BINARY CLASSIFICATION WITH TRAINING UNDER BOTH CLASSES

BY

YUN LI

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor Venugopal V. Veeravalli

ABSTRACT

This thesis focuses on the binary classification problem with training data under both classes. We first review binary hypothesis testing problems and present a new result on the case of countably infinite alphabet. The goal of binary hypothesis testing is to decide between the two underlying probabilistic processes. Asymptotic optimality of binary hypothesis testing can be achieved with the knowledge of only one of the processes. It is also shown that the finite sample performance could improve greatly with additional knowledge of the alternate process. Most previous work focuses on the case where the alphabet is finite. This thesis extends the existing results to the case of countably infinite alphabet. It is proved that, without knowledge of the alternate process, the worst-case performance of any test is arbitrarily bad, even when the alternate process is restricted to be "far" in the sense of relative entropy.

Binary classification problems arise in applications where a full probabilistic model of either of the processes is absent and pre-classified samples from both of the processes are available. It is known that asymptotic optimality can be achieved with the knowledge of only one pre-classified training sequence. We propose a classification function that depends on both training sequences. Then Stein's lemma for classification is proved using this new classification function. It states that the maximal error exponent under one class is given by the relative entropy between the conditional distributions of the two classes. Our results also shed light on how the classification errors depend on the relative size of the training and test data. It is shown in the simulation results that our classification method outperforms the asymptotically optimal one when the test samples are of limited size.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Hypothesis testing is a fundamental issue in statistics and many application areas. We restrict our attention to binary hypothesis testing problems in this work. There are two underlying processes that could have generated the samples. The goal of binary hypothesis testing is to test which one of the two processes generated the samples. There are mainly two scenarios for the problem. In one scenario, the probabilistic models of the two processes are fully specified. In the other scenario, only one of the two models is fully specified. The model of the other process is either partially known or completely unknown. In this work, we focus on the case where the underlying processes are stationary and memoryless. Thus the probabilistic model is fully specified by the marginal distribution of the samples. In the case where the two distributions are fully known, the loglikelihood ratio test is shown to be optimal under the Neyman-Pearson criterion. The test statistic depends on the two distributions and the samples to be tested. In the case where only one of the distributions is known and the other is completely unknown, Hoeffding proposed in the sixties a universal test statistic that only depends on the known distribution and the samples to be tested. The test is asymptotically optimal [1] under a modified Neyman-Pearson criterion. Though asymptotic optimality can be achieved with the knowledge of only one of the distributions, prior information of the other distribution is shown to be crucial to the finite sample size performance of the test [2].

Most of the previous works focus on the case where the alphabet size is finite. We study the case of countably infinite alphabet in this thesis. We prove that the worst-case performance of any test is arbitrarily bad, even when the unknown distribution is restricted to be "far" from the known distribution. The notion of "arbitrarily bad" is made clear in Chapter 3. This result implies that we need more prior information about the unknown distribution in order to guarantee uniform performance.

Classification problems arise in the areas where it is unrealistic to have full knowledge of either of the distributions. There are mainly two scenarios of the problem. In supervised classification, previously classified samples from one or both of the distributions are available. New samples are classified based on the knowledge of the classes learned from the previously classified samples. In unsupervised classification, even previously classified samples are not available. The classifier learns about the classes while classifying samples. The second part of this thesis focuses on the issue of supervised binary classification problems where samples are generated stationary and memoryless under both classes. It is proved in [3] that asymptotic optimality can be achieved by a classification function which depends only on the previously classified samples from one of the classes. The classification function is derived by formulating the classification problem as a composite versus composite hypothesis testing problem and a generalized likelihood ratio test is performed. The classification function is the same as the test statistics of the generalized likelihood ratio test.

We propose a classification function that utilizes previously classified samples from both classes. Stein's lemma for classification is proved in this thesis: the maximal error exponent under one class is characterized by the relative entropy between the two classes and the maximal error exponent is not achieved by the classification function in [3]. These results are consistent with the results in hypothesis testing problems [4]. We also found in the simulations that our method outperforms the method in [3] when the number of sample to be classified is limited. This shows the importance of utilizing both training sequences. Finite sample performance is a main concern in many applications because it can be expensive to accumulate samples, and sample size is usually associated with delay in decision making.

The rest of the thesis is organized as follows. In Chapter 2, we review the binary hypothesis testing problem with finite alphabet; a new result on countably infinite alphabet is also proved in Chapter 2. In Chapter 3, we show theoretical and simulation results for the binary classification problem. We conclude our work and discuss future work in Chapter 4.

CHAPTER 2

HYPOTHESIS TESTING

2.1 Definition and Background

In this section, we introduce background and definitions that we will use in future sections. We are concerned with the case where the underlying processes are stationary and memoryless. In other words, the samples to be tested X_1^n, $n \in \mathbb{N}$, are independent and identically distributed. We assume that the observations have a marginal which is absolutely continuous with respect to some measure μ. We denote the probability mass function of the observations as p_0 or p_1, both of which take values in the alphabet $\mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$. $\mathcal{A}$ can be either finite or countably infinite. We will refer to p_0 as the null distribution and p_1 as the alternate distribution. The hypotheses are

$$\begin{aligned} \mathrm{H}_0: \quad & X_1^n \sim p_0 \\ \mathrm{H}_1: \quad & X_1^n \sim p_1. \end{aligned} \tag{2.1}$$

A decision rule is characterized by a sequence of tests δ_1^n, where $\delta_n : X^n \to \{0, 1\}$ is a function that maps the observations to a binary decision. $\delta_n = 0$ represents a decision that is in favor of accepting p_0 as the true marginal distribution. The performance of a test is measured by the false alarm and missed-detection probability, which are defined as

$$\mathrm{P_F}(\delta_n, p_0) \triangleq \Pr(\delta_n(X_1^n) = 1 | X_1^n \sim p_0) \tag{2.2}$$

and

$$\mathrm{P_M}(\delta_n, p_1) \triangleq \Pr(\delta_n(X_1^n) = 0 | X_1^n \sim p_1). \tag{2.3}$$

We say a test is consistent if $\mathrm{P_F}$ and $\mathrm{P_M}$ converge to zero as the size of the test sequence goes to infinity.

We need to prescribe a performance criterion in order to compare various tests. A commonly used criterion is the Neyman-Pearson criterion:

$$\begin{aligned} \min \quad & \mathrm{P_M}(\delta_n, p_1) \\ \text{s.t.} \quad & \mathrm{P_F}(\delta_n, p_0) \ \leq \alpha. \end{aligned} \tag{2.4}$$

In the case where p_0 and p_1 are fully known, the loglikelihood ratio test is shown to be optimal under the Neyman-Pearson criterion. The test statistic of the loglikelihood ratio test is

$$\frac{1}{n} \log \frac{p_1(X_1^n)}{p_0(X_1^n)}. \tag{2.5}$$

In applications like anomaly detection, normal behavior is usually unique to the system, but abnormal behavior can be anything other than the normal behavior. For example, one may want to detect if there is a malicious entity (i.e., a Trojan horse) tampering a computer. But one is unaware of the skills that this malicious entity has. Or one may want to decide if a power network is working normally or not by observing its output. Anomalous behavior of a power network might be hard to model due to the fact that the network is vast in size and affected by various outside entities. As a result, the null distribution that characterizes the normal behavior of the system is usually unique, but the alternate distribution can sometimes be anything but the null distribution. In this case, we do not have a simple H_1 anymore, and the hypotheses become

$$\begin{aligned} \mathrm{H}_0 : \quad & X_1^n \sim p_0 \\ \mathrm{H}_1 : \quad & X_1^n \sim p_1 \in S \end{aligned} \tag{2.6}$$

where S is the class of possible alternate distributions. The above situation is called a simple versus composite hypothesis testing problem. And the assumption that p_1 is in a certain set S serves as the prior information about the alternate distribution. Most of the time, it is desirable to have uniform performance guarantee over the set S. We say a test is uniformly consistent against S if both error probabilities converge to zero under any $p_1 \in S$. We say a test is exponentially uniformly consistent if the worst-case error probabilities against S are exponentially small. Intuitively, if we restricted S to be far from P_0, we would have exponentially uniform consistency against S. But we will see in later sections that this is not always the case. In the anomaly detection example mentioned earlier, $S = \{p_1 | p_1 \neq p_0\}$ and this scenario is often referred to as the *universal hypothesis testing problem*. Since we do not have full knowledge of what p_1 is, a test that works in (2.6) can not depend on p_1. Instead, the test statistic can only be a function of the null distribution and the parameters of S. In universal hypothesis testing problems, the test statistic can only be a function of p_0

due to the fact that there is no structure at all on S.

The generalized likelihood ratio test (GLRT) is a popular test used in the above simple versus composite hypothesis testing problem. In this case, the test statistic of GLRT is given by

$$\sup_{p_1 \in S} \frac{1}{n} \log \frac{p_1(X_1^n)}{p_0(X_1^n)}. \tag{2.7}$$

The test also works in composite versus composite problems with the supremum taken on both the numerator and the denominator.

The test sequence is given by

$$\delta_n(X_1^n) = \mathbb{I}\{\sup_{p_1 \in S} \frac{1}{n} \log \frac{p_1(X_1^n)}{p_0(X_1^n)} \geq \tau_n\} \tag{2.8}$$

where $\mathbb{I}$ is the indicator function and τ_n is referred to as the threshold.

In the situation where $S = \{p_1 | p_1 \neq p_0\}$, the test is called the Hoeffding test which was proposed by Hoeffding in the sixties. We next show that the test statistic can be further simplified in this situation.

The relative entropy between two distributions p, $q \in \mathbb{P}(\mathcal{A})$ satisfying $p \prec q$ is defined as

$$D(p\|q) \triangleq \sum_{i \in A} p(i) \log \frac{p(i)}{q(i)}. \tag{2.9}$$

We define a divergence ball of radius τ around p is defined as

$$\mathcal{Q}_\tau(p) \triangleq \{\tilde{p} \in \mathbf{P}(\mathcal{A}) : D(\tilde{p}\|p) < \tau\}. \tag{2.10}$$

We use q to denote any empirical distribution. So the empirical distribution or type of the observations X_1^n is denoted by $q_{X_1^n} \in \mathbb{P}(\mathcal{A})$ where

$$q_{X_1^n}(i) \triangleq \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}(X_k = i). \tag{2.11}$$

In the case where $S = \{p_1 | p_1 \neq p_0\}$, it is not hard to see that the Hoeffding test can be equally written as

$$\delta_n(X_1^n) = \mathbb{I}\{q_{X_1^n} \notin \mathcal{Q}_{\tau_n}(p_0)\}. \tag{2.12}$$

The Hoeffding test is proved to be asymptotically optimal under a modified version of the Neyman-Pearson criterion:

Among all decision rules $\Delta = \{\delta | \delta = \{\delta_n, n = 1, 2, ...\}\}$ that do not depend on the unknown p_1 and at the same time make sure that the false alarm error exponent

$$\alpha = \liminf_{n\to\infty} -\frac{1}{n} \log p_{\mathrm{F}}(\delta_n, p_0) \geqslant \lambda, \tag{2.13}$$

select a sequence that maximizes the missed-detection error exponent

$$\beta = \liminf_{n\to\infty} -\frac{1}{n} \log p_{\mathrm{M}}(\delta_n, p_1) \text{ for all } p_1 \in \mathbb{P}(\mathcal{A}). \tag{2.14}$$

The quantities α and β are called the error exponents and the definition of them is justified by large deviation analysis. We say a test is exponentially consistent if these two error exponents are strictly positive. It is usually intractable to obtain a closed form expression of $\mathrm{P_F}$ and $\mathrm{P_M}$ as functions of n. The analysis of error exponents provides approximation of the test performance as a function of n. It is closely related to the channel reliability rate function [5]. So the analysis of error exponents is a key issue in studying the asymptotic behavior of a test.

Note that the test statistic of GLRT may not necessarily be written in terms of relative entropy for arbitrary S. In addition, GLRT may not achieve optimality for arbitrary S either. The sufficient condition for GLRT to be optimal, counter examples in which it is not, can be found in [6].

2.2 Finite Alphabet

2.2.1 Asymptotic Analysis of the Hoeffding Test

In this section, we study the asymptotic behavior of the Hoeffding test when $|\mathcal{A}| < \infty$. The analysis implies that the performance of the Hoeffding test is compromised when the number of observations is limited compared to $|\mathcal{A}|$. Then the next section shows how to incorporate prior information of the alternate distribution to improve the performance of the Hoeffding test.

The following theorem establishes the asymptotic behavior of the Hoeffding test.

