UNIVERSAL OUTLIER HYPOTHESIS TESTING WITH APPLICATIONS TO ANOMALY DETECTION

BY

YUN LI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Venugopal V. Veeravalli, Chair
Professor Pierre Moulin
Associate Professor Prashant Mehta
Assistant Professor Lav R. Varshney

ABSTRACT

Outlier hypothesis testing is studied in a universal setting. Multiple sequences of observations are collected, a small subset (possibly empty) of which are outliers. A sequence is considered an outlier if the observations in that sequence are distributed according to an "outlier" distribution, distinct from the "typical" distribution governing the observations in the majority of the sequences. The outlier and typical distributions are not fully known, and they can be arbitrarily close. The goal is to design a universal test to best discern the outlier sequence(s). Both fixed sample size and sequential settings are considered in this dissertation. In the fixed sample size setting, for models with exactly one outlier, the generalized likelihood test is shown to be universally exponentially consistent. A single letter characterization of the error exponent achieved by such a test is derived, and it is shown that the test achieves the optimal error exponent asymptotically as the number of sequences goes to infinity. When the null hypothesis with no outlier is included, a modification of the generalized likelihood test is shown to achieve the same error exponent under each non-null hypothesis, and also consistency under the null hypothesis. Then, models with multiple outliers are considered. When the outliers can be distinctly distributed, in order to achieve exponential consistency, it is shown that it is essential that the number of outliers be known at the outset. For the setting with a known number of distinctly distributed outliers, the generalized likelihood test is shown to be universally exponentially consistent. The limiting error exponent achieved by such a test is characterized, and the test is shown to be asymptotically exponentially consistent. For the setting with an unknown number of identically distributed outliers, a modification of the generalized likelihood test is shown to achieve a positive error exponent under each non-null hypothesis, and consistency under the null hypothesis. In the sequential setting, a test with the flavor of the repeated significance test is proposed. The test is shown to be universally consistent, and universally exponentially consistent under non-null hypotheses. In addition, with the typical distribution being known, the test is shown to be asymptotically optimal universally when the number of outliers is the largest possible. In all cases, the asymptotic performance of the proposed test when none of the underlying distributions is known is shown to converge to that when only the

typical distribution is known as the number of sequences goes to infinity. For models with continuous alphabets, a test with the same structure as the generalized likelihood test is proposed, and it is shown to be universally consistent. It is also demonstrated that there is a close connection between universal outlier hypothesis testing and cluster analysis. The performance of various proposed tests is evaluated against a synthetic data set, and contrasted with that of two popular clustering methods. Applied to a real data set for spam detection, the sequential test is shown to outperform the fixed sample size test when the lengths of the sequences exceed a certain value. In addition, the performance of the proposed tests is shown to be superior to that of another kernel-based test for large sample sizes.

To Baba and Mama, who have been there for me since day one. Thank you for everything you have done.

ACKNOWLEDGMENTS

First and foremost I would like to give special thanks to my advisor Prof. Venugopal Veeravalli for being a great mentor over the past three years. You are the example that it is not only top-notch research, but also enthusiasm, compassion, and genuine caring for academic advances that make a great scholar. Thank you for providing an encouraging research environment where I had the freedom to pursue various topics and approaches. I could never have made it without your guidance and support, and I am honored to be your student. I also cherish every moment that I shared with you, Starla, and the whole group in GM 3, and I look forward to another round of Pit anytime soon!

I would also like to mention the other members on my doctoral committee, Professors Pierre Moulin, Prashant Mehta, and Lav R. Varshney. Your academic input is greatly appreciated. And your humor made my defense a cheering and enjoyable moment. Thank you.

I would like to express my gratitude to my colleague, Dr. Sirin Nitinawarat. Thank you for your guidance, and your contribution to this work. I admire your humble nature, and your down-to-earth attitude toward research. I am very fortunate to have you as a friend.

A special thanks to my former academic advisor, Prof. Sean P. Meyn. You have been especially helpful in providing academic advices during my first three years as a graduate student. Thank you for being so patient in answering every "silly" question that I had as a fresh graduate.

Last but not least, I would like to thank all the faculty members in the ECE department for giving world-class lectures. I enjoyed every class I took, and I will continue to benefit from what I learnt as a student.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

We consider the following inference problem, which we term *outlier hypothesis testing.* Among a fixed number of independent and memoryless observation sequences, it is assumed that there is a small subset (possibly empty) of outlier sequences. Specifically, most of the sequences are assumed to be distributed according to a "typical" distribution, while an outlier sequence is distributed according to an "outlier distribution," distinct from the typical distribution. We are interested in *universal* settings of this problem, where the outlier and typical distributions are not fully known, and can be arbitrarily close. The goal is to design a test, which does not depend on any unknown distribution, to best identify *all* the outlier sequences. Outlier hypothesis testing finds possible applications in fraud and anomaly detection in large data sets [1, 2], severe weather prediction, environment monitoring in sensor networks [3], network intrusion and voting irregularity analysis. It also finds applications where the term "outlier" has a positive connotation, such as spectrum sensing and high-frequency trading.

We study both fixed sample size (FSS) and sequential settings of universal outlier hypothesis testing. In the FSS setting, the number of observations that are taken before a final decision is made is determined at the outset, and the goal is to identify the outlier sequences with a certain accuracy using as few observations as possible. In the sequential setting, observations are collected sequentially over a period of time. At each time, a test either decides to continue taking one more observation, or to stop and make a final decision. As a result, the number of observations that are collected before the test terminates is not fixed, but rather a random value. The goal in the sequential setting is to achieve a certain accuracy using the fewest observations *on average.*

1.1 Related Problems

Universal outlier hypothesis testing is related to a broader class of *composite hypothesis testing* problems in which there is uncertainty in the probabilistic laws associated with some or all of the hypotheses. To solve these problems, a popular approach is to apply the

generalized likelihood (GL) test [4, 5]. For example, in the *simple-versus-composite* case, the goal is to make a decision in favor of either the null distribution, which is known to the tester, or a family of alternative distributions. A fundamental result concerning the asymptotic optimality of the *generalized likelihood ratio test (GLRT)* in this case was shown in [6]. When some uncertainty is present in the null distribution as well, i.e., the *composite-versus-composite* setting, the optimality of the GLRT has been examined under various conditions [5].

Universal outlier hypothesis testing is also related to homogeneity testing and classification [7–11]. In homogeneity testing, one wishes to decide whether or not two samples come from the same probabilistic law. In classification problems, a set of test data is classified into one of multiple streams of training data with distinct labels. Metrics that are commonly used to quantify the performance of a test are *consistency* and *exponential consistency*. A universal test is *consistent* if the error probability approaches zero as the sample size goes to infinity, and is *exponentially consistent* if the decay is exponential with sample size. In [10, 11], a classifier based on the principle of the GL test was shown to be optimal under the asymptotic Neyman-Pearson criterion. In particular, in [10], the classifier is designed to minimize the error probability under the inhomogeneous hypothesis, under a predefined constraint on the exponent for the error probability under the homogeneous hypothesis. And, in [11], the classifier is designed to minimize the probability of rejection, under a constraint on the probability of misclassification. However, the aforementioned optimality is achieved only when the length of the training data grows at least linearly with that of the test data, and the distribution of the test data is separated enough from those of all unmatched training data.

There is a close connection between universal outlier hypothesis testing and cluster analysis. In fact, we can show that our proposed FSS test in Chapter 3 is equivalent to a clustering algorithm that performs cluster analysis over the probability simplex (cf. Chapter 6). The goal of cluster analysis is to partition a data set into subgroups, or clusters, such that data points within the same cluster are more closely related to one another than to those in different clusters [12–15]. A diverse collection of algorithms has been proposed for cluster analysis. For instance, the K-means algorithm (and also the K-medoids algorithm) is a classic prototype-based clustering technique that creates a one-level partition of the data set [16–18]. In contrast, hierarchical clustering produces nested clusters that can be organized as a tree. Methods for hierarchical clustering may be divided into two basic paradigms: *agglomerative* [13, 14] and *divisive* [19, 20]. Density-based clustering methods define a cluster as a dense region of data points, which is surrounded by a region of low density [21, 22]. Graph-based clustering techniques are appropriate if the closeness between different data

points can be represented by the edge structure of a (weighted) proximity graph [15, 23]. And the task of graph clustering is to group the vertices into disjoint components in such a way that there should be many more edges within each component compared with those between components.

It is to be noted that outlier hypothesis testing is distinct from statistical *outlier detection* [24, 25]. In outlier detection, the goal is to efficiently winnow out a few outlier observations from a single sequence of observations. The outlier observations are assumed to follow a different generating mechanism from that governing the normal observations. Statistical outlier detection is typically used to preprocess large data sets, to obtain clean data that is used for purposes such as inference and control. The main differences between statistical outlier detection and outlier hypothesis testing are: (i) in the former problem, the outlier observations constitute a much smaller fraction of the entire observations than in the latter problem, and (ii) these outlier observations can be arbitrarily spread out among all observations in the outlier detection problem, whereas all the outlier observations are concentrated in a *fixed* subset of sequences in the outlier hypothesis testing problem.

