14–16 July 1992; AAAI Press: San Jose, CA, USA, 1992; p. 223228.
- Aha, D.W. Lazy learning. In *Artificial Intelligent Review*; Kluwer Academic Publishers: Cambridge, MA, USA, 1997; p. 710.
- Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining*; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
- Kanaya, F.; Han, T.S. The asymptotics of posterior entropy and error probability for bayesian estimation. *IEEE Trans. Inf. Theory* **1995**, *41*, 1988–1992. [[CrossRef](#)]
- Gutman, M. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Inf. Theory* **1989**, *35*, 401–408. [[CrossRef](#)]
- Sason, I.; Verd, S. Arimoto-nyi conditional entropy and bayesian hypothesis testing. *IEEE Trans. Inf. Theory* **2018**, *64*, 4–25. [[CrossRef](#)]
- Wang, C.; She, Z.; Cao, L. Coupled attribute analysis on numerical data. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 2–9 August 2013; AAAI Press: Beijing, China, 2013; p. 17361742.
- Wang, C.; Dong, X.; Zhou, F.; Cao, L.; Chi, C.H. Coupled attribute similarity learning on categorical data. *IEEE Trans. Neural Networks Learn. Syst.* **2015**, *26*, 781–797. [[CrossRef](#)] [[PubMed](#)]
- Wang, C.; Cao, L.; Wang, M.; Li, J.; Wei, W.; Ou, Y. Coupled nominal similarity in unsupervised learning. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Scotland, UK, 24–28 October 2011; Association for Computing Machinery: Glasgow, Scotland, UK, 2011; p. 973978.

17. Liu, C.; Cao, L. A coupled k-nearest neighbor algorithm for multi-label classification. In *PAKDD*; Springer International Publishing: Cham, Switzerland, 2015.
18. Liu, C.; Cao, L.; Yu, P. A hybrid coupled k-nearest neighbor algorithm on imbalance data. In Proceedings of the International Joint Conference on Neural Networks, Beijing, China, 6–11 July 2014; pp. 2011–2018.
19. Liu, C.; Cao, L.; Yu, P.S. Coupled fuzzy k-nearest neighbors classification of imbalanced non-iid categorical data. In Proceedings of the 2014 International Joint Conference on Neural Networks IJCNN, Beijing, China, 8–13 July 2014; pp. 1122–1129.
20. Cao, L. Non-IIDness Learning in Behavioral and Social Data. *Comput. J.* **2013**, *57*, 1358–1370. [[CrossRef](#)]
21. Cao, L. Coupling learning of complex interactions. *Inf. Process. Manag.* **2015**, *51*, 167–186. [[CrossRef](#)]
22. Cao, L.; Zhang, H.; Zhao, Y.; Luo, D.; Zhang, C. Combined mining discovering informative knowledge in complex data. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2011**, *41*, 699–712.
23. Getoor, L.; Taskar, B. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2007.
24. Abed-Meraim, K.; Loubaton, P.; Moulines, E. A subspace algorithm for certain blind identification problems. *IEEE Trans. Inf. Theory* **2006**, *43*, 499–511. [[CrossRef](#)]
25. Almeida, L. Linear and nonlinear ica based on mutual information—The misep method. *IEEE Trans. Signal Process.* **2004**, *84*, 231–245. [[CrossRef](#)]
26. Hyvarinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411430. [[CrossRef](#)]
27. Yeredor, A. Independent component analysis over galois fields of prime order. *IEEE Trans. Inf. Theory* **2011**, *57*, 53425359. [[CrossRef](#)]
28. Nguyen, H.; Zheng, R. Binary independent component analysis with or mixtures. *IEEE Trans. Signal Processing* **2011**, *59*, 31683181. [[CrossRef](#)]
29. Barlow, H.B. Unsupervised learning. *Neural Comput.* **1989**, *1*, 295311. [[CrossRef](#)]
30. Nakhaeizadeh, G.; Taylor, C.C.; Kunisch, G. Dynamic Supervised Learning: Some Basic Issues and Application Aspects. In *Classification and Knowledge Organization*; Springer: Berlin/Heidelberg, Germany, 1997; pp. 123–135.
31. Koppen, M. *The Curse of Dimensionality*; SAGE: London, UK, 2000; pp. 4–8.
32. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2000.
33. Jain, A.K.; Duin, R.P.W.; Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 437. [[CrossRef](#)]
34. Xu, B.; Huang, J.; Williams, G.; Wang, Q.; Ye, Y. Classifying very high-dimensional data with random forests built from small subspaces. *Int. J. Data Warehous. Min.* **2012**, *8*, 44–63. [[CrossRef](#)]
35. Ye, Y.; Wu, Q.; Huang, J.Z.; Ng, M.K.; Li, X. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognit.* **2013**, *46*, 769–787. [[CrossRef](#)]
36. Kouiroukidis, N.; Evangelidis, G. The effects of dimensionality curse in high dimensional knn search. In Proceedings of the 15th Panhellenic Conference on Informatics, Kastoria, Greece, 30 September–2 October 2011; pp. 41–45.
37. Matusita, K. Classification based on distance in multivariate gaussian cases. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability; Volume 1: Statistics*; University of California Press: Regents, CA, USA, 1967; pp. 299–304.
38. Collins, M.; Schapire, R.E.; Singer, Y. Logistic regression, adaboost and bregman distances. *Mach. Learn.* **2002**, *48*, 253285. [[CrossRef](#)]
39. Deisenroth, M.P.; Faisal, A.A.; Ong, C.S. *Mathematics for Machine Learning*; Cambridge University Press: Warsaw, Poland, 2020.
40. Lebanon, G.; Lafferty, J. Cranking: Combining rankings using conditional probability models on permutations. In Proceedings of the 19th International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 2002; pp. 363–370.
41. Hoeffding, W. On the distribution of the number of successes in independent trials. *Ann. Math. Statist.* **1956**, *27*, 713–721. [[CrossRef](#)]
42. Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.* **1963**, *58*, 13–30. [[CrossRef](#)]