Theorem 2.1 (Asymptotic Behavior of the Hoeffding Test)**.** *Assume p_0 and p_1 have full support over $\mathcal{A}$.*

- *Suppose that the observations X_1^n are i.i.d. with marginal p_0. Then the normalized Hoeffding test statistic sequence $\{nD(q_{X_1^n}||p_0) : n \geq 1\}$ has the following asymptotic* *bias and variance*

$$\lim_{n\to\infty} \boldsymbol{E}[nD(q_{X_1^n}\|p_0)] \;=\; \frac{1}{2}(|\mathcal{A}|-1) \tag{2.15}$$

$$\lim_{n\to\infty} \boldsymbol{Var}[nD(q_{X_1^n}\|p_0)] \;=\; \frac{1}{2}(|\mathcal{A}|-1). \tag{2.16}$$

Furthermore, the following weak convergence result holds,

$$nD(q_{X_1^n}\|P_0) \;\xrightarrow[n\to\infty]{\text{d.}}\; \frac{1}{2}\chi^2_{|\mathcal{A}|-1} \tag{2.17}$$

where $\chi^2_{|\mathcal{A}|-1}$ *denotes the chi-square distribution with* $(|\mathcal{A}|-1)$ *degrees of freedom.*

- *Suppose the sequence* X_1^n *is i.i.d. under* $p_1 \neq p_0$*. We have with* $\sigma^2 \triangleq \mathrm{Var}_{p_1}(\log\frac{p_1}{p_0})$

$$\lim_{n\to\infty} \boldsymbol{E}[n(D(q_{X_1^n}\|p_0) - D(p_1\|p_0))] \;=\; \frac{1}{2}(|\mathcal{A}|-1) \tag{2.18}$$

$$\lim_{n\to\infty} \boldsymbol{Var}[n^{\frac{1}{2}}D(q_{X_1^n}\|p_0)] \;=\; \sigma^2 \tag{2.19}$$

$$n^{\frac{1}{2}}(D(q_{X_1^n}\|p_0) - D(p_1\|p_0)) \;\xrightarrow[n\to\infty]{\text{d.}}\; \mathcal{N}(0,\sigma^2). \tag{2.20}$$

The bias result of (2.15) follows from the unpublished report [7] and the weak convergence result of (2.17) follows from the result of [8]. The rest of the results follow from [2]. Unlike the well-known result of the error exponents of the Hoeffding test which follow from large deviation theory [4], the above results are derived from Taylor expansion and then a central limit theory analysis. As seen in [2], the weak convergence result can be used to set threshold for a finite sample size test based on a prescribed false alarm probability. And it turns out that, when the sample size is small, this approximation of error probabilities which follow from a central limit theorem analysis is more accurate than that from a large deviation analysis. Simulations of this can be seen in [2].

As we can see from (2.15)and (2.19), the bias of the test statistic is positive under either p_0 or p_1, and it is linear with $(|\mathcal{A}|-1)$. When the sample size is limited and $|\mathcal{A}|$ is large, the bias term can be significant. This can possibly be addressed by setting a higher threshold that incorporates this positive bias. However, the variance of the test statistic under p_0 is also linear with $(|\mathcal{A}|-1)$. The high variance implies that the decision region corresponding to p_0 needs to be large in order to guarantee the prescribed false alarm probability. As a result, the probability of missed-detection might be significant due to the fact that the decision region of p_1 is the complement of that of p_0. In other words, this test is not reliable in situations where the square root of the sample size is small compared to the alphabet size.

2.2.2 Performance Improvement of the Hoeffding Test

In this section, we show how prior information can improve the finite sample size performance of the Hoeffding test. In [2], the mismatched test is proposed, which is based on a relaxation of the Hoeffding test statistic. The relaxation itself relies on additional information about the set S that p_1 belongs to.

The relative entropy can be equally expressed as the convex dual of the log moment generating function [9]. For any p_0 and $p_1 \in \mathbb{P}(\mathcal{A})$

$$D(p_1||p_0) = \sup_f (p_1(f) - \Lambda_{p_0}(f)) \tag{2.21}$$

where the supremum is taken over the space of all real-valued functions on $\mathcal{A}$. Furthermore, if p_0 and p_1 have equal supports, the supremum is achieved by the log likelihood ratio function

$$f^* = \log \frac{p_1}{p_0}. \tag{2.22}$$

Also note that the above definition is invariant to an addition of a constant. So the supremum is also achieved by $(\log \frac{p_1}{p_0} + c)$ for any c a constant.

We can get a lower bound on the relative entropy if we fix $f \in \mathcal{F}$ for some function class $\mathcal{F}$. This lower bound is named as the mismatched divergence in [2]:

$$D^{MM}(p_1||p_0) = \sup_{f \in \mathcal{F}} \{p_1(f) - \Lambda_{p_0}(f)\}. \tag{2.23}$$

Then the mismatched test sequence is given by replacing the divergence by mismatched divergence in (2.8)

$$\delta_n^{MM}(X_1^n) = \mathbb{I}\{q_{X_1^n} \notin \mathcal{Q}_{\tau_n}^{MM}(p_0)\} \tag{2.24}$$

where

$$\mathcal{Q}_{\tau_n}^{MM}(p_0) \triangleq \{p \in \mathbb{P}(\mathcal{A}) : D^{MM}(p||p_0) < \tau_n\} \tag{2.25}$$

is the mismatched divergence ball around P_0 with radius τ_n.

Now we show how the mismatched test outperforms the Hoeffding test when the sample size is small. For the purpose of this thesis, we restrict our attention to linear function class $\mathcal{F}$. Note that the assumption of linear function class can be relaxed [2]. Let $\{\psi_i : 1 \leq i \leq d\}$ be d functions on $\mathcal{A}$. And $\psi = \{\psi_1, \psi_2, \cdots, \psi_d\}^T$ and let $\mathcal{F}$ be the linear function class with basis ψ.

$$\mathcal{F} = \{f_r = \sum_{i=1}^{d} r_r \psi_i : r \in \mathbb{R}^d\} \tag{2.26}$$

In addition, we assume the following assumptions hold:

- There exists an open neighborhood $B \subset \mathbb{P}(\mathcal{A})$ of q_0 such that for each $q \in B$, the supremum in the definition of $D^{MM}(q||q_0)$ is achieved at a unique point $r(q)$
- The vectors $\{\psi_1, \psi_1, \cdots, \psi_{d-1}\}$ are linearly independent over the support of p_0, where $\psi_1 = 1$

Then the following theorem holds.

Theorem 2.2 (Asymptotic Analysis of the Mismatched Test)**.** *Suppose that the observations* X_1^n *are i.i.d. with marginal* p*. Suppose that there exists* r^* *satisfying* $f_{r^*} = \log \frac{p}{p_0}$*. Further, suppose that the above assumptions hold with* $q_0 = p$*, then*

- *When* $p = p_0$*,*

$$\lim_{n\to\infty} \boldsymbol{E}[nD^{MM}(q_{X_1^n}\|p_0)] \;=\; \frac{1}{2}d \tag{2.27}$$

$$\lim_{n\to\infty} \boldsymbol{Var}[nD^{MM}(q_{X_1^n}\|p_0)] \;=\; \frac{1}{2}d \tag{2.28}$$

$$nD^{MM}(q_{X_1^n}\|p_0) \;\xrightarrow[n\to\infty]{\text{d.}}\; \frac{1}{2}\chi_d^2 \tag{2.29}$$

- *When* $p = p_1 \neq p_0$*, we have with* $\sigma^2 \triangleq \mathrm{Var}_{p_1}(\log \frac{p_1}{p_0})$

$$\lim_{n\to\infty} \boldsymbol{E}[n(D^{MM}(q_{X_1^n}\|p_0) - D^{MM}(p_1\|p_0))] \;=\; \frac{1}{2}d \tag{2.30}$$

$$\lim_{n\to\infty} \boldsymbol{Var}[n^{\frac{1}{2}}D^{MM}(q_{X_1^n}\|p_0)] \;=\; \sigma^2 \tag{2.31}$$

$$n^{\frac{1}{2}}(D^{MM}(q_{X_1^n}\|p_0) - D^{MM}(p_1\|p_0)) \;\xrightarrow[n\to\infty]{\text{d.}}\; \mathcal{N}(0,\sigma^2) \tag{2.32}$$

First note that the asymptotic bias and variance of the mismatched test statistic is linear with the dimension of the function class. Given the number of observations, we can choose the dimension to insure that the bias and variance of the test statistic is within certain tolerance.

Also note that if there exists r^* satisfying $f_{r^*} = \log \frac{p}{p_0}$, then $\log \frac{p^\lambda}{p_0} \in \mathcal{F}$ for any $\lambda \in [0,1]$. p^λ is the twisted distribution between P_1 and P_0

$$p^\lambda = \frac{p_0^\lambda p_1^{1-\lambda}}{\sum p_0^\lambda p_1^{1-\lambda}} s \tag{2.33}$$

and

$$\log \frac{p^\lambda}{p_0} = \lambda \log \frac{p_1}{p_0} - \log(\sum p_0^\lambda p_1^{1-\lambda}) \tag{2.34}$$

Because (2.21) is invariant to an addition of constant, $\lambda \log \frac{p_1}{p_0}$ also achieves the supremum. If $f_{r^*} = \log \frac{p}{p_0}$ is in the function class, $\lambda \log \frac{p_1}{p_0} \in \mathcal{F}$ because of the linearity of $\mathcal{F}$. So the mismatched divergence between p^λ and p_0 coincides with the relative entropy between them. With some large deviation analysis, it is shown that the mismatched test is still optimal under the Neyman-Pearson criterion (2.4) for this pair of p_0 and p_1. So if we know enough prior information about S, we are able to design $\mathcal{F}$ to include $\log \frac{p_1}{p_0}$ for all $p_1 \in S$. This improves the finite sample performance of the Hoeffding test without compromising asymptotic optimality. Moreover, the better we know about the possible alternate distributions, the further we can lower the dimensionality of $\mathcal{F}$. Thus, the more reliable the test statistic becomes. We refer readers to the simulation results in [2] for more information.

2.3 Countably Infinite Alphabet

The previous section focuses on the case that $|\mathcal{A}| < \infty$. We prove a new result on the case of countably infinite alphabet in this section.

We know from the asymptotic analysis of the Hoeffding test that the greater the difference between the alternate distribution and the null distribution, the larger the missed-detection error exponent is, given a certain false alarm error exponent. In other words, it is easier to detect the alternate distribution if it is a lot different from the null distribution. The difference between the alternate distribution and the null distribution is measured by the relative entropy between those two. Given any prescribed false alarm error exponent λ, the parameter of the twisted distribution can be determined accordingly. Then the missed-detection error exponent is determined by the relative entropy between the twisted distribution and the alternate distribution p_1. Apparently, the missed-detection error exponent can be zero for some choices of the alternate distribution p_1. This means that the Hoeffding test is not uniformly exponentially consistent if p_1 can be any distribution other than p_0. However, if we restrict our attention to the following alternate distributions,

$$S = \{p_1 | D(p_1 \| p_0) \geqslant \epsilon\} \tag{2.35}$$

where $\epsilon > \lambda$, the worst-case missed-detection error exponent of the Hoeffding test is strictly bounded away from zero.

Theorem 2.3 (Uniformly Exponential Consistency of the Hoeffding Test with $|\mathcal{A}| < \infty$). *Assume p_0 has full support over $\mathcal{A}$. Consider the Hoeffding test with threshold λ and the class of alternate distributions $S = \{p_1 | D(p_1 \| p_0) \geqslant \epsilon\}$ with $\epsilon > \lambda$, then the worst-case* *missed-detection error exponent of the Hoeffding test over S is strictly bounded away from zero,*

$$\inf_{p_1 \in S}\{\liminf_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_M(\delta_n, p_1)\} > 0 \tag{2.36}$$

Proof. We begin by proving the set S is closed under the assumption that p_0 has full support over $\mathcal{A}$. Let $q_n \in$ S be any sequence of distributions that converges in L_1. Let q^* be the limit of q_n. We need to prove that $q^* \in S$. Consider any $i \in \mathcal{A}$ such that $q^*(i) \neq 0$.

$$\begin{aligned} q^*(i)\log\frac{q^*(i)}{p_0(i)} &= q_n(i)\log\frac{q_n(i)}{p_0(i)} + (q^*(i) - q_n(i))\log\frac{q_n(i)}{q_0(i)} \\ &\quad + q_n(i)\log\frac{q^*(i)}{q_n(i)} + (q^*(i) - q_n(i))\log\frac{q^*(i)}{q_n(i)} \end{aligned}$$

The last three terms on the right all converges to zero since $q_n \to q^*$ in L_1 and $q^*(i) \neq 0$. For i that $q^*(i) = 0$

$$\lim_{n\to\infty} q_n(i)\log\frac{q_n(i)}{p_0(i)} = q^*(i)\log\frac{q^*(i)}{p_0(i)} = 0$$

since p_0 has full support. Sum $q^*(i)\log\frac{q^*(i)}{p_0(i)}$ over all i, there is $\lim_{n\to\infty}\mathrm{D}(q_n\|p_0) = \mathrm{D}(q^*\|p_0)$.