1.2 Dissertation Outline

We now provide a brief overview of each chapter.

- In Chapter 2, we introduce notations, and provide some useful identities and well-known technical facts.

- The FSS setting is studied in Chapter 3, where we show that the GL test is far more efficient for universal outlier hypothesis testing than for the other inference problems, such as homogeneity testing and classification [7–11]. In particular, the GL test is universally exponentially consistent as long as the outlier distributions are distinct from the common typical distribution, and there is indeed an outlier among the sequences. Furthermore, we prove that the GL test is *asymptotically efficient* in the limit of a large number of sequences in certain settings. When it is also possible that there is no outlier present, a modification of the GL test is shown to be consistent under all hypotheses, and exponentially consistent under every non-null hypothesis.

- In Chapter 4, we generalize our findings in the FSS setting to the sequential setting. We propose a sequential test that has the flavor of the Multihypothesis Sequential Probability Ratio Test [26, 27] and the repeated significance test [28, 29]. The sequential test is shown to be universally consistent, and universally exponentially consistent

conditioned on an outlier being present. In addition, when the outliers are identically distributed, it is shown to be asymptotically optimal when the number of outliers is the largest possible, and with the typical distribution being known.

- In Chapter 5, we extend our results to models with continuous alphabets. We propose an FSS test that is of the same spirit as the GL test, and uses non-parametric estimates of the Kullback-Leibler (KL) divergence. The proposed test is shown to be universally consistent for various settings. In addition, we compare the performance of the proposed test with that of a kernel-based test against a synthetic data set.

- In Chapter 6, we elaborate on our discussion on the connection between universal outlier hypothesis testing and cluster analysis. We evaluate the performance of the FSS test and two other clustering algorithms on a synthetic data set, where it is discovered that the FSS test outperforms the clustering algorithms when the sample size is sufficiently large.

- In Chapter 7, we apply the proposed tests to a real data set for spam detection. The performance of another kernel-based universal test is also evaluated for contrast. The FSS test outperforms three different versions of the kernel-based test for large sample size. And the performance of the sequential test is superior to that of the FSS test when the average stopping time is sufficiently large.

- We provide concluding remarks and comment on future work in Chapter 8.

CHAPTER 2

PRELIMINARIES

Throughout the dissertation, random variables are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All random variables are assumed to take values in finite sets if not specified otherwise, and all logarithms are the natural ones.

For a finite set $\mathcal{Y}$, let $\mathcal{Y}^m$ denote the m Cartesian product of $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ denote the set of all probability mass functions (pmfs) on $\mathcal{Y}$. The empirical distribution of a sequence $\boldsymbol{y} = y^m = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, denoted by $\gamma = \gamma_{\boldsymbol{y}} \in \mathcal{P}(\mathcal{Y})$, is defined as

$$\gamma(y) \triangleq \frac{1}{m} \big| \{k = 1, \ldots, m : y_k = y\} \big|,$$

$y \in \mathcal{Y}$.

Our results will be stated in terms of various distance metrics between a pair of distribution $p, q \in \mathcal{P}(\mathcal{Y})$. In particular, we shall consider two symmetric distance metrics: the *Bhattacharyya distance* and *Chernoff information*, denoted respectively by $B(p, q)$ and $C(p, q)$, and defined as (see, e.g., [30])

$$B(p, q) \triangleq -\log \left(\sum_{y \in \mathcal{Y}} p(y)^{\frac{1}{2}} q(y)^{\frac{1}{2}} \right) \tag{2.1}$$

and

$$C(p, q) \triangleq \max_{s \in [0,1]} -\log \left(\sum_{y \in \mathcal{Y}} p(y)^s q(y)^{1-s} \right), \tag{2.2}$$

respectively. Another distance metric, which will be key to our study, is the relative entropy, denoted by $D(p\|q)$ and defined as

$$D(p\|q) \triangleq \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}. \tag{2.3}$$

Unlike the Bhattacharyya distance (2.1) and Chernoff information (2.2), the relative entropy in (2.3) is a *non-symmetric* distance [30].

The following technical facts will be useful; their derivations can be found in [30, Theorem 11.1.2]. Consider random variables Y^n which are i.i.d. according to $p \in \mathcal{P}(\mathcal{Y})$. Let $y^n \in \mathcal{Y}^n$ be a sequence with an empirical distribution $\gamma \in \mathcal{P}(\mathcal{Y})$. It follows that the probability of such sequence y^n, under p and under the i.i.d. assumption, is

$$p(y^n) \;=\; \exp\Big\{ -n\left(D(\gamma\|p) + H(\gamma)\right)\Big\}, \tag{2.4}$$

where $D(\gamma\|p)$ and $H(\gamma)$ are the relative entropy of γ and p, and entropy of γ, defined as

$$D(\gamma\|p) \;\triangleq\; \sum_{y\in\mathcal{Y}} \gamma(y)\log\frac{\gamma(y)}{p(y)},$$

and

$$H(\gamma) \;\triangleq\; -\sum_{y\in\mathcal{Y}} \gamma(y)\log\gamma(y),$$

respectively. Consequently, it holds that for each y^n, the pmf p that maximizes $p(y^n)$ is $p=\gamma$, and the associated maximal probability of y^n is

$$\gamma(y^n) \;=\; \exp\big\{ -nH(\gamma)\big\}. \tag{2.5}$$

Next, for each $n \geq 1$, the number of all possible empirical distributions from a sequence of length n in $\mathcal{Y}^n$ is upper bounded by $(n+1)^{|\mathcal{Y}|}$, where $|\mathcal{Y}|$ denotes the (finite) size of $\mathcal{Y}$. Using this fact, (2.4) and a bound on the size of the set of sequences with the same empirical distribution (see, e.g., [30, Theorem 11.1.3] for details), it can be shown that the probability that the i.i.d. sequence Y^n that is distributed according to p has the empirical distribution $\gamma = q$, (for a feasible q) satisfies

$$\mathbb{P}\left\{\gamma = q\right\} \;\leq\; e^{-nD(q\|p)}. \tag{2.6}$$

We shall also find the following "sum centroid" inequality and its consequence useful. Consider any collection $\mathcal{C}$ of distributions on $\mathcal{Y}: p_i, i\in\mathcal{C}$. Then, for any arbitrary distribution q,

$$\sum_{i\in\mathcal{C}} D\left(p_i \middle\| \frac{\sum_{j\in\mathcal{C}} p_j}{|\mathcal{C}|}\right) \leq \sum_{i\in\mathcal{C}} D\left(p_i\|q\right). \tag{2.7}$$

The proof of (2.7) follows from the fact that for any distribution q,

$$\sum_{i\in\mathcal{C}} D\left(p_i\middle\|\frac{\sum_{j\in\mathcal{C}} p_j}{|\mathcal{C}|}\right) \;=\; \sum_{i\in\mathcal{C}} D\left(p_i\|q\right) - |\mathcal{C}|D\left(\frac{\sum_{i\in\mathcal{C}} p_i}{|\mathcal{C}|}\middle\|q\right).$$

Now for a pair of distributions $p,\bar{p}$ on $\mathcal{Y}$, particularizing (2.7) to the special case, where $\mathcal{C}$ comprises one p distribution and L copies of the $\bar{p}$ distribution, and with q in (2.7) being $\bar{p}$, we have that

$$D\left(p\middle\|\frac{p+L\bar{p}}{L+1}\right) + LD\left(\bar{p}\middle\|\frac{p+L\bar{p}}{L+1}\right) \;\leq\; D(p\|\bar{p}). \tag{2.8}$$

The proofs in future sections rely on the following lemmas.

Lemma 1. *Let $\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(J)}$ be mutually independent random vectors with each $\boldsymbol{Y}^{(j)}$, $j=1,\ldots,J$, being n i.i.d. repetitions of a random variable distributed according to $p_j\in\mathcal{P}(\mathcal{Y})$. Let A_n be the set of all J tuples $\left(\boldsymbol{y}^{(1)},\ldots,\boldsymbol{y}^{(J)}\right)\in\mathcal{Y}^{Jn}$ whose empirical distributions $(\gamma_1,\ldots,\gamma_J)=\left(\gamma_{\boldsymbol{y}^{(1)}},\ldots,\gamma_{\boldsymbol{y}^{(J)}}\right)$ lie in a closed set $E\in\mathcal{P}(\mathcal{Y})^J$. Then, it holds that*

$$\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\Big\{\left(\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(J)}\right)\in A_n\Big\} = \min_{(q_1,\ldots,q_J)\,\in\,E}\;\sum_{j=1}^{J}\; D\left(q_j\|p_j\right). \tag{2.9}$$

Proof. Let $\overline{E}$ be the set of all joint distributions in $\mathcal{P}\left(\mathcal{Y}^J\right)$ with the tuple of their corresponding marginal distributions lying in E. It now follows from the closeness of E in $\mathcal{P}(\mathcal{Y})^J$ and the compactness of $\mathcal{P}\left(\mathcal{Y}^J\right)$ that $\overline{E}$ is also closed in $\mathcal{P}\left(\mathcal{Y}^J\right)$. Let $\overline{A}_n$ be the set of all J tuples $\left(\boldsymbol{y}^{(1)},\ldots,\boldsymbol{y}^{(J)}\right)=\left(\left(y_1^{(1)},\ldots,y_n^{(1)}\right),\ldots,\left(y_1^{(J)},\ldots,y_n^{(J)}\right)\right)\in\mathcal{Y}^{Jn}$ whose joint empirical distribution lies in a closed set $\overline{E}\in\mathcal{P}\left(\mathcal{Y}^J\right)$. The lemma then follows by observing that $\mathbb{P}\Big\{\left(\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(J)}\right)\in A_n\Big\} \;=\; \mathbb{P}\Big\{\left(\left(y_1^{(1)},\ldots,y_n^{(1)}\right),\ldots,\left(y_1^{(J)},\ldots,y_n^{(J)}\right)\right)\in\overline{A}_n\Big\}$, and by

invoking Sanov's theorem to compute the exponent of the latter probability, i.e.,

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\Big\{\Big(\boldsymbol{Y}^{(1)},\ldots,\boldsymbol{Y}^{(J)}\Big)\in A_n\Big\}\\
&=\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}\Big\{\Big(\big(y_1^{(1)},\ldots,y_n^{(1)}\big),\ldots,\\
&\qquad\qquad\big(y_1^{(J)},\ldots,y_n^{(J)}\big)\Big)\in\overline{A}_n\Big\}\\
&=\min_{q\,\in\,\overline{E}}\ D\left(q\|p_1\times\ \ \ldots\ \ \times p_J\right)\\
&=\min_{(q_1,\ldots,q_J)\,\in\,E}\ \sum_{j=1}^{J}\ D\left(q_j\|p_j\right).
\end{aligned}$$

□

Lemma 2. *For any two pmfs p_1, $p_2\in\mathcal{P}(\mathcal{Y})$ with full supports, it holds that*

$$2B\left(p_1\,,\,p_2\right)=\min_{q\,\in\,\mathcal{P}(\mathcal{Y})}\Big(D\left(q\|p_1\right)+D\left(q\|p_2\right)\Big). \tag{2.10}$$

In particular, the minimum on the right side of (2.10) is achieved by

$$q^{\star}\ =\ \frac{p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)}{\sum\limits_{y\in\mathcal{Y}}p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)},\ y\in\mathcal{Y}. \tag{2.11}$$

Proof. It follows from the concavity of the logarithm function that

$$\begin{aligned}
D\left(q\|p_1\right)+D\left(q\|p_2\right)&=\sum_{y\in\mathcal{Y}}q(y)\log\frac{q^2(y)}{p_1(y)p_2(y)}\\
&=-2\sum_{y\in\mathcal{Y}}q(y)\log\frac{p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)}{q(y)}\\
&\geq-2\log\left(\sum_{y\in\mathcal{Y}}p_1^{\frac{1}{2}}(y)p_2^{\frac{1}{2}}(y)\right)\\
&=2B(p_1,p_2).
\end{aligned} \tag{2.12}$$

In particular, equality is achieved in (2.12) by $q(y)=q^{\star}(y)$ in (2.11).

It is interesting to note that from (2.10), we recover the known inequality discovered in [31]:

$$2B\left(p_1\,,\,p_2\right)\ \leq\ \min\left(D\left(p_2\|p_1\right),D\left(p_1\|p_2\right)\right), \tag{2.13}$$

by evaluating the argument distribution q on the right side of (2.10) by p_1 and p_2, respectively. □

Lemma 3. *For any two pmfs p_1, $p_2 \in \mathcal{P}(\mathcal{Y})$ with full supports, it holds that*

$$C(p_1, p_2) \;\leq\; 2B(p_1, p_2).$$

Proof. The proof follows from an alternative characterization (instead of (2.2)) of the $C(p_1, p_2)$ as (cf. [32])

$$C(p_1, p_2) \;=\; \min_{q \in \mathcal{P}(\mathcal{Y})} \max\left(D(q\|p_1), D(q\|p_2)\right), \tag{2.14}$$

and upon noting that the objective function for the optimization problem in (2.14) is always no larger than that for the one in (2.10). □

CHAPTER 3

FIXED SAMPLE SIZE SETTING

3.1 Exactly One Outlier

Consider $M \geq 3$ independent sequences of observations, each of which consists of n independent and identically distributed (i.i.d.) observations. We denote the k-th observation of the i-th sequence by $Y_k^{(i)}$, which takes values in a finite set denoted by $\mathcal{Y}$. It is assumed that only one sequence is the "outlier," i.e., the observations in that sequence are uniquely distributed (i.i.d.) according to the "outlier" distribution $\mu \in \mathcal{P}(\mathcal{Y})$, while all the other sequences are commonly distributed according to the "typical" distribution $\pi \in \mathcal{P}(\mathcal{Y})$. *We are interested in a non-parametric setting, in which μ and π are not fully known and can be arbitrarily close. We further assume that both μ and π have full support over the finite alphabet $\mathcal{Y}$.* The assumption of μ, π having full support rules out trivial cases where it is straightforward to identify the outlier sequence. Clearly, if $M = 2$, either sequence can be considered as an outlier; hence, it becomes degenerate to consider outlier hypothesis testing in this case.

It is assumed throughout this section that the outlier distribution is unknown but is independent of the identity of the outlier. In certain applications, it may be natural to consider the model where the outlier distribution can vary depending on the identity of the outlier. This scenario can be viewed as a special case (with the number of outlier sequences being exactly one) of the multiple outlier hypothesis testing problem studied in Section 3.3.

Conditioned on the hypothesis that the i-th sequence is the outlier, the joint distribution of all the observations is

$$\begin{aligned} p_i\big(y^{Mn}\big) &= p_i\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \\ &= \prod_{k=1}^{n} \Big\{ \mu\big(y_k^{(i)}\big) \prod_{j \neq i} \pi\big(y_k^{(j)}\big) \Big\} \\ &\triangleq L_i\left(y^{Mn}, \mu, \pi\right), \end{aligned} \tag{3.1}$$

where

$$\boldsymbol{y}^{(i)} \;=\; \left(y_1^{(i)}, \ldots, y_n^{(i)}\right),\ i = 1, \ldots, M.$$

The test for the outlier sequence is done based on a *universal* rule $\delta : \mathcal{Y}^{Mn} \to \{1, \ldots, M\}$. In particular, the test δ is not allowed to be a function of (μ, π).

For a universal test, the maximal error probability, which will be a function of the test and (μ, π), is

$$e\big(\delta, (\mu, \pi)\big) \;\triangleq\; \max_{i=1,\ldots,M} \sum_{y^{Mn}:\ \delta(y^{Mn}) \,\neq\, i} p_i\big(y^{Mn}\big),$$

and the corresponding error exponent is defined as

$$\alpha\big(\delta, (\mu, \pi)\big) \;\triangleq\; \lim_{n\to\infty} -\frac{1}{n} \log e\big(\delta, (\mu, \pi)\big). \tag{3.2}$$

Throughout the dissertation, we consider the error exponent as n goes to infinity, while M, and hence the number of hypotheses, is kept fixed. Consequently, the error exponent in (3.2) also coincides with the one for the average probability of error.

A test is termed *universally consistent* if the maximal error probability converges to zero as the number of samples goes to infinity, i.e.,

$$e\big(\delta, (\mu, \pi)\big) \to 0, \tag{3.3}$$

for any (μ, π), $\mu \neq \pi$ as $n \to \infty$. It is termed *universally exponentially consistent* if the exponent for the maximal error probability is strictly positive, i.e.,

$$\alpha\big(\delta, (\mu, \pi)\big) > 0, \tag{3.4}$$

for any (μ, π), $\mu \neq \pi$.

3.1.1 Generalized Likelihood Test

We now describe the generalized likelihood test in two setups when only π is known, and when neither μ nor π is known, respectively.