Then we prove that $\inf_{q\in S}\{\inf_{p\in\mathcal{Q}_\lambda(p_0)}\mathrm{D}(p\|q)\ \} > 0$. Suppose it is not true, then there is a sequence of (q_n, p_n), $q_n \in S$ and $p_n \in \mathcal{Q}_\lambda(p_0)$ such that

$$\lim_{n\to\infty}\mathrm{D}(p_n\|q_n) = 0$$

S is closed and bounded as proved before. So there is a subsequence q_{n_k} of q_n and $q_{n_k} \to q_0$ in L_1.

$$\|p_{n_k} - q_0\|_{L_1} \leqslant \|p_{n_k} - q_{n_k}\|_{L_1} + \|q_{n_k} - q_0\|$$

The two terms on the right all converge to zero since $\lim_{n\to\infty}\mathrm{D}(p_n\|q_n) = 0$ and $q_{n_k} \to q_0$ in L_1. As a consequence, $\|p_{n_k} - q_0\|_{L_1} \to 0$. This contradicts the fact that $\mathcal{Q}_\lambda(p_0)$ is compact since p_{n_k} converges to $q_0 \notin \mathcal{Q}_\lambda(p_0)$. By Sanov's theorem, $\inf_{q\in S}\{\liminf_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_M(\delta_n, q)\} = \inf_{q\in S}\{\inf_{p\in\mathcal{Q}_\lambda(p_0)}\mathrm{D}(p\|q)\}$. The theorem follows directly. □

We shall see that the above theorem falls apart when $\mathcal{A}$ is countably infinite. And that we need to better model the class of alternate distributions, if we would like to have any worst-case performance guarantee under the alternate distribution.

From now on, we assume that $\mathcal{A} = \{1, 2, ...\}$ is countably infinite. We are still considering the same universal hypothesis testing problem as in (2.6) except for that p_0 and p_1 take values in countably infinite alphabet and S is defined in (2.35). We have proved that S is a compact set when $\mathcal{A}$ is finite in size, and that the worst-case missed-detection error

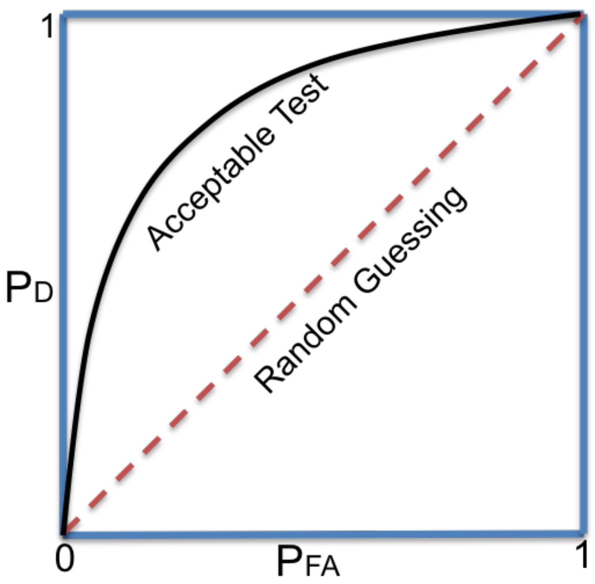

Figure 2.1: Receiver Operation Characteristic, Acceptable Test and Coin Flipping

exponent of the Hoeffding test over S is strictly bounded away from zero. Interestingly, this falls apart when the alphabet is countably infinite. The next theorem shows that the worst-case missed-detection error probability of any test is arbitrarily bad over the set of alternate distributions that are arbitrarily far away from p_0. This is due to the fact that S is not closed when $\mathcal{A}$ is countably infinite. The result asserts that, in order to have any worst-case performance guarantee of the Hoeffding test when $\mathcal{A}$ is countably infinite, we need a model of the alternate distribution which provides more information than S does.

Theorem 2.4 (Worst-Case Performance of Any Test with $|\mathcal{A}| = \infty$)**.** *Assume p_0 has full support over $\mathcal{A}$ and $\mathcal{A} = \{1, 2, \cdots\}$ is countably infinite. Consider the worst-case performance of an arbitrary test sequence $\delta = \{\delta_n : n = 1, 2, ...\}$ over the class of alternate distributions S. Then the performance of δ is arbitrarily bad in the sense that*

$$\inf_{p_1 \in S} \mathrm{P}_D(\delta_n, p_1) \leq \mathrm{P}_F(\delta_n, p_0) \quad \text{for all } n \tag{2.37}$$

where $\mathrm{P}_D(\delta_n, p_1) = 1 - \mathrm{P}_M(\delta_n, p_1)$ is the probability of detection.

Figure 2.1 shows what we mean by arbitrarily bad performance. The solid curve is the receiver operation characteristic (ROC) curve for an acceptable test. The concavity follows from the fact that we can improve any non-concave ROC curve by using randomized decision rules. The dotted curve corresponds to the test done by blindly flipping coins. Given a false alarm probability α, we can achieve $\mathrm{P}_F = \mathrm{P}_D$ simply by flipping biased coins with the probability of heads α. Moreover, if a test gives $\mathrm{P}_D \leqslant \mathrm{P}_F$, it can be improved simply by coin flipping. So the ROC curve of any acceptable test should be concave and above the dotted line. We call the performance of a test arbitrarily bad if $\mathrm{P}_F = \mathrm{P}_D$ because it can be replaced with simple coin flipping.

We begin the proof by constructing a sequence of distributions that converges to p_0 pointwisely and in L_1. For any sequence $\alpha_n \to 0$, construct a sequence of distribution p_n

$$p_n(i) \triangleq (1-\alpha_n)p_0(i) + \mathbb{I}(i=n)\alpha_n. \tag{2.38}$$

The next two lemmas show that we are able to choose α_n carefully to make $p_n \to p_0$ in L_1, and $p_n \in S$ for all n. We can do so simply because the relative entropy between any two probability measure dominates the L_1 distance.

Lemma 2.1. *Assume p satisfies $\sum_{n=1}^{\infty} p(n) = 1$ and $p(n) \neq 0$ for any $n \in \mathcal{A}$. There exists a sequence of $\alpha_n \to 0$ that satisfies $\log(\frac{\alpha_n}{p(n)}) \geqslant \frac{\delta}{\alpha_n}$ for all n and any $\delta > 0$.*

Proof. Let $\alpha_n = \frac{1}{\log\log(\frac{1}{p(n)})}$. It is obvious that $\alpha_n \to 0$ since $p(n) \to 0$.

$$\begin{aligned}\log(\frac{\alpha_n}{p(n)}) &= \log(\frac{1}{p(n)\log\log(\frac{1}{p(n)})}) \\ &= \frac{1}{2}log(\frac{1}{p(n)}) + \log(\frac{(\frac{1}{p(n)})^{\frac{1}{2}}}{\log\log(\frac{1}{p(n)})}) \\ &> \frac{1}{2}\log(\frac{1}{p(n)}) \quad \text{when } n \text{ large enough}\end{aligned} \tag{2.39}$$

The right hand side is

$$\frac{\delta}{\alpha_n} = \delta\log\log(\frac{1}{p(n)}) \tag{2.40}$$

Note that $\frac{1}{2}\log(\frac{1}{p(n)}) > \delta\log\log(\frac{1}{p(n)})$ eventually for any $\delta > 0$.
So let $\alpha_n = \frac{1}{\log\log(\frac{1}{p(n)})}$ for n large enough. Let α_n be the solution to $\log(\frac{\alpha_n}{p(n)}) = \frac{\delta}{\alpha_n}$ when n small. Such $\{\alpha_n\}$ satisfies that $\alpha_n \to 0$ and $\log(\frac{\alpha_n}{p(n)}) \geqslant \frac{\delta}{\alpha_n}$ by construction. □

The following lemma shows that every p_n constructed in (2.38) is at least ϵ away from p if the α_n in Lemma 2.1 is adopted.

Lemma 2.2. *p_n is given in the construction above. Let $\alpha_n \to 0$ be given in Lemma 2.1. Then $D(p_n \| p) \geqslant \epsilon$ for all n large enough.*

Proof. Adopt the $\{\alpha_n\}$ given in Lemma 2.1. Apparently p_n is guaranteed to be a valid p.m.f. when n large enough since $\alpha_n = \frac{1}{\log\log\frac{1}{p(n)}} \to 0$. Now calculate the divergence between p_n

and p.

$$\begin{aligned}
D(p_n\|p) &= \sum_{i=1}^{\infty} p_n(i)\log\frac{p_n(i)}{p(i)} \\
&= \sum_{i=1}^{\infty} p_n(i)\log((1-\alpha_n)+\frac{\delta_n(i)\alpha_n}{p(i)}) \\
&= \sum_{i\neq n}^{\infty} p_n(i)\log(1-\alpha_n)+p_n(n)\log(1-\alpha_n+\frac{\alpha_n}{p(n)}) \\
&\geqslant \log(1-\alpha_n)+p_n(n)\log(1-\alpha_n+\frac{\alpha_n}{p(n)}) \text{ for } n \text{ large} \\
&\geqslant \log\frac{1}{2}+(p(n)-\alpha_n p(n)+\alpha_n)\log(1-\alpha_n+\frac{\alpha_n}{p(n)}) \text{ for } n \text{ large} \\
&\geqslant \log\frac{1}{2}+\alpha_n(1-p(n))\log(\frac{\alpha_n}{p(n)}) \text{ for } n \text{ large since } \alpha_n \gg p(n) \\
&\geqslant \log\frac{1}{2}+\frac{1}{2}\alpha_n\log\frac{\alpha_n}{p(n)} \text{ for } n \text{ large}
\end{aligned} \tag{2.41}$$

Now note that if we let $\delta = \frac{\epsilon-\log\frac{1}{2}}{\frac{1}{2}}$ in Lemma 2.1, we get $\log\frac{\alpha_n}{p(n)} \geqslant \frac{\epsilon-\log\frac{1}{2}}{\frac{1}{2}\alpha_n}$. Combine this and the above calculation, we get $D(p_n\|p) \geqslant \epsilon$ when n is large enough with the choice $\alpha_n = \frac{1}{\log\log\frac{1}{p(n)}}$. □

Before proceeding to prove the theorem, we first introduce coupling on two random variables X and Y in the following way

$$Y = \begin{cases} X & \textit{with prob. } (1-\alpha_n) \\ n & \textit{with prob. } \alpha_n \end{cases} \tag{2.42}$$

Now we begin to prove Theorem 2.4. The proof techniques are similar to the techniques used in [10] and [11].

Proof. We begin by constructing two random variables X and Y on the same probability space. X has marginal p_0 and Y has marginal p_n.

Then couple X and Y in the above way. And repeat the coupling k times to get X_1^k i.i.d. with marginal p_0 and Y_1^k i.i.d. with marginal p_n.

For any test sequence $\delta = \{\delta_n : n = 1, 2, ...\}$, the following holds:

$$\begin{aligned} \Pr\{\delta_k(X_1^k) = \delta_k(Y_1^k)\} &\geqslant \Pr\{X_1^k = Y_1^k\} \\ &\geqslant \Pr\{\textit{symbol } n \textit{ does not appear in either } X_1^k \textit{ or } Y_1^k\} \\ &= [1 - ((1-\alpha_n)p_0(n) + \alpha_n)]^k \\ &\geqslant (1 - (p_0(n) + \alpha_n))^k \to 1 \qquad \text{as } n \to \infty \end{aligned} \tag{2.43}$$

Now prove that the performance of any test can be arbitrarily bad.

$$\mathrm{P}_F(\delta_k, p_0) = \Pr(\delta_k(X_1^k) = 1) \tag{2.44}$$

The probability of missed-detection for p_n with k samples can be calculated as follows

$$\begin{aligned} \mathrm{P}_M(\delta_k, p_n) &= \Pr\{\delta_k(Y_1^k) = 0\} \\ &\geqslant \Pr\{\delta(Y_1^k) = 0, X_1^k = Y_1^k\} \\ &= \Pr\{\delta_k(X_1^k) = 0, X_1^k = Y_1^k\} \\ &\geqslant (1 - (p_0(n) + \alpha_n))^k - \mathrm{P}_F(\delta_k, p_0) \end{aligned} \tag{2.45}$$

For any k, let $n \to \infty$ we have

$$\sup_{p_1 \in S} \Pr\{\delta_k(Y_1^k) = 0\} \geqslant 1 - \mathrm{P}_F(\delta_k, p_0) \tag{2.46}$$

and

$$\inf_{p_1 \in S} \mathrm{P}_D(\delta_k, p_1) \leqslant \mathrm{P}_F(\delta_k, p_0) \tag{2.47}$$

where $\mathrm{P}_D(\delta_k, p_1) = 1 - \mathrm{P}_M(\delta_k, p_1)$ is the probability of detection under p_1. So the performance of any test over the set S is arbitrarily bad in the sense that the corresponding ROC curve is a straight line. □

Note that the above theorem works for any choice of positive ϵ, any arbitrary test sequence, and any n. Thus we have no guarantee on the performance, even if we restrict our attention to the alternate distributions that are far from the null distribution in terms of relative entropy. This justifies the necessity that we need to incorporate more prior information of the alternate distribution. Or in other words, we need a better modeling of the alternate distributions.

CHAPTER 3

CLASSIFICATION

3.1 Definition and Background

In this chapter, we focus on the design of the classification method. We show that utilizing both training sequences improves the finite sample performance even though asymptotic optimality is achieved with one training sequence. We prove Stein's lemma for classification using a new classification function. Our results also shed light on how the classification errors depend on the relative size of the training and test data.