For each $i = 1, \ldots, M$, denote the empirical distributions of $\boldsymbol{y}^{(i)}$ by γ_i. When π is known and μ is unknown, conditioned on the i-th sequence being the outlier, $i = 1, \ldots, M$, we compute the generalized likelihood of y^{Mn} by replacing μ in (3.1) with its maximum likelihood

(ML) estimate $\hat{\mu}_i \triangleq \gamma_i$, as

$$\hat{p}_i^{\text{typ}}\left(y^{Mn}\right) \;=\; L_i\left(y^{Mn}, \hat{\mu}_i,\, \pi\right). \tag{3.5}$$

Similarly, when neither μ nor π is known, we compute the generalized likelihood of y^{Mn} by replacing the μ and π in (3.1) with their ML estimates $\hat{\mu}_i \triangleq \gamma_i$, and $\hat{\pi}_i \triangleq \frac{\sum_{k\neq i}\gamma_k}{M-1}$, $i = 1, \ldots, M$, as

$$\hat{p}_i^{\text{univ}}\left(y^{Mn}\right) \;=\; L_i\left(y^{Mn}, \hat{\mu}_i,\, \hat{\pi}_i\right). \tag{3.6}$$

Finally, we decide upon the sequence corresponding to the largest generalized likelihood to be the outlier. Using (3.5), (3.6), the GL tests in the two cases can be described respectively as

$$\delta\big(y^{Mn}\big) \;=\; \underset{i=1,\ldots,M}{\operatorname{argmax}}\; \hat{p}_i^{\text{typ}}\left(y^{Mn}\right) \tag{3.7}$$

when only π is known, and

$$\delta\big(y^{Mn}\big) \;=\; \underset{i=1,\ldots,M}{\operatorname{argmax}}\; \hat{p}_i^{\text{univ}}\left(y^{Mn}\right) \tag{3.8}$$

when neither μ nor π is known. In (3.7) and (3.8), should there be multiple maximizers, we pick one of them arbitrarily. Using the identity in (2.4), it is straightforward to show that when only π is known, the GL test in (3.7) is equivalent to

$$\begin{aligned}\delta\big(y^{Mn}\big) \;&=\; \underset{i=1,\ldots,M}{\operatorname{argmin}}\; H\left(\gamma_i\right) + \sum_{j\neq i}\left[H\left(\gamma_j\right) + D\left(\gamma_j\|\pi\right)\right] \\ &=\; \underset{i=1,\ldots,M}{\operatorname{argmax}}\; D(\gamma_i\|\pi),\end{aligned} \tag{3.9}$$

and when neither π nor μ is known, the test in (3.8) is equivalent to

$$\begin{aligned}\delta\big(y^{Mn}\big) &= \underset{i=1,\ldots,M}{\operatorname{argmin}} H\left(\gamma_i\right) + \sum_{j\neq i}\left[H\left(\gamma_j\right) + D\big(\gamma_j\big\|\tfrac{\sum_{k\neq i}\gamma_k}{M-1}\big)\right] \\ &= \underset{i=1,\ldots,M}{\operatorname{argmin}} \sum_{j\neq i} D\big(\gamma_j\big\|\tfrac{\sum_{k\neq i}\gamma_k}{M-1}\big).\end{aligned} \tag{3.10}$$

3.1.2 Performance of Generalized Likelihood Test

Our first theorem for models with one outlier characterizes the optimal exponent for the maximal error probability when both μ and π are known, and when only π is known.

Theorem 1. *When μ and π are both known, the optimal exponent for the maximal error probability is equal to*

$$2B(\mu, \pi). \tag{3.11}$$

Furthermore, the error exponent in (3.11) is achievable by the GL test in (3.7), which uses only the knowledge of π.

Proof. Since we consider the error exponent as n goes to infinity, while M and hence the number of hypotheses is fixed, the ML test, which maximizes the error exponent for the average error probability (averaged over all hypotheses), will also achieve the best error exponent for the maximal error probability. In particular, for any $y^{Mn} = \left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \in \mathcal{Y}^{Mn}$, with $\gamma_{\boldsymbol{y}^{(i)}} = \gamma_i$, $i = 1, \ldots, M$, conditioned on the i-th sequence being the outlier, applying the identity in (2.4), it now follows from (3.1) that the ML test is

$$\delta(y^{Mn}) \;=\; \underset{i=1,\ldots,M}{\operatorname{argmin}} \; U_i(y^{Mn}),$$

where for each $i = 1, \ldots, M$,

$$U_i(y^{Mn}) \;\triangleq\; D\left(\gamma_i \| \mu\right) + \sum_{j \neq i} D\left(\gamma_j \| \pi\right). \tag{3.12}$$

By the symmetry of the problem, it is clear that $\mathbb{P}_i\left\{\delta \neq i\right\}$ is the same for every $i = 1, \ldots, M$; hence,

$$\max_{i=1,\ldots,M} \mathbb{P}_i\left\{\delta \neq i\right\} \;=\; \mathbb{P}_1\left\{\delta \neq 1\right\}.$$

It now follows from

$$\mathbb{P}_1\left\{\delta \neq 1\right\} \;=\; \mathbb{P}_1\left(\cup_{j \neq 1}\{U_1 \geq U_j\}\right), \tag{3.13}$$

that

$$\mathbb{P}_1\left\{U_1 \geq U_2\right\} \;\leq\; \mathbb{P}_1\left\{\delta \neq 1\right\} \;\leq\; \sum_{j=2}^{M} \mathbb{P}_1\left\{U_1 \geq U_j\right\}. \tag{3.14}$$

Next, we get from (3.12) that

$$\begin{aligned}\mathbb{P}_1\left\{U_1 \geq U_2\right\} \;&=\; \mathbb{P}_1\{D\left(\gamma_1\|\mu\right) + D\left(\gamma_2\|\pi\right) \\ &\qquad \geq D\left(\gamma_1\|\pi\right) + D\left(\gamma_2\|\mu\right)\}.\end{aligned}$$

Applying Lemma 1 with $J = 2$, $p_1 = \mu$, $p_2 = \pi$, and

$$\begin{aligned}E = \Big\{ (q_1, q_2) :\; &D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) \\ &\geq D\left(q_1\|\pi\right) + D\left(q_2\|\mu\right) \Big\},\end{aligned}$$

we get that the exponent for $\mathbb{P}_1\left\{U_1 \geq U_2\right\}$ is given by the value of the following optimization problem

$$\min_{q_1, q_2 \,\in\, \mathcal{P}(\mathcal{Y})} \Big(D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) \Big), \tag{3.15}$$

where the minimum above is over the set of q_1, q_2 such that

$$D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) \;\geq\; D\left(q_1\|\pi\right) + D\left(q_2\|\mu\right).$$

Note that the objective function in (3.15) is convex in (q_1, q_2), and the constraint is linear in (q_1, q_2). It then follows that the optimization problem in (3.15) is convex. Consequently, strong duality holds for the optimization problem (3.15) [33]. Then by solving the Lagrangian dual of (3.15), its solution can be easily computed to be $2B(\mu, \pi)$.

By the symmetry of the problem, the exponents of $\mathbb{P}_1\left\{U_1 \geq U_i\right\}$, $i \neq 1$, are the same, i.e., for every $i = 2, \ldots, M$, we get

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1\left\{U_1 \geq U_i\right\} \;=\; 2B(\mu, \pi). \tag{3.16}$$

From (3.14) and (3.16), using the union bound and that $\lim_{n\to\infty} \frac{\log M}{n} = 0$, we get that

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1\left\{\delta \neq 1\right\} \;=\; 2B(\mu, \pi). \tag{3.17}$$

It is now left to prove that when only π is known, the GL test in (3.7) and (3.9) also achieves the optimal error exponent $2B(\mu, \pi)$.

For each $i = 1, \ldots, M$, denote the test statistic in (3.9) as

$$U_i^{\text{typ}} \triangleq D(\gamma_i\|\pi).$$

It follows from the same argument leading to (3.17) that

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1\{\delta' \neq 1\}\\ &= \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1\Big\{U_1^{\text{typ}} \ \leq \ U_2^{\text{typ}}\Big\}.\end{aligned} \tag{3.18}$$

The exponent on the right side of (3.18) can be computed by applying Lemma 1 with $J = 2$, $p_1 = \mu$, $p_2 = \pi$, and

$$E \ = \ \big\{\, (q_1, q_2) : \ D(q_2\|\pi) \geq D(q_1\|\pi)\big\}$$

to be

$$\min_{\substack{q_1,q_2\in\mathcal{P}(\mathcal{Y})\\ D(q_2\|\pi) \ \geq \ D(q_1\|\pi)}} \Big(D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right)\Big) \tag{3.19}$$

The optimal value of (3.19) can be computed as follows:

$$\min_{\substack{q_1,q_2\in\mathcal{P}(\mathcal{Y})\\ D(q_2\|\pi)\geq D(q_1\|\pi)}} \Big(D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right)\Big) \tag{3.20}$$

$$\geq \ \min_{q_1} \ \Big(D\left(q_1\|\mu\right) + D\left(q_1\|\pi\right)\Big) \tag{3.21}$$

$$= \ 2B(\mu,\pi), \tag{3.22}$$

where the inequality in (3.21) stems from substituting the constraint in (3.20) into the objective function, and the equality in (3.22) follows from Lemma 2. Since the minimum in (3.21) is achieved by $q_1 = q^\star$ in (2.11) with $p_1 = \mu$, $p_2 = \pi$, and $q_1 = q_2 = q^\star$ satisfy the constraint in (3.20), the inequality in (3.21) is in fact an equality. □

Remark 1. It is interesting to note that when only μ is known, one can also achieve the optimal error exponent in (3.11) using a different test that will be presented in Section 3.5. However, we do not yet know if the corresponding version of the GL test, wherein the π in (3.1) is replaced with $\hat{\pi}_i = \frac{\sum_{k\neq i}\gamma_k}{M-1}$, $i = 1, \ldots, M$, is optimal.

Consequently, in the completely universal setting, when nothing is known about μ and π except that $\mu \neq \pi$, and both μ and π have full supports, it holds that for any universal test δ,

$$\alpha\big(\delta, (\mu,\pi)\big) \ \leq \ \ 2B(\mu,\pi). \tag{3.23}$$

Given the second assertion in Theorem 1, it might be tempting to think that it would be possible to design a test to achieve the optimal error exponent of $2B(\mu,\pi)$ universally when neither μ nor π is known. Our first example shows that such a goal cannot be fulfilled, and hence we need to be content with a lesser goal.

Example 1: Consider the model with $M=3$, and a distinct pair of distributions $p \neq \overline{p}$ on $\mathcal{Y}$ with full supports. We now show that there cannot exist a universal test that achieves the optimal error exponent of $2B(\mu,\pi)$ *even just for the two models* when $\mu = p, \pi = \overline{p}$, and when $\mu = \overline{p}, \pi = p$, both of which have $2B(\mu,\pi) = 2B(p,\overline{p})$. To this end, let us look at the region when a universal test δ decides that the first sequence is the outlier, i.e., $A_1 = \{y^{3n} : \delta(y^{3n}) = 1\}$. Let $\mathbb{P}_{p,\overline{p},\overline{p}}$ denote the distribution corresponding to the first hypothesis of the first model, i.e., when $\boldsymbol{y}^{(1)}$ are i.i.d. according to p, and $\boldsymbol{y}^{(2)}$ and $\boldsymbol{y}^{(3)}$ are i.i.d. according to $\overline{p}$. Similarly, let $\mathbb{P}_{p,\overline{p},p}$ denote the distribution corresponding to the second hypothesis of the second model, i.e., when $\boldsymbol{y}^{(2)}$ are i.i.d. according to $\overline{p}$, and $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(3)}$ are i.i.d. according to p. Suppose that δ achieves the best error exponent of $2B(p,\overline{p})$ for the first model when $\mu = p, \pi = \overline{p}$. Then, it must hold that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_{p,\overline{p},\overline{p}}\{A_1^c\} \;\geq\; 2B(p,\overline{p}). \tag{3.24}$$

It now follows from (3.24) and the classic result of Hoeffding [6] in binary hypothesis testing (see, e.g., [34][Exercise 2.13 (b)]) that

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_{p,\overline{p},p}\{A_1\} \\
&\leq \Big[\min_{q(y_1,y_2,y_3)} D(q(y_1)\|p) + D(q(y_2)\|\overline{p}) \\
&\qquad\qquad + D(q(y_3)\|p) \Big]^+ \\
&\leq \Big[\min_{q(y_1,y_2,y_3)} 2B(p,\overline{p}) + D(q(y_3)\|p) \\
&\qquad\qquad - D(q(y_3)\|\overline{p}) \Big]^+ \\
&\leq \big(2B(p,\overline{p}) - D(p\|\overline{p})\big)^+ \;=\; 0,
\end{aligned} \tag{3.25}$$

where each minimum on the right side above is taken over the set of $q(y_1,y_2,y_3)$ such that

$$D(q(y_1)\|p) + D(q(y_2)\|\overline{p}) + D(q(y_3)\|\overline{p}) \leq 2B(p,\overline{p}).$$

The last equality in (3.25) follows from Lemma 2 in Chapter 2. Consequently, the test cannot yield even a *positive* error exponent for the second model when $\mu = \overline{p}, \pi = p$.

Remark 2. It is interesting to contrast this example for outlier hypothesis testing with the results (Theorems 2 and 3 in [35]) for universal coding over discrete memoryless channels (DMCs). Specifically, Theorems 2 and 3 in [35] establish that the optimal error exponent at zero rate is universally achieved for all DMCs, whereas the optimal error exponent $2B(\mu,\pi)$ for outlier hypothesis testing here *cannot* be universally achieved. The difference between these two results stems from the following distinctions between the nature of these two problems. First, in universal coding, the encoder and decoder are jointly optimized to achieve universality. On the other hand, in outlier hypothesis testing, when properly interpreted, only the decoding is allowed to be optimized, while the encoding scheme is fixed by the structure of the distributions of observations among all hypotheses, and cannot be chosen. Second, the zero-rate error exponent in [35] applies only for the case when the number of messages *grows* to infinity with the blocklength sub-exponentially. In contrast, the number of hypotheses in outlier hypothesis testing is fixed and does not grow with the number of observations in each sequence.

To summarize, the results in [35] cannot be applied to our problem. Had the results in [35] been applicable, Theorems 2 and 3 in [35] would have implied that the optimal error exponent $2B(\mu,\pi)$ is achieved universally for outlier hypothesis testing as well. However, Example 1 proves otherwise.

Example 1 shows explicitly that when neither μ nor π is known, it is impossible to construct a test that achieves $2B(\mu,\pi)$ universally. In fact, the example shows that had we insisted on achieving the best error exponent of $2B(\mu,\pi)$ for some pairs of μ,π, it might not be possible to achieve even *positive* error exponents for some other pairs of μ,π. This motivates us to seek instead a test that yields just a positive (no matter how small) error exponent $\alpha(\delta,(\mu,\pi))>0$ for *every* μ,π, $\mu\neq\pi$, i.e., achieving universally exponential consistency. One of our main contributions in this chapter is to show that GL tests are indeed *universally exponentially consistent* under various settings, including the current single outlier setting for every *fixed* M.

Theorem 2. *The GL test δ in (3.8) is universally exponentially consistent. Furthermore, for every pair of distributions $\mu,\pi,\mu\neq\pi$, it holds that*

$$\alpha\big(\delta,(\mu,\pi)\big)=\min_{q_1,\ldots,q_M} D(q_1\|\mu)+D(q_2\|\pi) + \ldots + D(q_M\|\pi), \tag{3.26}$$

where the minimum above is over the set of $(q_1, \ldots, q_M)$ *such that*

$$\sum_{j\neq 1} D\left(q_j \,\middle\|\, \tfrac{\sum_{k\neq 1} q_k}{M-1}\right) \geq \sum_{j\neq 2} D\left(q_j \,\middle\|\, \tfrac{\sum_{k\neq 2} q_k}{M-1}\right). \tag{3.27}$$

Proof. For each $i = 1, \ldots, M$, denote the test statistic in (3.10) as

$$U_i^{\text{univ}} \triangleq \sum_{j\neq i} D\big(\gamma_j \big\| \tfrac{\sum_{k\neq i} \gamma_k}{M-1}\big). \tag{3.28}$$

The same argument leading to (3.17) yields that

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1\{\delta \neq 1\} \\ &= \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_1\Big\{U_1^{\text{univ}} \;\geq\; U_2^{\text{univ}}\Big\}. \end{aligned} \tag{3.29}$$

By applying Lemma 1 with $J = M,\ p_1 = \mu,\ p_j = \pi,\ j = 2, \ldots, M$, and

$$\begin{aligned} E = \Bigg\{ (q_1, \ldots, q_M) \,:\, &\sum_{j\neq 1} D\Big(q_j \,\Big\|\, \tfrac{\sum_{k\neq 1} q_k}{M-1}\Big) \\ &\geq \sum_{j\neq 2} D\Big(q_j \,\Big\|\, \tfrac{\sum_{k\neq 2} q_k}{M-1}\Big)\Bigg\}, \end{aligned} \tag{3.30}$$

the exponent on the right side of (3.29) can be computed to be

$$\min_{(q_1,\ldots,q_M)\,\in\, E} D\left(q_1 \| \mu\right) + D\left(q_2 \| \pi\right) + \ldots + D\left(q_M \| \pi\right). \tag{3.31}$$

Unlike the convex optimization problems in (3.15) and (3.19), the optimization problem in (3.31) for the completely universal setting is much more complicated, and a closed-form solution is not available. However, we show that the value of (3.31) is strictly positive for every $\mu \neq \pi$. In particular, it is not hard to see that the objective function is continuous in $q_1, \ldots, q_M$ and the constraint set E is compact. Therefore the minimum in (3.31) is achieved by some $(q_1^\star, \ldots, q_M^\star) \in E$. Note that the objective function in (3.31) is always non-negative. In order for the objective function in (3.31) to be zero, the minimizing $(q_1^\star, \ldots, q_M^\star)$ has to satisfy that $q_1^\star = \mu,\ q_i^\star = \pi,\ i = 2, \ldots, M$. Since this collection of distributions is not in the constraint set E in (3.30), we get that the optimal value of (3.31) is strictly positive for every $\mu \neq \pi$. □

Note that for any fixed $M \geq 3$, $\epsilon > 0$, regardless of which sequence is the outlier, it holds

that the random empirical distributions $(\gamma_1, \ldots, \gamma_M)$ satisfy

$$\lim_{n\to\infty} \mathbb{P}_i \left\{ \left\| \tfrac{1}{M} \sum_{j=1}^{M} \gamma_j - \left(\tfrac{1}{M}\mu + \tfrac{M-1}{M}\pi \right) \right\|_1 > \epsilon \right\} = 0, \tag{3.32}$$

where $\|\cdot\|_1$ denotes the 1-norm of the argument distribution. Since $\frac{1}{M}\mu + \frac{M-1}{M}\pi \to \pi$ as $M \to \infty$, heuristically speaking, a consistent estimate of the typical distribution can readily be obtained asymptotically in M from the entire observations before deciding upon which sequence is the outlier. This observation and the second assertion of Theorem 1 motivate our study of the asymptotic performance (achievable error exponent) of the GL test in (3.8) when $M \to \infty$ (after having taken the limit as n goes to infinity first).