A full probabilistic model of the system can be too costly to obtain in some cases. The modeling may not even be possible for some intricate and large-scaled systems. The problem of classification arises naturally in those applications. The goal of classification is to identify which of the M classes a new observation belongs to. A class is characterized by the probabilistic model from which the observations are generated. The probabilistic models are not known but can be learned in various ways. In supervised classification problems, previously classified samples are available to the classifier and future observations are classified based on the information learned from these previously classified samples. In unsupervised classification problems, information about each class is learned as samples are classified. In this work, we focus on the problem of supervised classification with $M = 2$. Without any knowledge of the marginals, a classification rule can only depend on the previously classified samples. We further assume that the observations are i.i.d. with different marginals from different classes. Anomaly detection is one example of binary classification problems. There are two states of a system, normal and abnormal. Observations generated by the normal behavior follow a different distribution from those generated from any abnormal behavior. For the cases where $M > 2$, the readers are referred to [3] for more information.

From now on, we refer to the previously classified sample sequences as the training data and the samples to be classified as the test data. We assume that all the training data and test data are i.i.d. with a marginal distribution p_1 if they are from class one and p_2 from class two. p_1 and p_2 are discrete over the alphabet $\mathcal{A}$ and $|\mathcal{A}| < \infty$. Without loss of generality, assume $\mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$. Let $\{T_1\}_1^N$ be the training data from class one and $\{T_2\}_1^N$ class two. Let X_1^n be the test data with marginal p which is either p_1 or p_2. We use N to denote the size of the training data and n for the size of the test data. We use q to denote the empirical distributions. For instance, $q_{X_1^n}$ is the empirical distribution constructed from the test sequence. A classification rule $\delta = \mathbb{I}\{h(\{T_1\}_1^N, \{T_2\}_1^N, X_1^n) > 0\} + 1 : \mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^n \to \{1, 2\}$ maps the training and test data to a decision on the two classes. $h(\{T_1\}_1^N, \{T_2\}_1^N, X_1^n)$ is the classification function. A classification rule essentially divides the space of training and testing data $\mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^n$ into two decision regions Λ_1 and Λ_2 and $\Lambda_1 \cup \Lambda_2 = \mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^n$. The error probability under each of the classes is defined similar to (2.2) and (2.3).

$$\mathrm{P}_{\mathrm{err}1} \triangleq \mathrm{P}(\delta = 2 | X_1^n \sim p_1) \tag{3.1}$$

$$\mathrm{P}_{\mathrm{err}2} \triangleq \mathrm{P}(\delta = 1 | X_1^n \sim p_2) \tag{3.2}$$

Intuitively, there are two sources of error in supervised classification problems. The misclassification error comes from both the false modeling of the classes and the classification itself. If we have unlimited training data under both classes, the problem becomes a binary hypothesis testing problem with fully known p_1 and p_1. Then it is known that the error is exponentially small with respect to the size of the test data. If the training data are limited, we do not have full knowledge of the classes. The false modeling is another source of error because future observations are classified based on an inaccurate model. It is not clear at this point how the classification errors decay. In other words, we do not know how to normalize the errors to get the error exponent. As we can see from the optimality criterion introduced next, the errors are normalized by $\frac{1}{m}$ for some $m \to \infty$ as (n, N) tend to infinity. It will be clear what m is once we study the asymptotic behavior of the errors. We shall also see that it is the relative size of (n, N) that determines the rate at which the errors tend to zero. So we postpone the definition of error exponent until later.

We adopt a similar definition of asymptotic optimality as the modified Neyman-Pearson criterion:

> Among all classification rules $\Delta = \{\delta | \delta = \{\delta_n, n = 1, 2, ...\}\}$ that do not depend on the unknown p_1 and p_2 and with the error exponent under class one
>
> $$\liminf_{n\to\infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err}1}(\delta_n, p_1) \geqslant \lambda$$
>
> select a sequence that maximizes error exponent under class two
>
> $$\liminf_{n\to\infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err}2}(\delta_n, p_2) \text{ for all } p_2 \in \mathbb{P}(\mathcal{A})$$

where $m = m(n, N)$ is a function of (n, N) and tends to infinity as (n, N) tend to infinity. We will specify m later as we know more about the behavior of error probabilities in classification.

This chapter is organized as follows. We first prove a theorem that characterizes asymptotically optimal classification functions. Then we study two classification functions under which the error probabilities of the two classes are exponentially small. We will see that the error probability depends on both n and N. These two functions depend only on the training samples from class one. The first classification function is inspired by [12]. The second classification function is proposed by [3] and is proved to be asymptotically optimal under the above criterion. This result is very similar to the case in hypothesis testing problems where the knowledge of only one distribution is needed to achieve asymptotic optimality. Then we propose a different classification function which resembles the test statistic of the loglikelihood test. Our method utilizes both training sequences. We prove Stein's lemma for classification using the new classification function. In the end, we present simulation results which show that the classification function we proposed outperforms the one in [3] when the number of samples are limited. This justifies using both training sequences for additional information about the two classes.

3.2 Asymptotically Optimal Classification Rules

3.2.1 Characteristic of the Asymptotically Optimal Classification Rules

We start by proving a theorem that characterizes asymptotically optimal classification functions. Not surprisingly, the asymptotically optimal classification function depends on the training and test data only through their types. The proof techniques are similar to the Lemma 1 in [13].

Theorem 3.1 (Characteristic of Asymptotically Optimal Classification Functions)**.** *Any asymptotically optimal classification function depends on the training and test data only through their types.*

Proof. First prove that all classification functions can be replaced by the ones that only depend on X_1^n through its type without compromising its asymptotic performance.

Let $\Lambda = \Lambda_1 \cup \Lambda_2$ be the decision region specified by any classification function. Let $\Lambda_1(t_1, t_2), \Lambda_2(t_1, t_2) \in \mathcal{A}^n$ be the decision regions conditioned on that the training data $\{T_1\}_1^N = t_1$ and $\{T_2\}_1^n = t_2$. Let $X_\mu = \{X_1^n | q_X = \mu\}$ be the set of test data that has empirical distribution μ. Let $B_\mu(t_1, t_2) = X_\mu \cap \Lambda_1(t_1, t_2)$ be the part of X_μ that is included in decision region one. Let $C_\mu(t_1,t_2) = X_\mu \cap \Lambda_2(t_1,t_2)$ be the rest of X_μ which is in decision region two. The basic idea is to compare the size of these two sets and construct new decision regions based on majority vote. Define

$$\Omega_\mu(t_1,t_2) = \begin{cases} X_\mu & \text{if } |B_\mu| \geq |C_\mu|; \\ \varnothing & \text{else.} \end{cases}$$

The new decision region is defined as $\Omega_1(t_1,t_2) = \cup\Omega_\mu(t_1,t_2)$ and $\Omega_2(t_1,t_2) = \mathcal{A}^n/\Omega_1$. It is easy to see that

$$p_1(X_1^n \in \Omega_2(t_1,t_2)|X_1^n \in X_\mu) \leq 2p_1(X_1^n \in \Lambda_2(t_1,t_2)|X_1^n \in X_\mu)$$

and

$$p_2(X_1^n \in \Omega_1(t_1,t_2)|X_1^n \in X_\mu) \leq 2p_2(X_1^n \in \Lambda_1(t_1,t_2)|X_1^n \in X_\mu)$$

Do this for every pair of t_1 and t_2. And construct the new decision region $\Omega = \Omega_1 \cup \Omega_2$. The error under p_1 with decision region Ω is

$$\begin{aligned} \mathrm{P}_{\text{err1}}(\Omega) &= \sum_{t_1,t_2} p_1(t_1)p_2(t_2)\sum_{X_\mu} p_1(x_1^n \in \Omega_2(t_1,t_2)|x_1^n \in X_\mu)p_1(x_1^n \in X_\mu) \\ &\leq 2\sum_{t_1,t_2} p_1(t_1)p_2(t_2)\sum_{X_\mu} p_1(x_1^n \in \Lambda_1(t_1,t_2)|x_1^n \in X_\mu)p_1(x_1^n \in X_\mu) \\ &= 2\mathrm{P}_{\text{err1}}(\Lambda) \end{aligned}$$

where $\mathrm{P}_{\text{err1}}(\Omega)$ is the error under class one with decision regions Ω and $\mathrm{P}_{\text{err1}}(\Lambda)$ is the error under class one with decision region Λ. With the same argument,

$$\mathrm{P}_{\text{err2}}(\Omega) \leq 2\mathrm{P}_{\text{err2}}(\Lambda)$$

If we calculate the error exponent,

$$\lim_{m\to\infty} -\frac{1}{m}\log \mathrm{P}_{\text{err1}}(\Omega) \geq \lim_{m\to\infty} -\frac{1}{m}\log \mathrm{P}_{\text{err1}}(\Lambda)$$

and

$$\lim_{m\to\infty} -\frac{1}{m}\log \mathrm{P}_{\text{err2}}(\Omega) \geq \lim_{m\to\infty} -\frac{1}{m}\log \mathrm{P}_{\text{err2}}(\Lambda)$$

So the constructed new decision region Ω which depends on X_1^n only through its type has no worse error exponents under both classes.

The second part proves that all decision regions like Ω can be replaced by decision regions that depend on $\{T_1\}_1^N$ and $\{T_2\}_1^N$ only through their types. $\Lambda_1(t_1, t_2)$ is the decision region conditional on the training data $\{T_1\}_1^N = t_1$ and $\{T_2\}_1^N = t_2$. Let $T_{\nu_1,\nu_2} = \{t_1, t_2 | q_{t_1} = \nu_1, q_{t_2} = \nu_2\}$ be the set of training sequence pairs which has empirical distributions (ν_1, ν_2). Construct a new decision region Θ as follows. Let

$$B(X_\mu, T_{\nu_1,\nu_2}) = \{t_1, t_2 \in T_{\nu_1,\nu_2} | X_\mu \in \Omega_1(t_1, t_2)\}$$

and

$$C(X_\mu, T_{\nu_1,\nu_2}) = \{t_1, t_2 \in T_{\nu_1,\nu_2} | X_\mu \in \Omega_2(t_1, t_2)\}$$

Let

$$\Theta_\mu(T_{\nu_1,\nu_2}) = \begin{cases} X_\mu & \text{if } |B(X_\mu, T_{\nu_1,\nu_2})| \geq |C(X_\mu, T_{\nu_1,\nu_2})|; \\ \varnothing & \text{else.} \end{cases}$$

Let $\Theta_1(T_{\nu_1,\nu_2}) = \cup_\mu \Theta_\mu(T_{\nu_1,\nu_2})$ and $\Theta_2(T_{\nu_1,\nu_2}) = \mathcal{A}^n / \Theta_1(T_{\nu_1,\nu_2})$. Note that Θ only depends on the training and test data through their types.

$$\begin{aligned} \mathrm{P}_{\mathrm{err1}}(\Theta) &= \sum_{T_{\nu_1,\nu_2}} \sum_{(t_1,t_2) \in T_{\nu_1,\nu_2}} \sum_{T_\mu \in \Theta_2(T_{\nu_1,\nu_2})} p_1(t_1) p_2(t_2) p_1(T_\mu) \\ &= \sum_{T_{\nu_1,\nu_2}} \sum_{T_\mu \in \Theta_2(T_{\nu_1,\nu_2})} p_1(T_\mu) \sum_{t_1,t_2 \in T_{\nu_1,\nu_2}} p_1(t_1) p_2(t_2) \\ &= \sum_{T_{\nu_1,\nu_2}} \sum_{T_\mu \in \Theta_2(T_\nu)} p_1(T_\mu) (\sum_{(t_1,t_2) \in T_\nu, T_\mu \in \Omega_2(t_1,t_2)} p_1(t_1) p_2(t_2) \\ &+ \sum_{(t_1,t_2) \in T_{\nu_1,\nu_2}, T_\mu \in \Omega_1(t_1,t_2)} p_1(t_1)) p_2(t_2) \end{aligned} \tag{3.3}$$

Note that

$$\sum_{t_1,t_2 \in T_{\nu_1,\nu_2}, X_\mu \in \Omega_1(t_1,t_2)} p_1(t_1) p_2(t_2) = p_1(B(X_\mu, T_{\nu_1,\nu_2}))$$

$$\sum_{t_1,t_2 \in T_{\nu_1,\nu_2}, X_\mu \in \Omega_2(t_1,t_2)} p_1(t_1) p_2(t_2) = p_1(C(X_\mu, T_{\nu_1,\nu_2}))$$

Note that $X_\mu \in \Theta_2(T_{\nu_1,\nu_2})$ implies that $|B(T_\mu, T_{\nu_1,\nu_2})| \leq |C(T_\mu, T_{\nu_1,\nu_2})|$. Thus,

$$p_1(B(T_\mu, T_{\nu_1,\nu_2})) + p_1(B(T_\mu, T_{\nu_1,\nu_2})) \leq 2 p_1(C(T_\mu, T_{\nu_1,\nu_2})$$