Our last result for models with one outlier shows that for the completely universal setting, the GL test in (3.8) is *asymptotically efficient*, i.e., as $M \to \infty$, it achieves the optimal error exponent in (3.11) corresponding to the case in which *both μ and π are known*.

Theorem 3. *For each $M \geq 3$, the exponent for the maximal error probability achievable by the GL test δ in (3.8) is lower bounded by*

$$\min_{\substack{q \in \mathcal{P}(\mathcal{Y}) \\ D(q\|\pi) \leq \frac{1}{M-1}\left(2B(\mu,\pi)+C_\pi\right)}} 2\,B(\mu\,,\,q)\,, \tag{3.33}$$

where $C_\pi \triangleq -\log\left(\min_{y\in\mathcal{Y}} \pi(y)\right) < \infty$ by the fact that π has a full support.

The lower bound for the error exponent in (3.33) is nondecreasing in $M \geq 3$. Furthermore, as $M \to \infty$, this lower bound converges to the optimal error exponent $2B(\mu, \pi)$; hence, the GL test δ in (3.8) achieves the optimal error exponent asymptotically as the number of sequences approaches infinity, i.e.,

$$\lim_{M\to\infty} \alpha\big(\delta, (\mu, \pi)\big) \;=\; 2B(\mu, \pi), \tag{3.34}$$

which from Theorem 1 is equal to the optimal error exponent when both μ and π are known.

Proof. By the continuity of the objective function on the right side of (3.26) and the compactness of the constraint set (3.27), for each $M \geq 3$, the optimal value on the right side of (3.26), denoted by $V^\star$, is achieved by some $(q_1^\star, \ldots, q_M^\star)$. It then follows from (3.26) and (3.27) that

$$
\begin{aligned}
V^\star \geq\; & D(q_1^\star \,\|\, \mu) + \sum_{j\neq 1} D\left(q_j^\star \,\|\, \pi\right) \\
& - \sum_{j\neq 1} D\Big(q_j^\star \,\Big\|\, \tfrac{\sum_{k\neq 1} q_k^\star}{M-1}\Big) + \sum_{j\neq 2} D\Big(q_j^\star \,\Big\|\, \tfrac{\sum_{k\neq 2} q_k^\star}{M-1}\Big) \\
=\; & D(q_1^\star \,\|\, \mu) + \sum_{j\neq 2} D\Big(q_j^\star \,\Big\|\, \tfrac{\sum_{k\neq 2} q_k^\star}{M-1}\Big) \\
& + \sum_{j\neq 1} \sum_{y\in\mathcal{Y}} q_j^\star(y) \log\left(\frac{\frac{1}{M-1}\sum_{k\neq 1} q_k^\star(y)}{\pi}\right) \\
=\; & D(q_1^\star \,\|\, \mu) + \sum_{j\neq 2} D\Big(q_j^\star \,\Big\|\, \tfrac{\sum_{k\neq 2} q_k^\star}{M-1}\Big) \\
& + (M-1) D\Big(\tfrac{\sum_{k\neq 1} q_k^\star}{M-1} \,\Big\|\, \pi\Big) \\
\geq\; & D(q_1^\star \,\|\, \mu) + D\Big(q_1^\star \,\Big\|\, \tfrac{\sum_{k\neq 2} q_k^\star}{M-1}\Big) \\
\geq\; & 2B\Big(\mu\,,\, \tfrac{\sum_{k\neq 2} q_k^\star}{M-1}\Big) \\
=\; & 2B\Big(\mu\,,\, \tfrac{q_1^\star}{M-1} + \tfrac{M-2}{M-1}\big(\tfrac{\sum_{k=3}^M q_k^\star}{M-2}\big)\Big),
\end{aligned}
\tag{3.35}
$$

where the last inequality follows Lemma 2.

On the other hand, it follows from (3.23) that the value on the right side of (3.26), $V^\star$, satisfies

$$
\begin{aligned}
2B(\mu,\pi) \;\geq\; & V^\star \\
=\; & D\left(q_1^\star \,\|\, \mu\right) + \sum_{j\neq 1} D\left(q_j^\star \,\|\, \pi\right) \\
\geq\; & \sum_{j=3}^{M} D\left(q_j^\star \,\|\, \pi\right) \\
\geq\; & (M-2)\, D\Big(\tfrac{1}{M-2}\sum_{k=3}^M q_k^\star \,\Big\|\, \pi\Big),
\end{aligned}
\tag{3.36}
$$

where the last inequality follows from the convexity of relative entropy.

Combining (3.35) and (3.36), we get that the value $V^\star$ on the right side of (3.26) is lower bounded by

$$
\min_{\substack{q_1, q\in\mathcal{P}(\mathcal{Y}) \\ (M-2)D(q\|\pi)\leq 2B(\mu,\pi)}} \quad 2B\Big(\mu\,,\, \tfrac{1}{M-1}q_1 + \tfrac{M-2}{M-1}q\Big). \tag{3.37}
$$

Note that the constraint in (3.37) can be equally written as

$$D\left(q_1\|\pi\right)+(M-2)D\left(q\|\pi\right) \;\leq\; 2B(\mu,\pi)+D\left(q_1\|\pi\right).$$

Also by the convexity of relative entropy, it follows that

$$\begin{aligned} D\left(q_1\|\pi\right)+&(M-2)D\left(q\|\pi\right) \;\geq \\ &(M-1)D\Big(\tfrac{q_1+(M-2)q}{M-1}\Big\|\pi\Big). \end{aligned}$$

As a result, the optimal value of (3.37) is further lower bounded by the optimal value of

$$\min_{\substack{q_1,q\,\in\,\mathcal{P}(\mathcal{Y}) \\ (M-1)D\left(\frac{1}{M-1}q_1+\frac{M-2}{M-1}q\|\pi\right) \\ \leq\; 2B(\mu,\pi)+D(q_1\|\pi)}} 2B\Big(\mu\,,\,\tfrac{1}{M-1}q_1+\tfrac{M-2}{M-1}q\Big). \tag{3.38}$$

By the fact that π has full support, it holds that

$$D\left(q_1\|\pi\right) \;\leq\; -\log\Big(\min_{y\in\mathcal{Y}}\pi(y)\Big) \;=\; C_\pi \;\leq\; \infty. \tag{3.39}$$

Proceeding from (3.38), by using (3.39), we get that the optimal value of (3.26) is lower bounded by

$$\min_{\substack{q'\,\in\,\mathcal{P}(\mathcal{Y}) \\ D(q'\|\pi)\,\leq\,\frac{1}{M-1}(2B(\mu,\pi)+C_\pi)}} 2\,B\left(\mu\,,\,q'\right). \tag{3.40}$$

For any $\mu,\pi\in\mathcal{P}(\mathcal{Y})$ with full supports, it holds that

$$\lim_{M\to\infty}\frac{1}{M-1}\big(2B(\mu,\pi)+C_\pi\big) \;=\; 0.$$

This and the continuity of $D\left(q\|\pi\right)$ in q (π has a full support) establish (3.34): the asymptotic optimality of the GL test in the regime of large number of sequences.

Furthermore, for any $\mu,\pi\in\mathcal{P}(\mathcal{Y})$, $\mu\neq\pi$, the value of $\frac{1}{M-1}(2B(\mu,\pi)+C(\pi))$ is strictly decreasing with M. Consequently, the feasible set in (3.33) is nonincreasing with M, and hence the optimal value of (3.33) is nondecreasing with M. □

Example 2: We now provide some numerical results for an example with $\mathcal{Y}=\{0,1\}$. Specifically, the three plots in Figure 3.1 are for three pairs of outlier and typical distributions being $\mu=(p(0)=0.3,\ p(1)=0.7)$, $\pi=(0.7,0.3)$; $\mu=(0.35,0.65)$, $\pi=(0.65,0.35)$; and

$\mu = (0.4, 0.6)$, $\pi = (0.6, 0.4)$, respectively. Each horizontal line corresponds to $2B(\mu, \pi)$, and each curve line corresponds to the lower bound in (3.33) for the error exponent achievable by the GL test in (3.8). As shown in these plots, the lower bounds converge to $2B(\mu, \pi)$ as $M \to \infty$, i.e., the GL test in (3.8) is asymptotically optimal for all three pairs of μ, π, and, indeed, for all $\mu \neq \pi$.