So

$$\begin{aligned}
\mathrm{P}_{\mathrm{err1}}(\Theta) &\leq 2 \sum_{T_{\nu_1,\nu 2}} \sum_{X_\mu \in \Theta_2(T_{\nu_1,\nu_2})} p_1(X_\mu) \sum_{t_1,t_2 \in T_{\nu_1,\nu_2}, X_\mu \in \Omega_2(t_1,t_2)} p_1(t_1) p_2(t_2) \\
&\leq 2 \sum_{T_{\nu_1,\nu_2}} \sum_{t_1 t_2 \in T_{\nu_1,\nu_2}} \sum_{X_\mu \in \Omega_2(t_1,t_2)} p_1(X_\mu) p_1(t_1) p_2(t_2) \\
&= 2 \sum_{T_{\nu_1,\nu_2}} \sum_{t_1,t_2 \in T_{\nu_1,\nu_2}} p_1(t_1) p_2(t_2) \sum_{X_\mu \in \Omega_2(t_1,t_2)} p_1(X_\mu) \\
&= 2\mathrm{P}_{\mathrm{err1}}|(\Omega)
\end{aligned}$$

By the same argument, $\mathrm{P}_{\mathrm{err2}}(\Theta) \leq 2\mathrm{P}_{\mathrm{err2}}(\Omega)$. Asymptotically,

$$\lim_{m \to \infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err1}}(\Theta) \geq \lim_{m \to \infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err1}}(\Lambda)$$

and

$$\lim_{m \to \infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err2}}(\Theta) \geq \lim_{m \to \infty} -\frac{1}{m} \log \mathrm{P}_{\mathrm{err2}}(\Lambda)$$

where Λ is any decision region and Θ is constructed from Λ which depends on the training and test sequences only through their types. So the asymptotically optimal classification function depends on the data only through their types. □

3.2.2 Classification Functions Based on Total Variation Distance

According to the last theorem, we can focus our attention on classification functions that depend on the training and test data through their types. We first study a classification function that is similar in nature to the test statistic in [12]. The classification function takes the total variation distance between the empirical distribution of the training data under class one and that of the test data.

$$h(t_1, t_2, x) = d(q_{t_1}, q_x) \tag{3.4}$$

where d denotes the total variation distance. For any P, Q defined over $(\Omega, \mathcal{F})$, the total variation distance is defined as

$$d(\mathrm{P}, \mathrm{Q}) \triangleq \sup_{A \in \mathcal{F}} |\mathrm{P}(A) - \mathrm{Q}(A)|. \tag{3.5}$$

For discrete $\mathcal{A}$, the total variation distance is simply

$$d(P,Q) = \frac{1}{2}\sum_{i\in\mathcal{A}} |P(i) - Q(i)|. \tag{3.6}$$

The decision $\delta(t_1, t_2, x)$ is made by comparing $h(t_1, t_2, x)$ with a threshold r

$$\delta(t_1, t_2, x) = \begin{cases} \text{class one} & \text{if } h(t_1, t_2, x) - r < 0; \\ \text{class two} & \text{else} \end{cases} \tag{3.7}$$

Note that this classification function depends only on the training data from class one.

We extend the notion of consistency and uniform consistency to classification problems. We say a classification method is consistent if the error probabilities under all classes converge to zero as the size of the training and test data increases. In examples like anomaly detection, the normal behavior of a system is unique but abnormal behaviors can be anything in the set S. It is desirable that a classification method is consistent against all the possible abnormal behaviors in S. We say a classification method is uniformly consistent against the set S if the misclassification errors converge to zero under any $p_2 \in S$. Uniform exponential consistency can be defined similarly.

Consider the set of all possible marginals p_2 under class two which is at least δ away from the marginal p_1 under class one,

$$S \triangleq \{p_2 | d(p_1, p_2) > \delta\} \tag{3.8}$$

where the choice of δ will be specified later. We will prove that the classification rule based on total variation distance is uniform exponentially consistent over S.

Theorem 3.2 (Uniform Exponential Consistency of the Classifier Based on Total Variation Distance)**.** *The classification function $h(t_1, t_2, x) = d(q_{t_1}, q_x)$ with threshold r is uniform exponentially consistent against the set of marginals $S \triangleq \{p_2 | d(p_1, p_2) > \delta\}$ for any $\delta > r$.*

Proof. Choose any r_1 such that $\sqrt{\frac{1}{2}r_1} < r < \delta$ and $(\sqrt{\frac{1}{2}r_1} + r)^2 < \delta$. Let C^{r_1} be the set of training sequences from class one that have empirical distribution close to the true marginal p_1 in terms of relative entropy,

$$C^{r_1} = \{t_1 | D(q_{t_1} \,||\, p_1) \le r_1\}$$

and define the decision region corresponding to class one

$$A^r = \{x | d(q_x, q_{t_1}) \le r\}$$

We first calculate a few bounds on the total variation distance between any two sequences in the above two sets. If $q_{t_1} \in C^{r_1}$, then by the definition of C^{r_1} and Pinsker's inequality, there is $2d^2(q_{t_1}, p_1) \leq D(q_{t_1} \,||\, p_1) \leq r_1$. If in addition that $X_1^n \in (A^r)^c$ and $r > \sqrt{\frac{1}{2} r_1}$, then by the same argument,

$$\begin{aligned} D(q_x \,||\, p_1) &\geq 2d^2(q_x, p_1) \\ &\geq 2(d(q_x, q_{t_1}) - d(q_{t_1}, p_1))^2 \\ &\geq 2(r - \sqrt{\frac{1}{2} r_1})^2 \end{aligned}$$

If $q_{t_1} \in C^{r_1}$ and $q_x \in A^r$, for any $p_2 \in S$,

$$\begin{aligned} D(q_x \,||\, p_2) &\geq 2d^2(q_x, p_2) \\ &\geq 2(d(p_2, p_1) - d(q_x, p_2))^2 \\ &\geq 2(\delta - (r + \sqrt{\frac{1}{2} r_1}))^2 \end{aligned}$$

provided $\delta > (r + \sqrt{\frac{1}{2} r_1})$. The inequality follows from the fact that

$$d(q_x, p_1) \leq d(q_x, q_{t_1}) + d(q_{t_1}, p_1) \leq r + \sqrt{\frac{1}{2} r_1}$$

and $d(p_2, p_1) \geq \delta$.

Now we proceed to calculate the missed classification error of class one,

$$\begin{aligned}
p_1(A^r) &= p_1(A^r, T_1 \in C^{r_1}) + p_1(A^r, T_1 \in (C^{r_1})^c) \\
&\geq p_1(A^r, T_1 \in C^{r_1}) \\
&= p_1(A^r | T_1 \in C^{r_1}) p_1(T_1 \in C^{r_1}) \\
&= (1 - p_1((A^r)^c | T_1 \in C^{r_1})) p_1(T_1 \in C^{r_1}) \\
&= (1 - \frac{p_1((A^r)^c, T_1 \in C^{r_1})}{p_1(T_1 \in C^{r_1})}) p_1(T_1 \in C^{r_1}) \\
&\geq (1 - \frac{p_1(D(q_X \,||\, p_0) \geq 2(r_0 - \sqrt{\frac{1}{2}r_1})^2)}{1 - p_1(T_1 \in (C^{r_1})^c)}) p_1(T_1 \in C^{r_1}) \\
&\geq (1 - \frac{\binom{n+|\mathcal{A}|}{|\mathcal{A}|} e^{-n2(r-\sqrt{\frac{1}{2}r^1})^2}}{1 - \binom{N+|\mathcal{A}|}{|\mathcal{A}|} e^{-Nr_1}})(1 - \binom{N+|\mathcal{A}|}{|\mathcal{A}|} e^{-Nr_1}) \\
&= 1 - \binom{N+|A|}{|A|} e^{-Nr_1} - \binom{n+|A|}{|A|} e^{-n2(r-\sqrt{\frac{1}{2}r^1})^2}
\end{aligned}$$

So the missed classification error under class one is upper bounded by

$$\mathrm{P}_{\mathrm{err1}} \leq \binom{N+|\mathcal{A}|}{|\mathcal{A}|} e^{-Nr_1} + \binom{n+|\mathcal{A}|}{|\mathcal{A}|} e^{-n2(r-\sqrt{\frac{1}{2}r_1})^2}$$

Use the following bound on $\binom{n+m}{m}$,

$$\binom{m+n}{m} \leq e^{(n+m)H(\frac{m}{n+m})}$$

where $H(\frac{m}{m+n})$ is the entropy of binomial distribution $(\frac{m}{m+n}, \frac{n}{m+n})$. Plug the above bound in $\mathrm{P}_{\mathrm{err1}}$; finally we get that

$$\begin{aligned}
\mathrm{P}_{\mathrm{err1}} &\leq \exp\{-N[r_1 - (1 + \frac{|\mathcal{A}|}{N})H(\frac{|\mathcal{A}|}{N+|\mathcal{A}|})]\} \\
&+ \exp\{-n[2(r - \sqrt{\frac{1}{2}r_1})^2 - (1 + \frac{|\mathcal{A}|}{n})H(\frac{|\mathcal{A}|}{|\mathcal{A}|+n})]\}
\end{aligned}$$

It is easy to see that $(1 + \frac{|\mathcal{A}|}{N})H(\frac{|\mathcal{A}|}{N+|\mathcal{A}|}) \to 0$ as $N \to \infty$. And $(1 + \frac{|\mathcal{A}|}{n})H(\frac{|\mathcal{A}|}{|\mathcal{A}|+n}) \to 0$ as

$n \to \infty$. So the missed classification error under class one decays exponentially fast with respect to the term among (N, n) that grows slower.

The error under any $p_2 \in S$ can be bounded similarly and finally we get

$$\begin{aligned} \mathrm{P}_{\mathrm{err2}} &\leq \binom{N+|\mathcal{A}|}{|\mathcal{A}|}\exp\{-Nr_1\} + \binom{n+|\mathcal{A}|}{|\mathcal{A}|}\exp\{-n2(\delta-(r+\sqrt{\frac{1}{2}r_1})^2\} \\ &= \exp\{-N[r_1-(1+\frac{|\mathcal{A}|}{N})H(\frac{|\mathcal{A}|}{n+|\mathcal{A}|})]\} \\ &+ \exp\{-n[2(\delta-(r+\sqrt{\frac{1}{2}r_1})^2)-(1+\frac{|\mathcal{A}|}{n})H(\frac{|\mathcal{A}|}{n+|\mathcal{A}|})]\} \end{aligned}$$

Note that since the bound does not depend on a particular p_2 but only the fact that $p_2 \in S$, we have proved that the missed classification error is uniformly exponentially small over the set S. □

Recall the intuitive observation we made in the beginning of this chapter. We can see the two sources of error in the upper bound of $\mathrm{P}_{\mathrm{err1}}$ and $\mathrm{P}_{\mathrm{err2}}$. There are two terms in the expression. One of the terms is exponentially small in the size of the training data. The other is exponentially small in the size of the test data. It is the relative size of (n, N) that determines how the error decays. For example, if $N = o(n)$, the error is exponentially small with respect to the size of the training samples. If $n = o(N)$, the error is exponentially small with respect to the size of the test data. If N and n are of the same order, for instance $N = Kn$, the error exponent depends on r, r_1 and δ and the choice of the growth rate K. The best error exponent is achieved by choosing K that balances the two terms in the error. So if we would like the error probability to decay exponentially with the size of the test data, we need to have enough training samples to get a good understanding of the difference between the classes so the errors brought by the training samples are negligible compared to the error brought by the test samples. The relative size of the training data with respect to the test data seems to be a universal issue in classification problems. The same situation also appears in the classification function we study next.

Similarly to [12], if we look at a general probability space and continuous probability measures, the classification function can be modified to work in the new scenario. We could partition the space with pi_n and use the variational distance $d_{\pi_n}(q_x, q_{t_1})$ in the new classification function. Also note that if we let the size of the partition $|\pi_n|$ grow sub-linearly with respect to the smaller term among (N, n), the classification function still has uniform exponentially consistent performance over the set of alternatives $S = \{p|d_{\pi_n}(p_2, p_1) \geq \delta\}$. The difference is that we are including more and more alternative models in S as the sizes of training and test data grow. This result is not surprising in the sense that more samples provide more discerning power so more models of class two can be discerned from class one. More information about this issue can be found in [12].

3.2.3 Asymptotically Optimal Classification Rules

We have a good understanding of the asymptotic behavior of the errors in hypothesis testing problems. We know that there is a tradeoff between the error probabilities under the two hypotheses. It is also known that if a test is exponentially consistent then the errors decay exponentially fast with respect to the size of the test samples. Recall that we did not specify the normalization term m in the definition of asymptotic optimality because it is not clear how the errors decay. First, there are two sample sizes involved in the analysis. Does the error decay exponentially fast with respect to the size of the training sample or the test sample? What is the relation between the training sample size and test sample size in order to guarantee exponential decay? How fast can the error under class one go to zero under the constraint that the error under class two converge to zero? Intuitively, if the training sample size is too small to provide enough information about the difference between the two classes, the error brought by false modeling of the two classes would dominate. If the number of training samples goes to infinity, we would know the exact distribution p_1 and p_2, so the error should be exponentially small with respect to the size of the test samples. We show in this section that it is possible to make the error under class one decay exponentially fast with m for any choice of m. But if m is of higher order than the smaller term in (n, N), no classifier is consistent under class two. This result is consistent with the analysis of the classifier based on the total variation distance. We will make the above statement rigorous in the following theorems.