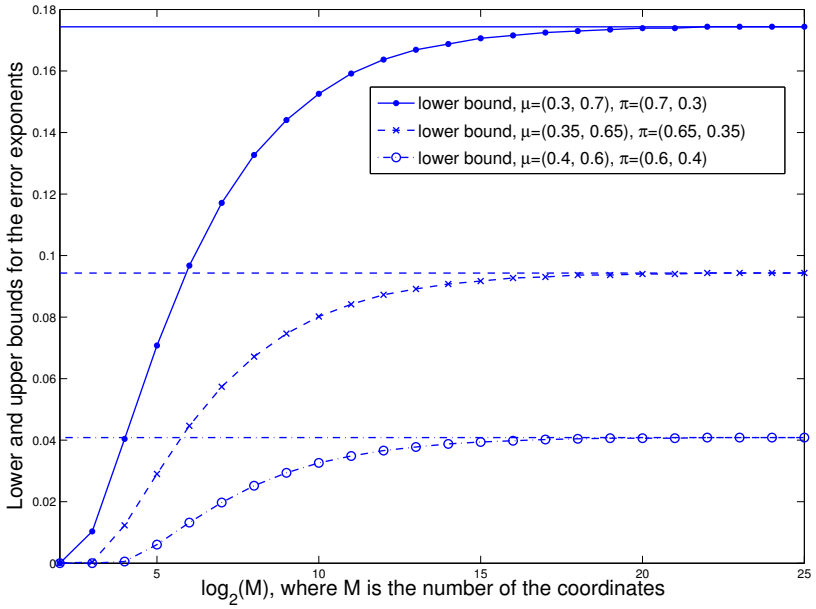

Figure 3.1: Lower and upper bounds for the achievable error exponent of the GL test

3.2 At Most One Outlier Sequence

A natural question that arises at this point is what would happen if it is also possible that no outlier is present. To answer this question, we now consider models that append an additional *null* hypothesis with no outlier to the previous models consider in Section 3.1. In particular, under the null hypothesis, the joint distribution of all the observations is given by

$$p_0\big(y^{Mn}\big) \;=\; \prod_{k=1}^{n} \prod_{i=1}^{M} \pi\big(y_k^{(i)}\big).$$

A universal test $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$ will now also accommodate for an additional decision for the null hypothesis. Correspondingly, the maximal error probability is now

computed with the inclusion of the null hypothesis according to

$$e\big(\delta,(\mu,\pi)\big) \triangleq \max_{i=0,1,\ldots,M} \sum_{y^{Mn}:\ \delta(y^{Mn})\ \neq\ i} p_i\big(y^{Mn}\big).$$

With just one additional null hypothesis, contrary to the previous models with one outlier, it becomes impossible to achieve universally exponential consistency *even with the knowledge of the typical distribution.*

Proposition 4. *For the setting with the additional null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known.*

Proof. The proposition follows as a special case of the second assertion of Theorem 10, the proof of which is deferred to Section 3.4 □

In typical applications such as environment monitoring and fraud detection, the consequence of a missed detection of the outlier can be much more catastrophic than that of a false positive. In addition, Proposition 4 tells us that there cannot exist a universal test that yields exponential decays for both the conditional probability of false positive (under the null hypothesis) and the conditional probabilities of missed detection (under all non-null hypotheses). Consequently, it is natural to look for a universal test fulfilling a lesser objective: attaining universally exponential consistency for conditional error probabilities under *all the non-null* hypotheses, while seeking *only* universal consistency for the conditional error probability under the null hypothesis. We now show that such a test can be obtained by slightly modifying the GL test in (3.8). Furthermore, in addition to achieving universal consistency under the null hypothesis, this new test achieves the same exponent as in (3.26), (3.27) in Theorem 2 for the conditional error probabilities under *all* non-null hypotheses.

3.2.1 Proposed Universal Test

We modify the previous test in (3.8) to allow for the possibility of deciding for the null hypothesis as follows:

$$\delta(y^{Mn}) \ = \ \begin{cases} \underset{i=1,\ldots,M}{\arg\max}\ \hat{p}_i^{\text{univ}}(y^{Mn}), & \text{if} \quad \max\limits_{j\neq k} \frac{1}{n}\big(\log \hat{p}_j^{\text{univ}}(y^{Mn}) \\ & \qquad - \log \hat{p}_k^{\text{univ}}(y^{Mn})\big) > \lambda_n, \\ 0, & \text{otherwise,} \end{cases} \tag{3.41}$$

where $\lambda_n = \Theta(\frac{\log n}{n})$ and the ties in the first case of (3.41) are broken arbitrarily. Using the identity in (2.4), it is straightforward to show that test in (3.41) can be equivalently written

as

$$\delta(y^{Mn}) \;=\; \begin{cases} \underset{i=1,\ldots,M}{\arg\min} \sum\limits_{k\neq i} D\left(\gamma_k \,\middle\|\, \frac{\sum_{l\neq i}\gamma_l}{M-1}\right), & \text{if} \quad \max\limits_{j\neq j'}\left[\sum\limits_{k\neq j} D\left(\gamma_k\middle\|\frac{\sum_{l\neq j}\gamma_l}{M-1}\right)\right. \\ & \left. -\sum\limits_{k\neq j'} D\left(\gamma_k\middle\|\frac{\sum_{l\neq j'}\gamma_l}{M-1}\right)\right] > \lambda_n, \\ 0, & \text{otherwise.} \end{cases} \tag{3.42}$$

3.2.2 Performance of Proposed Test

Theorem 5. *For every pair of distributions $\mu, \pi,\ \mu \neq \pi$, the test in (3.41) yields a positive exponent for the conditional probability of error under every non-null hypothesis $i = 1, \ldots, M$, and a vanishing conditional probability of error under the null hypothesis. In particular, the achievable error exponent under every non-null hypothesis is the same as that given in (3.26), (3.27), i.e., for each $i = 1, \ldots, M$, the test in (3.41) achieves*

$$\begin{aligned} &\lim_{n\to\infty} -\frac{1}{n}\log\left(\mathbb{P}_i\left\{\delta \neq i\right\}\right) \\ &\quad = \min_{q_1,\ldots,q_M} D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) + \ldots + D\left(q_M\|\pi\right), \end{aligned} \tag{3.43}$$

where the minimum above is over the set of $(q_1, \ldots, q_M)$ satisfying (3.27). In addition, the test also yields that

$$\lim_{n\to\infty} \mathbb{P}_0\left\{\delta \neq 0\right\} \;=\; 0. \tag{3.44}$$

Proof. We start by establishing universal consistency of the test under the null hypothesis. Applying the identity in (2.4) to the test statistics in (3.41), it holds that

$$\begin{aligned} \mathbb{P}_0\{\delta \neq 0\} \;&\leq\; \mathbb{P}_0\Big(\cup_{j=1}^{M}\left\{U_j^{\text{univ}} \geq \lambda_n\right\}\Big) \\ &\leq\; \sum_{j=1}^{M} \mathbb{P}_0\left\{U_j^{\text{univ}} \geq \lambda_n\right\} \\ &=\; M\mathbb{P}_0\left\{U_1^{\text{univ}} \geq \lambda_n\right\}, \end{aligned} \tag{3.45}$$

where U_j^{univ} is defined in (3.28), and the last equality follows from the fact that all $\boldsymbol{y}^{(i)}$, $i = 1, \ldots, M$, are identically distributed according to π.