With some abuse of notation, let $\min(n, N)$ be the one between n and N that is of smaller order. And if $N = Kn$ for a constant K, we let $\min(n, N) = n$. The following theorem states that if we require that the error probability under class one is exponentially small with respect to m with $\frac{\min(n,N)}{m} \to 0$, the error probability under class two does not vanish. Thus in the optimality criterion (3.9), $m = min(n, N)$.

Theorem 3.3 (Necessary Condition for Consistency). *Assume p_1 and $p_2 \in \mathbb{P}(\mathcal{A})$ have full support over $\mathcal{A}$. Let $t_1 \sim p_1$, $t_2 \sim p_2$ be the training samples of length N. Let $\delta(t_1, t_2, x)$ be any classification rule that does not depend on p_1 and p_2. Let m be any sequence that $\frac{\min(n,N)}{m} \to 0$. For any $\lambda > 0$, if $\delta(t_1, t_2, x)$ satisfies,*

$$\mathrm{P}_{\mathrm{err1}}(\delta(t_1, t_2, x)) \leq 2^{-\lambda m}$$

Then for any $p_1 \in \mathrm{P}(\mathcal{A})$ *and any* $p_2 \neq p_1$

$$\lim_{m\to\infty} \mathrm{P}_{\mathrm{err2}}(h(t_1, t_2, x)) > 0$$

The proof of the theorem is very similar to the one of theorem 3 in [14] and we do not repeat it here. In application, we are usually interested in classification rules that are consistent under both classes. We have seen in the analysis of the classifier based on total variation distance that the errors are exponentially small with $\min(n, N)$. This theorem states that we cannot do better than this if we would like the classifier to be consistent under both classes. Now we can specify the choice of m in the optimality criterion.

> Among all classification rules $\Delta = \{\delta | \delta = \{\delta_n, n = 1, 2, ...\}\}$ that do not depend on the unknown p_1 and p_2 and with the error exponent under class one
>
> $$\liminf_{\min(n,N)\to\infty} -\frac{1}{\min(n, N)} \log \mathrm{P}_{\mathrm{err1}}(\delta_n, p_1) \geqslant \lambda \qquad (3.9)$$
>
> select a sequence that maximizes error exponent under class two
>
> $$\liminf_{\min\to\infty} -\frac{1}{\min} \log \mathrm{P}_{\mathrm{err2}}(\delta_n, p_2) \text{ for all } p_2 \in \mathbb{P}(\mathcal{A}) \qquad (3.10)$$

We are not able to prove if the classifier based on total variation distance is asymptotically optimal under this criterion. Note that the total variation distance and the test statistics of the Hoeffding test are of a different nature. And the fact that the Hoeffding test is asymptotically optimal motivates us to look for classification functions that are similar to the generalized likelihood ratio. This is exactly how the next classification function is formulated. The classification function we study next first appeared in [14] and was later clarified by [3]. The classification function is constructed by formulating the classification problem as a composite versus composite hypothesis testing problem. Thus GLRT can be applied to it. [14] also points out the connection between the classification function and universal data compression algorithms. The connection helps to avoid constructing empirical distributions from training and test data, which can be tedious if the underlying processes are not stationary and memoryless. The optimality result also extends to more general processes, i.e., all finite alphabet ergodic measures with a certain fading memory. This work focuses on i.i.d. models. So we would stick with constructing empirical distributions. The classification problem can be formulated as a composite versus composite hypothesis testing problem as follows:

- $H_1 :\quad \{T_1\}_1^N$ and X_1^n are i.i.d. with the same distribution p

- H_2 : $\{T_1\}_1^N$ and X_1^n are i.i.d. with different distribution p_1 and p_2 .

Apply the generalized likelihood ratio test to this composite versus composite hypothesis testing problem

$$h(t_1, t_2, x) = \log \frac{\sup_{p_1,p_2} p_1(x)p_2(t_1)}{\sup_p p(x, t_1)} \tag{3.11}$$

It is easy to show that the supremum in h is achieved by the empirical distributions $p_1^* = q_x$, $p_2^* = q_{t_1}$ and $p^* = q_{x,t_1}$ where $q_{x,t_1} = \frac{N}{N+n} q_{t_1} + \frac{n}{N+n} q_x$ is the empirical distribution of the concatenation of x and t_1. Plug in the maximizer, and we see that h can be equally written as

$$h(t_1, t_2, x) = (n + N)H(q_{x,t_1}) - nH(q_x) - NH(q_{t_1}) \tag{3.12}$$

where $H(p)$ is the entropy of distribution p.

Note that h can also be equally written as

$$h(t_1, t_2, x) = nD(q_x||q_{x,t_1}) + ND(q_{t_1}||q_{x,t_1}) \tag{3.13}$$

which is the sum of the total relative entropy between the training samples T_1 and (X, T_1) and the total relative entropy between X and (X, T). If X and T_1 have the same marginal, and if the size of N and n are large enough, $D(q_X||q_{X,T_1})$ and $D(q_{T_1}||q_{X,T_1})$ are both close to 0 in probability. If X and T_1 have different marginals, we would expect $D(q_{T_1}||q_{X,T_1})$ and $D(q_X||q_{X,T_1})$ converge to positive numbers in probability by the law of large numbers. The convergence obviously depends on the relative size of N and n since q_{X,T_1} is the empirical distribution of the concatenation of X and T_1. This is intuitively how the classification function has discerning power.

To be consistent with our definition of asymptotic optimality, we use a decision rule which is slightly different from the one in [3]. The decision $\delta(t_1, t_2, x)$ is made by comparing $\frac{1}{m} h(t_1, t_2, x) + \rho(n, M)$ with a threshold λ. $\rho(n, N) = o(1)$.

$$\delta(t_1, t_2, x) = \begin{cases} \text{class one} & \text{if } \frac{1}{m} h(t_1, t_2, x) + \rho(n) - \lambda < 0; \\ \text{class two} & \text{otherwise} \end{cases} \tag{3.14}$$

The following lemma states an upper bound for the probability of error under class one of (3.14).

Lemma 3.1 (Misclassification Error under Class One). *Consider the classification rule defined in* (3.14). *Assume that* $\frac{\log n}{m} \to 0$ *and* $\frac{\log N}{m} \to 0$, *then the probability of error under* *class one is upper bounded by*

$$-\frac{1}{m}\log \mathrm{P}_{\mathrm{err1}}(\delta(t_1,t_2,x)) \geq \lambda$$

Proof. Let $\Lambda_1 = \{x, t_1 | nD(q_x||q_{x,t_1}) + ND(q_{t_1}||q_{x,t_1}) - m\lambda < 0\}$ be the acceptance region of class one and $\Lambda_2 = \mathcal{A}^N \times \mathcal{A}^n/\Lambda_0$ be the acceptance region of class two. The error probability under class one is

$$\begin{aligned} \mathrm{P}_{\mathrm{err1}}(\Lambda) &= \sum_{(x,t_1)\in\Lambda_1} p_1(x)p_1(t_1) \\ &\leq \sum_{(x,t)\in\Lambda_1} q_{x,t_1}(x,t_1) \end{aligned}$$

Note that $q_{x,t_1}(x,t_1) = 2^{-(n+N)H(q_{x,t_1})}$. If $(x,t_1)\in\Lambda_1$, then $(n+N)H(q_{x,t_1}) - nH(q_x) - NH(q_{t_1}) + m\rho(n) - m\lambda \geq 0$. So $2^{-(n+N)H(q_{x,t_1})} \leq 2^{-nH(q_x)}2^{-NH(q_{t_1})}2^{-m(\lambda-\rho(n))}$. Plug this in $P_{err1}(\Lambda)$,

$$\begin{aligned} \mathrm{P}_{\mathrm{err1}}(\Lambda) &\leq \sum_{(x,t_1)\in\Lambda_1} 2^{-NH(q_{t_1})}2^{-nH(q_x)}2^{-m(\lambda-\rho(n))} \\ &\leq 2^{-n(\lambda-\rho(n))}\sum_{t_1\in\mathcal{A}^n} 2^{-NH(q_{t_1})}\sum_{x\in\mathcal{A}^n} 2^{-nH(q_x)} \end{aligned}$$

Let $T(p_1)$ be the set of all possible types constructed from x under p_1. For any $T\in T(p_1)$, let $q(T)$ be any empirical distribution associated with the particular type T. Let $|T|$ be the number of sequences that are of the same type T. We have

$$\sum_{x\in\mathcal{A}^n} 2^{-nH(q_x)} = \sum_{T\in T(p_1)} 2^{-nH(q_T)}|T|$$

By direct calculation in [4], $|T|$ can be bounded by $2^{nH(q_T)}$. The size of $T(p_1)$ can be bounded by $(n+1)^{|\mathcal{A}|}$. So

$$\sum_{x\in\mathcal{A}^n} 2^{-nH(q_x)} \leq (n+1)^{|\mathcal{A}|}$$

and for the same reason

$$\sum_{t_1\in\mathcal{A}^N} 2^{-NH(q_{t_1})} \leq (N+1)^{|\mathcal{A}|}$$

So the error under p_1 can be upper bounded by

$$\begin{aligned} \mathrm{P}_{\mathrm{err1}}(\Lambda) &\leq 2^{-m(\lambda-\rho(n))}(n+1)^{|\mathcal{A}|}(N+1)^{|\mathcal{A}|} \\ &= 2^{-m(\lambda-\rho(n))}2^{|\mathcal{A}|(\log(n+1)+\log(N+1))} \\ &= 2^{-m(\lambda-\rho(n)-|\mathcal{A}|\frac{\log(n+1)}{m}-|\mathcal{A}|\frac{\log(N+1)}{m})} \end{aligned}$$

We assume that $\frac{\log(n)}{m} \to 0$ as $n \to \infty$, $\frac{log(N+1)}{m} \to 0$ as $n \to \infty$ and $\rho(n) \to 0$. So the error exponent under class one is bounded by

$$\lim_{n\to\infty} -\frac{1}{m}\log(p_{\mathrm{err1}}) \geq \lambda$$

□

So the error under class one can be exponentially small with respect to m regardless of the relative size of m and (n, N). But we know from Theorem 3.3 that we need to set $m = \min(n, N)$ so that the classifier is consistent under class two.

The following theorem [3] states that the classification function defined above is asymptotically optimal even though it uses only the training data from class one. The expression of $\rho(n)$ is specified in the proof.

Theorem 3.4 (Asymptotically Optimal Classifier). *For any p_1, $p_2 \in \mathbb{P}(\mathcal{A})$ and any $\lambda > 0$. Let $\Omega = (\Omega_1, \Omega_2 = \Omega_1^c)$ be the decision regions specified by any decision rule that is independent of* p_1 *and* p_2 *such that*

$$\lim_{\min(n,N)\to\infty} -\frac{1}{\min(n,N)}\log \mathrm{P}_{\mathrm{err1}}(\Omega) \geq \lambda \tag{3.15}$$

Let the decision region $\Lambda = (\Lambda_1, \Lambda_2)$ be specified by (3.14). *Then*

$$\lim_{\min(n,N)\to\infty} -\frac{1}{\min(n,N)}\log \mathrm{P}_{\mathrm{err1}}(\Lambda) \geq \lambda \tag{3.16}$$

and

$$P_{\mathrm{err2}}(\Lambda) \leq P_{\mathrm{err2}}(\Omega) \tag{3.17}$$

The proof follows directly from the proof in [3] by replacing n with $\min(n, N)$. Theorem 3.4 shows a result that is very similar to the optimality of the Hoeffding test that asymptotic optimality is achieved even though only the training data from class one is used. It does not answer the following questions. What are the error exponents under both classes? What is the relative size between n and N such that the error under class two is exponentially small with respect to $\min(n, N)$? From [14], it turns out that if N and n are of the same order, N needs to satisfy that $N \geq K^\star n$ where $K^\star$ depends on λ, p_1 and p_2. Readers who are interested can find more information in [14].

3.3 Classification Function with Two Training Sequences

In this section, we propose a classification function that depends on both training sequences. We also prove Stein's lemma using the new classification function. We also present simulation results which show that our classification function outperforms the one in [3].

We have seen that asymptotic optimality can be achieved using only the training data from class one. In this section, we propose a classification function that depends on both T_1 and T_2 and show that

- The classification rule in [3] does not work for the special case $\lambda = 0$. This can be solved by using the function we propose and the best error exponent under class two is characterized by $D(p_2||p_1)$. Thus we prove Stein's lemma for classification.

- The classification rule that we propose incorporates additional prior information by utilizing both training data. It outperforms the asymptotically optimal one when the test data is limited.

Let $h(t_1, t_2, x) = \frac{1}{n} \log \frac{q_{t_1}(x)}{q_{t_2}(x)}$ and classification rule we study next is

$$\delta(t_1, t_2, x) = \begin{cases} \text{class one} & \text{if } h(t_1, t_2, x) - \lambda < 0; \\ \text{class two} & \text{otherwise} \end{cases} \tag{3.18}$$

Note that the classification function is inspired by the loglikelihood ratio test.

3.3.1 Stein's Lemma for Classification

Recall that the Hoeffding test is proved to be asymptotically optimal for any positive error exponent under H_0. But when the error exponent under H_0 is zero, a different test needs to be constructed in order to prove Stein's lemma [4]. The same issue also appears in classification problems. We have seen that the classification rule in [3] is asymptotically optimal for any positive error exponent under class one. But a different classification rule needs to be constructed in order to achieve zero error exponent under class one. We adopt the same classification function as (3.18) with thresholds adapted to the training data.

$$\delta(t_1, t_2, x) = \begin{cases} \text{class one} & \text{if } D(q_{t_1}||q_{t_2}) - \epsilon \leq \frac{1}{n}\log\frac{q_{t_1(x)}}{q_{t_2(x)}} \leq D(q_{t_1}||q_{t_2}) + \epsilon \ ; \\ \text{class two} & \text{otherwise} \end{cases} \tag{3.19}$$

Theorem 3.5 (Stein's Lemma for Classification)**.** *Assume* $\lim_{n\to\infty}\frac{n}{N} = 0$*. Let* $\Lambda_1 \subseteq \mathcal{A}^N \times \mathcal{A}^N \times \mathcal{A}^n$ *be the acceptance region for class one and* $\Lambda_2 = \Lambda_1^c$ *for class two. Let the error probabilities be*

$$\alpha_n = \Pr(\Lambda_2|H_1)$$

and

$$\beta_n = \Pr(\Lambda_1|H_2)$$

for any $0 < \eta < \frac{1}{2}$*, define*

$$\beta_n^\star = \min_{\substack{\Lambda_1 \subseteq \mathbf{A}^m \times \mathbf{A}^m \times \mathbf{A}^n \\ \alpha_n < \eta}} \beta_n \tag{3.20}$$

Then

$$\lim_{\eta\to 0}\lim_{n\to\infty} -\frac{1}{n}\log\beta_n^\star = D(p_1||p_2) \tag{3.21}$$

Proof. First prove that the test is consistent under H_1.

Choose $\delta < \epsilon$ and define the following regions.

$$A_\epsilon = \{t_1, t_2, x | D(q_{t_1}||q_{t_2}) - \epsilon \leq \frac{1}{n}\log\frac{q_{t_1(x)}}{q_{t_2(x)}} \leq D(q_{t_1}||q_{t_2}) + \epsilon\} \tag{3.22}$$

$$B_\delta = \{t_1, t_2 | D(p_1||p_2) - \delta \leq D(q_{t_1}||q_{t_2}) \leq D(p_1||p_2) + \delta\} \tag{3.23}$$

By definition, B_δ is the set of possible training sequences that the relative entropy between their empirical distributions is close to $D(p_1||p_2)$.

$$C_{\delta,\epsilon} = \{t_1, t_2, x | D(p_1||p_2) + \delta - \epsilon \leq \frac{1}{n}\log\frac{q_{t_1}(x)}{q_{t_2}(x)} \leq D(p_1||p_2) - \delta + \epsilon\} \tag{3.24}$$

It is not hard to see that $B_\delta \cap C_{\delta,\epsilon} \subseteq A_\epsilon$.

Next show that for any $\epsilon > 0$, the test is consistent under H_1. The test statistics can be equally written as

$$\frac{1}{n}\log\frac{q_{t_1}(x)}{q_{t_2}(x)} = -D(q_x||q_{t_1}) + D(q_x||q_{t_2}) \tag{3.25}$$

Under H_1, $||q_{t_1} - p_1||_{l_1} \to 0$ in probability because of the weak law of large numbers. So are

$||q_{t_2}-p_2||_{l_1}$ and $||q_x-p_1||_{l_1}$ for the same reason.

$$\begin{aligned}
|D(q_x||q_{t_2})-D(p_1||p_2)| &= |\sum_i q_x(i)\log\frac{q_x(i)}{q_{t_2}(i)}-\sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}| \\
&= |\sum_i (p_1(i)+q_x(i)-p_1(i))\log\frac{q_x(i)}{q_{t_2}(i)}-\sum_i p_1(i)\log\frac{p_1(i)}{p_2(i)}| \\
&= |\sum_i p_1(i)\log\frac{p_2(i)}{q_{t_2}(i)}+\sum_i p_1(i)\log\frac{q_x(i)}{p_1(i)}+\sum_i (q_x-p_1)\log\frac{q_x}{q_{t_2}}| \\
&\leq |\sum_i p_1(i)\log\frac{p_2(i)}{q_{t_2}(i)}| \\
&+ |\sum_i p_1(i)\log\frac{q_x(i)}{p_1(i)}|+|\sum_i (q_x-p_1)\log\frac{q_x}{q_{t_2}}|
\end{aligned}$$

$$|\sum_i p_1(i)\log\frac{p_2(i)}{q_{t_2}(i)}| \xrightarrow{\text{in Prob.}} 0$$

$$|\sum_i p_1(i)\log\frac{q_x(i)}{p_1(i)}| \xrightarrow{\text{in Prob.}} 0$$

$$|\sum_i (q_x-p_1)(i)\log\frac{q_x(i)}{q_{t_2}(i)}| \xrightarrow{\text{in Prob.}} 0$$

all because of the weak law of large numbers. With the same argument,

$$|D(q_x||q_{t_1})| \xrightarrow{\text{in Prob.}} 0$$

Putting together the above convergence result, we get that for any $\epsilon > \delta > 0$,

$$\begin{aligned}
\lim_{n,\ N\to\infty} \Pr(|-D(q_x||q_{t_1})+D(q_x||q_{t_2})-D(p_1||p_2)| \geq \epsilon-\delta|\ H_1) &= \lim_{n,\ N\to\infty} \Pr(C^c_{\delta,\epsilon}|\ H_1) \\
&= 0
\end{aligned}$$

With a similar argument,

$$\lim_{n,\ N\to\infty} \Pr(B^c_\delta|\ H_1)=0$$

So under H_1,

$$\begin{aligned}
\lim_{n,\ N\to\infty} \Pr(A_\epsilon|\ H_1) &\geq \lim_{n,\ N\to\infty} \Pr(B_\delta \cap C_{\epsilon,\delta}|\ H_1) \\
&\geq 1-\lim_{n,\ N\to\infty} \Pr(B^c_\delta|H_1)-\lim_{n,\ N\to\infty} \Pr(C^c_{\epsilon,\delta}|H_1) \\
&= 1
\end{aligned}$$

This proves that for any ϵ, the test is consistent under H_1.

Then prove a lower bound on the error exponent under H_2 of the proposed test. Let $\beta_\epsilon = \Pr(A_\epsilon|H_2)$. Pick ϵ small enough such that $\epsilon < \min p_1$ and $\epsilon < \min p_2$ and define two regions as follows,

$$Q_1 = \{t_1|D(q_{t_1}||p_1) < \frac{1}{2}\epsilon^2\}$$

$$Q_2 = \{t_2|D(q_{t_2}||p_2) < \frac{1}{2}\epsilon^2\}$$

A simple observation gives that

$$\begin{aligned}\beta_\epsilon &= \sum_{\{t_1,t_2,x\}\in A_\epsilon} p_1(t_1)p_2(t_2)p_2(x) \\ &\leq (\sum_{A_\epsilon\cap Q_1\cap Q_2} + \sum_{A_\epsilon\cap Q_1^c} + \sum_{\epsilon\cap Q_2^c})p_1(t_1)p_2(t_2)p_2(x)\end{aligned}$$

First deal with the last two summations.

$$\begin{aligned}\sum_{A_\epsilon\cap Q_1^c} p_1(t_1)p_2(t_2)p_2(x) &\leq \sum_{Q_1^c} p_1(t_1)p_2(t_2)p_2(x) \\ &= \sum_{Q_1^c} p_1(t_1) \\ &\leq (N+1)^M 2^{-N\frac{1}{2}\epsilon^2}\end{aligned}$$

$$\begin{aligned}\sum_{A_\epsilon\cap Q_2^c} p_1(t_1)p_2(t_2)p_2(x) &\leq \sum_{Q_2^c} p_1(t_1)p_2(t_2)p_2(x) \\ &= \sum_{Q_2^c} p_2(t_2) \\ &\leq (N+1)^M 2^{-N\frac{1}{2}\epsilon^2}\end{aligned}$$

Both are exponentially small with respect to N which is the size of the training sequences. Now we deal with the first summation. Using Pinsker's inequality,

$$||q_{t_1} - p_1||_{l_1} < \epsilon \quad \text{for} \;\; t_1 \in Q_1$$

and

$$||q_{t_2} - p_2||_{l_1} < \epsilon \quad \text{for} \;\; t_2 \in Q_2$$

Rewrite the empirical divergence rate between t_1 and t_2 as

$$D(q_{t_1}||q_{t_2}) = D(p_1||p_2) + \sum_i p_1 \log \frac{q_{t_1}}{p_1} + \sum_i p_1 \log \frac{p_1}{q_{t_2}} + \sum_i (q_{t_1} - p_1) \log \frac{q_{t_1}}{q_2}$$

We give upper and lower bound on each of the last three terms:

$$\begin{aligned} \left|\sum_i p_1 \log \frac{q_{t_1}}{p_1}\right| &\leq \sum_i p_1 \left|\log \frac{q_{t_1}}{p_1}\right| \\ &\leq \sum_i p_1 \max\{\log \frac{p_1+\epsilon}{p_1}, \log \frac{p_1}{p_1-\epsilon}\} \\ &\leq \sum_i p_1 \max\{\frac{\epsilon}{\min p_1}, \frac{\epsilon}{\min p_1 - \epsilon}\} \\ &= \frac{\epsilon}{\min p_1 - \epsilon} = O(\epsilon) \end{aligned}$$

and

$$\begin{aligned} \left|\sum_i p_1 \log \frac{p_2}{q_{t_2}}\right| &\leq \sum_i p_1 \left|\log \frac{q_{t_1}}{p_1}\right| \\ &\leq \frac{\epsilon}{\min p_2 - \epsilon} = O(\epsilon) \end{aligned}$$

and

$$\begin{aligned} \left|\sum_i (q_{t_1} - p_1) \log \frac{q_{t_1}}{q_{t_2}}\right| &\leq \sum_i |q_{t_1} - p_1| \left|\log \frac{q_{t_1}}{q_{t_2}}\right| \\ &\leq \sum_i |q_{t_1} - p_1| \max\{\log \frac{\max p_1 + \epsilon}{\min p_2 - \epsilon}, \log \frac{\max p_2 + \epsilon}{\min p_1 - \epsilon}\} \\ &\leq \epsilon \max\{\log \frac{\max p_1 + \epsilon}{\min p_2 - \epsilon}, \log \frac{\max p_2 + \epsilon}{\min p_1 - \epsilon}\} = O(\epsilon) \end{aligned}$$

So for any $t_1 \in Q_1$ and $t_2 \in Q_2$,

$$D(p_1||p_2) - O(\epsilon) \leq D(q_{t_1}||q_{t_2}) \leq D(p_1||p_2) + O(\epsilon)$$

For any $\{t_1, t_2, x\} \in A_\epsilon \cap Q_1 \cap Q_2$, combine the above

$$D(p_1||p_2) - O(\epsilon) \leq \frac{1}{n} \log \frac{q_{t_1}(x)}{q_{t_2}(x)} \leq D(p_1||p_2) + O(\epsilon)$$

The test statistic can be rewritten as

$$\frac{1}{n}\log\frac{q_{t_1}(x)}{q_{t_2}(x)} = \frac{1}{n}\log\frac{p_1(x)}{p_2(x)} + \frac{1}{n}\log\frac{q_{t_1}(x)}{p_1(x)} + \frac{1}{n}\log\frac{p_2(x)}{q_{t_2}(x)}$$

Use the same method to bound the last two terms,

$$\begin{aligned}\left|\frac{1}{n}\log\frac{q_{t_1}(x)}{p_1(x)}\right| &\leq \frac{1}{n}\sum_{j=1}^{n}\left|\log\frac{q_{t_1}(x_j)}{p_1(x_j)}\right| \\ &\leq \frac{1}{n}\sum_{j=1}^{n}\max\{\log(1+\frac{\epsilon}{p_1(x_j)}),\log(1+\frac{\epsilon}{p_1(x_j)-\epsilon}\} \\ &\leq \log(1+\frac{\epsilon}{\min p_1-\epsilon}) \leq \frac{\epsilon}{\min p_1-\epsilon} = O(\epsilon)\end{aligned}$$

and

$$\begin{aligned}\left|\frac{1}{n}\log\frac{p_2(x)}{q_{t_2}(x)}\right| &\leq \frac{1}{n}\sum_{j=1}^{n}\left|\log\frac{p_2(x_j)}{q_{t_2}(x_j)}\right| \\ &\leq \frac{\epsilon}{\min p_2-\epsilon} = O(\epsilon)\end{aligned}$$

Putting the above results together, we get that for any $\{t_1, t_2, x\} \in A_\epsilon \cap Q_1 \cap Q_2$,

$$D(p_1||p_2) - O(\epsilon) \leq \frac{1}{n}\log\frac{p_1(x)}{p_2(x)} \leq D(p_1||p_2) + O(\epsilon)$$

So $p_2(x)$ can be bounded by

$$p_1(x)2^{-n(D(p_1||p_2)+O(\epsilon))} \leq p_2(x) \leq p_1(x)2^{-n(D(p_1||p_2)-O(\epsilon))}$$

So the error under H_2 can be bounded as follows,

$$\begin{aligned}\sum_{A_\epsilon\cap Q_1\cap Q_2} p_1(t_1)p_2(t_2)p_2(x) &\leq \sum_{A_\epsilon\cap Q_1\cap Q_2} p_2(x) \\ &\leq \sum p_1(x)2^{-n(D(p_1||p_2)-O(\epsilon))} \\ &\leq 2^{-n(D(p_1||p_2)-O(\epsilon))}\end{aligned}$$

$$\beta_\epsilon \leq 2^{-n(D(p_1||p_2)-O(\epsilon))} + 2^{-N(\frac{1}{2}\epsilon^2 - M\frac{\log(N+1)}{N})} + 2^{-N(\frac{1}{2}\epsilon^2 - M\frac{\log(N+1)}{N})}$$

Since N is of greater order of n, take log of the error and normalize by $\frac{1}{n}$,

$$-\frac{1}{n}\log\beta_\epsilon \geq D(p_1||p_2) - O(\epsilon)$$

Lastly, we prove the converse that the error exponents cannot be better. Assume A^* is the optimal decision region corresponding to H_1 which satisfies that

$$\Pr((A^*)^c|H_1) \leq \alpha$$

Let β^* be the error under H_2

$$\begin{aligned}
\beta^* &= \sum_{\{t_1,t_2,x\}\in A^*} p_1(t_1)p_2(t_2)p_2(x) \\
&\geq \sum_{\{t_1,t_2,x\}\in A^*\cap A_\epsilon\cap Q_1\cap Q_2} p_1(t_1)p_2(t_2)p_2(x) \\
&\geq 2^{-n(D(p_1||p_2)+O(\epsilon))} \sum_{\{t_1,t_2,x\}\in A^*\cap A_\epsilon\cap Q_1\cap Q_2} p_1(t_1)p_2(t_2)p_1(x)
\end{aligned}$$

The summation can be bounded by

$$\begin{aligned}
\sum_{A^*\cap A_\epsilon\cap Q_1\cap Q_2} p_1(t_1)p_2(t_2)p_1(x) &\geq 1 - \Pr((A^*)^c|H_1) - \Pr(A_\epsilon^c|H_1) - p_1(Q_1^c) - p_2(Q_2^c) \\
&\geq 1 - \alpha - \alpha - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})} - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})}
\end{aligned}$$

Combine the above two inequalities,

$$\beta^* \geq 2^{-n(D(p_1||p_2)+O(\epsilon))}(1 - \alpha - \alpha - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})} - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})})$$

Since N is of greater order of n, take log of β^* and normalize it by $\frac{1}{n}$,

$$-\frac{1}{n}\log\beta^* \leq -\frac{1}{n}\log(1 - 2\alpha - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})} - 2^{-N(\frac{1}{2}\epsilon^2 - \frac{M\log(N+1)}{N})}) + D(p_1||p_2) + O(\epsilon)$$

Let $\epsilon \to 0$, we get the desired result. □

Note that the above result is achieved by letting the training data grow much faster than the test data. So the error brought by false modeling of the classes is negligible compared to the error brought by the test data. In this case, the classifier basically has full knowledge of the classes.

3.3.2 Improvement of Finite Sample Performance

Recall that though the Hoeffding test is proved to be asymptotically optimal, the mismatched test, which incorporates additional prior information, has better finite sample performance [2]. The improvement is implied by the asymptotic mean and variance analysis of the test statistic (2.27) – (2.32). Simulation results also show that the mismatched test outperforms the Hoeffding test with finite samples. In this section, we present simulation results to compare the finite sample size performance of the Gutman classifier and the loglikelihood ratio classifier.

Figure 3.1 shows the receiver operational curve (ROC) of both classifiers when the training sequences from both classes are of the same length. The larger the area under the curve, the better the performance is. The solid blue line corresponds to the ROC of the loglikelihood ratio classifier. The dotted green line corresponds to the Gutman classifier. The two underlying distributions are generated randomly on alphabet $\mathcal{A}$ with $|\mathcal{A}| = 5$. Both of the distributions have full support over $\mathcal{A}$. The size of the test data is $n = 50$. The training sequences from both classes are of size $N = 50$. The Gutman classifier utilizes only the training sequence from class one. The loglikelihood ratio classifier utilizes both of the training sequences. As we can see in Figure 3.1, our classifier outperforms the Gutman classifier when the test samples are limited. The improvement comes from incorporating additional information that characterizes class two.

The fact that these two training sequences are of the same size merits them similar importance in classifying future samples. If one of the training sequences is a lot shorter than the other, the information it provides suffers from more serious inaccuracy. Recall that the classification errors come from both false modeling of the classes and classification itself, and that the final error is determined by the dominating term among those two. Intuitively, if one of the training sequences is too short, the false modeling error it brings may dominate the total error. So the classifier may benefit from not utilizing that training data. Figure 3.2 verifies this argument. In Figure 3.2, the training sequence from class one and the test data are of the same size as in Figure 3.1. But the training sequence from class two is reduced to only 15 samples. Figure 3.2 does not compare the high false alarm region of the ROCs. This is because a very high false alarm region is not achievable for the loglikelihood ratio classifier. Recall that the classification function for the loglikelihood ratio classifier is

$$h(t_1, t_2, x) = \log \frac{q_{t_1}(x)}{q_{t_2}(x)}. \tag{3.26}$$

Though both of the underlying distributions have full support over $\mathcal{A}$, there is a relatively high probability that q_{t_2} does not have full support because of the estimation error between

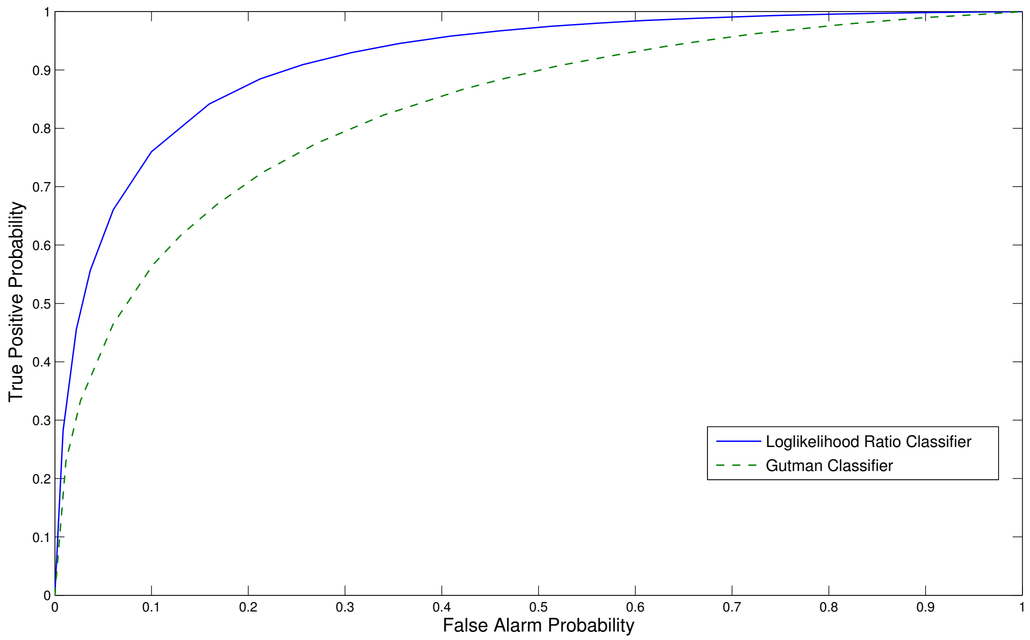

Figure 3.1: Receiver Operation Characteristic, the Gutman Classifier and the Likelihood Ratio Classifier, Training Sequence with Equal Size

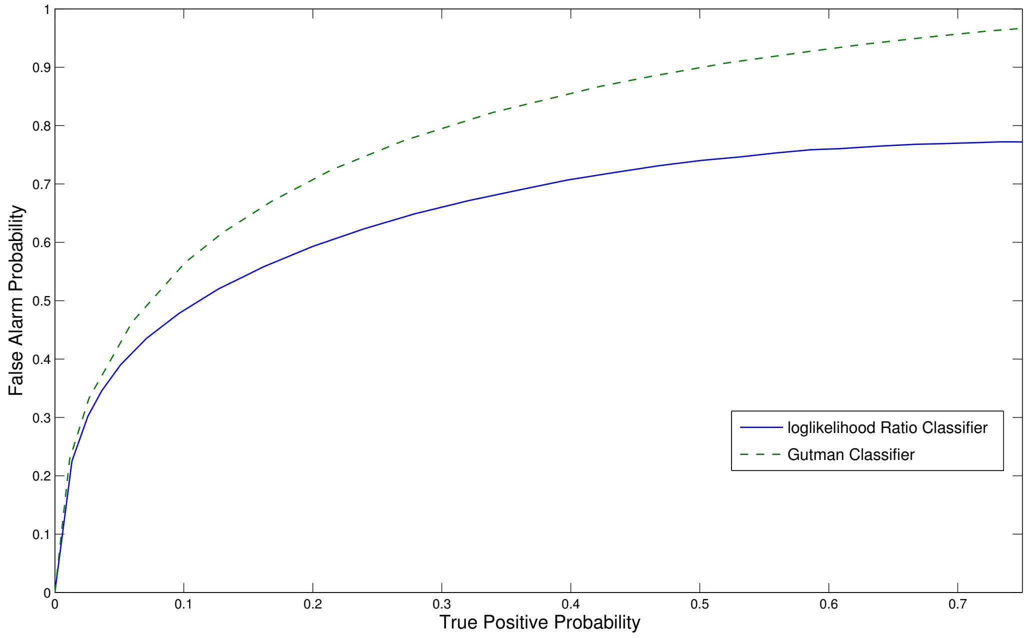

Figure 3.2: Receiver Operation Characteristic, the Gutman Classifier and the Likelihood Ratio Classifier, Less Training Data from Class Two

q_{t_2} and p_2. The estimation error is made worse by the limited size of T_2. Whenever q_{t_2} does not have full support over $\mathcal{A}$, the classification function is infinity, which is greater than any threshold. So the high false alarm region is not achievable. As we can see from Figure 3.2, the loglikelihood ratio classifier is outperformed by the Gutman classifier. It is not clear how fast the training sequence from class two should grow such that incorporating it would benefit the finite sample performance.

CHAPTER 4

CONCLUDING REMARKS

We summarize the main contributions of this thesis in Section 4.1, and in Section 4.2 we briefly discuss some promising directions based on the results presented in Chapter 2–3.

4.1 Summary of Contributions

The three main contributions of this thesis are:

- demonstration that in binary classification problems, it may be rewarding to utilize training sequences from both classes;
- proof of Stein's lemma for classification which characterizes the maximal error exponent under one class;
- an account of the performance limit of hypothesis testing in the case of countably infinite alphabet.

4.2 Future Extension

In Figure 3.1, we can see that utilizing training sequences from both classes improves the performance of classification. In Figure 3.2, the simulation yields an opposite result. It may be rewarding to utilize training sequences from both classes, and the relative size of the two training sequences is a factor. A meaningful extension will be to study the conditions under which the improvement is guaranteed.

We prove Stein's lemma for classification under the condition that the size of the training sequence is of higher order than that of the test sequence. The maximal error exponent under one class is given by the relative entropy. This result is not surprising given Stein's lemma for hypothesis testing. This result relies on the fact that the training data grows much faster than the test data. In practice, this means that a large amount of training data needs to be collected before a test can be done. A useful extension will be to study the maximal error exponent under one class, under the condition that the training and test data are of the same size.

REFERENCES

[1] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, Apr. 1965.

[2] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, “Universal and composite hypothesis testing via mismatched divergence,” *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1587–1603, Mar. 2011.

[3] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *Information Theory, IEEE Transactions on*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 99th ed. Wiley-Interscience, Aug. 1991.

[5] R. Blahut, “Hypothesis testing and information theory,” *Information Theory, IEEE Transactions on*, vol. 20, no. 4, pp. 405–417, Jul. 1974.

[6] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *Information Theory, IEEE Transactions on*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.

[7] B. Clarke and A. Barron, “Information-theoretic asymptotics of bayes methods,” *Information Theory, IEEE Transactions on*, vol. 36, no. 3, pp. 453–471, May. 1990.

[8] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[9] I. Csiszar and P. Shields, *Information Theory and Statistics: A Tutorial.* Foundations and Trends in Communications and Information Theory, 2004, vol. 1, no. 4.

[10] A. Antos and I. Kontoyiannis, “Convergence properties of functional estimates for discrete distributions,” *Random Structures and Algorithms*, vol. 19, pp. 163–193, Dec. 2001.

[11] A. J. Wyner and D. Foster, “On the lower limits of entropy estimation,” Technical report, University of Pennsylvania, Jun. 2003.

[12] A. R. Barron, “Uniformly powerful goodness of fit tests,” *The Annals of Statistics*, vol. 17, no. 1, pp. 107–124, 1989.

[13] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *Information Theory, IEEE Transactions on*, vol. 37, no. 2, pp. 285–290, Mar. 1991.

[14] J. Ziv, "On classification with empirically observed statistics and universal data compression," *Information Theory, IEEE Transactions on*, vol. 34, no. 2, pp. 278–286, Mar. 1988.