We now proceed to bound $\mathbb{P}_0\{U_1^{\text{univ}} \geq \lambda_n\}$ as follows:

$$\begin{aligned}
&\mathbb{P}_0\{U_1^{\text{univ}} \geq \lambda_n\} \\
&= \mathbb{P}_0\Big\{ \sum_{j\neq 1} D\left(\gamma_j \Big\| \tfrac{\sum_{k\neq 1}\gamma_k}{M-1}\right) \geq \lambda_n \Big\} \\
&= \mathbb{P}_0\Big\{ \sum_{j\neq 1} D\left(\gamma_j \| \pi\right) - (M-1) D\left(\tfrac{\sum_{k\neq 1}\gamma_k}{M-1} \Big\| \pi\right) \geq \lambda_n \Big\} \\
&\leq \mathbb{P}_0\Big\{ \sum_{j\neq 1} D\left(\gamma_j \| \pi\right) \geq \lambda_n \Big\} \\
&\leq \mathbb{P}_0\left(\cup_{j\neq 1} \Big\{ D\left(\gamma_j \| \pi\right) \geq \frac{1}{M-1}\lambda_n \Big\} \right) \\
&\leq (M-1)\mathbb{P}_0\Big\{ D\left(\gamma_2 \| \pi\right) \geq \frac{1}{M-1}\lambda_n \Big\}, \qquad (3.46)
\end{aligned}$$

where the first inequality follows from the non-negativity of the relative entropy, and the last inequality follows from the fact that all $\boldsymbol{y}^{(j)}$, $j \neq 1$, are identically distributed according to π. By the fact that the set of all possible empirical distributions of $(y_1, \ldots, y_n)$ is upper bounded by $(n+1)^{|\mathcal{Y}|}$ (cf. [30][Theorem 11.1.1]), and (2.4), we get that

$$\begin{aligned}
&\mathbb{P}_0\Big\{ D\left(\gamma_2 \| \pi\right) \geq \frac{1}{M-1}\lambda_n \Big\} \\
&\qquad \leq \ (n+1)^{|\mathcal{Y}|} \exp(-\frac{n}{M-1}\lambda_n). \qquad (3.47)
\end{aligned}$$

It then follows from (3.45), (3.46) and (3.47) that

$$\mathbb{P}_0\{\delta \neq 0\} \leq M^2 \exp\Big\{ -\frac{n}{M-1}\lambda_n + |\mathcal{Y}| \log(n+1) \Big\}. \qquad (3.48)$$

By choosing $\lambda_n = 2(M-1)|\mathcal{Y}|\frac{\log(n+1)}{n}$, we get from (3.48) that

$$\lim_{n\to\infty} \mathbb{P}_0\{\delta \neq 0\} \ = \ 0.$$

Next we treat the exponent for the conditional probability of error under every non-null hypothesis. In particular, by the symmetry of the test (3.41) among all the M non-null hypotheses, it suffices to consider the conditional error probability under just the first

hypothesis, which can be upper bounded as follows:

$$\begin{aligned}\mathbb{P}_1\{\delta \neq 1\} &\leq \mathbb{P}_1\left(\cup_{j\neq 1}\Big\{U_1^{\text{univ}} \geq U_j^{\text{univ}} - \lambda_n\Big\}\right)\\ &\leq \sum_{j\neq 1}\mathbb{P}_1\Big\{U_1^{\text{univ}} \geq U_j^{\text{univ}} - \lambda_n\Big\}\\ &\leq (M-1)\mathbb{P}_1\Big\{U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_n\Big\}. \end{aligned} \tag{3.49}$$

For an arbitrary $\lambda_0 > 0$, as $\lambda_n \to 0$, it holds that $\lambda_n \leq \lambda_0$ for n sufficiently large and hence that

$$\mathbb{P}_1\Big\{U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_n\Big\} \leq \mathbb{P}_1\Big\{U_1^{\text{univ}} \geq U_2^{\text{univ}} - \lambda_0\Big\}. \tag{3.50}$$

The exponent of the right side of (3.50) can be computed by applying Lemma 1 with $J = M$, $p_1 = \mu$, $p_j = \pi$, $j = 2, \ldots, M$ and (cf.(3.28))

$$\begin{aligned}E(\lambda_0) \triangleq \Bigg\{ (q_1, \ldots, q_M) : &\sum_{j\neq 1} D\Big(q_j \Big\| \tfrac{\sum_{k\neq 1} q_k}{M-1}\Big)\\ &\geq \sum_{j\neq 2} D\Big(q_j \Big\| \tfrac{\sum_{k\neq 2} q_k}{M-1}\Big) - \lambda_0\Bigg\}\end{aligned}$$

to be

$$\min_{(q_1,\ldots,q_M)\in E(\lambda_0)} D(q_1\|\mu) + D(q_2\|\pi) + \ldots + D(q_M\|\pi). \tag{3.51}$$

Since λ_0 can be arbitrarily close to zero, the exponent for the left side of (3.50) is lower bounded by

$$\lim_{\lambda_0\to 0} \min_{(q_1,\ldots,q_M)\,\in\, E(\lambda_0)} D(q_1\|\mu) + D(q_2\|\pi) + \ldots + D(q_M\|\pi).$$

Let

$$\begin{aligned}E \triangleq \Bigg\{ (q_1, \ldots, q_M) : &\sum_{j\neq 1} D\Big(q_j \Big\| \tfrac{\sum_{k\neq 1} q_k}{M-1}\Big)\\ &\geq \sum_{j\neq 2} D\Big(q_j \Big\| \tfrac{\sum_{k\neq 2} q_k}{M-1}\Big)\Bigg\}.\end{aligned}$$

By the fact that $E(\lambda_0)$ is closed and compact for any $\lambda_0 > 0$, and that the objective function

in (3.51) is continous, the exponent for the left side of (3.50) is lower bounded by

$$\min_{(q_1,\ldots,q_M)\,\in\,E} D\left(q_1\|\mu\right) + D\left(q_2\|\pi\right) + \ldots + D\left(q_M\|\pi\right), \tag{3.52}$$

as required. □

Since under every non-null hypothesis, the modified test in (3.41) achieves the same exponent (the value of the optimization problem in (3.26) and (3.27)) for the conditional error probability as the GL test in (3.8) when the null hypothesis is absent, we get the following corollary by just observing that Theorem 3 was proved by finding a suitable lower bound for the value of the optimization problem in (3.26) and (3.27).

Corollary 6. *For each $M \geq 3$ and under every non-null hypothesis $i = 1, \ldots, M$, the exponent for the conditional error probability achievable by the test in (3.41) is lower bounded as*

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(\mathbb{P}_i\left\{\delta \neq i\right\}\right) \geq \min_{q\,\in\,\mathcal{P}(\mathcal{Y})} 2\,B(\mu\,,\,q)\,, \tag{3.53}$$

where the minimum above is over the set of q such that

$$D(q\|\pi) \;\leq\; \frac{1}{M-1}\big(2B(\mu,\pi) + C_\pi\big),$$

and $C_\pi \triangleq -\log\Big(\min\limits_{y\in\mathcal{Y}} \pi(y)\Big) < \infty$. Consequently, as $M \to \infty$, this lower bound converges to the optimal error exponent $2B(\mu,\pi)$, i.e., for every $i = 1, \ldots, M$, the test in (3.41) achieves

$$\lim_{M\to\infty}\;\lim_{n\to\infty} -\frac{1}{n}\;\log\left(\mathbb{P}_i\left\{\delta \neq i\right\}\right) \;=\; 2B(\mu,\pi),$$

while also yielding that

$$\lim_{n\to\infty} \mathbb{P}_0\left\{\delta \neq 0\right\} \;=\; 0.$$

3.3 Multiple Distinctly Distributed Outliers

We now generalize our results in the previous sections to models with multiple outlier sequences. With more than one outlier sequence, it may be more natural to consider models for which the different outlier sequences are distinctly distributed, and therefore our models will allow for this possibility.

We start by describing a generic model with possibly distinctly distributed outliers, the number of which is not known exactly. Specifically, it is assumed that there are up to $K \geq 1$ outliers. Note that the current model with $K = 1$ corresponds to the single outlier setting where the outlier distribution can vary according to the index of the outlier sequence. As in Section 3.1, we denote the k-th observation of the i-th sequence by $Y_k^{(i)} \in \mathcal{Y}, i = 1, \ldots, M,\ k = 1, \ldots, n$. Most of the sequences are commonly distributed according to the "typical" distribution $\pi \in \mathcal{P}(\mathcal{Y})$ except for a small (possibly empty) subset $S \subset \{1, \ldots, M\}$ of "outlier" sequences, each of which is assumed to be distributed according to an outlier distribution $\mu_i,\ i \in S$. *Nothing is known about* $\{\mu_i\}_{i=1}^M$ *and* π *except that each* $\mu_i \neq \pi,\ i = 1, \ldots, M$, *and that all* $\mu_i,\ i = 1, \ldots, M$, *and* π *have full supports.* In the following presentation, we sometimes consider the special case when all the outlier sequences are identically distributed, i.e., $\mu_i = \mu$, $i = 1, \ldots, M$.

For the hypothesis corresponding to an outlier subset $S \subset \{1, \ldots, M\}$, $|S| < \frac{M}{2}$, the joint distribution of all the observations is given by

$$\begin{aligned} p_S\left(y^{Mn}\right) &= p_S\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \\ &= \prod_{k=1}^{n} \left\{ \prod_{i \in S} \mu_i\left(y_k^{(i)}\right) \prod_{j \notin S} \pi\left(y_k^{(j)}\right) \right\}, \end{aligned} \tag{3.54}$$

where

$$\boldsymbol{y}^{(i)} = \left(y_1^{(i)}, \ldots, y_n^{(i)}\right),\ i = 1, \ldots, M.$$

We refer to the unique hypothesis corresponding to the case with *no* outlier, i.e., $S = \emptyset$, as the *null* hypothesis. In the following subsections, we shall consider different settings, each being described by a suitable set $\mathcal{S}$ comprising all possible outlier subsets.

The test for the outlier subset is done based on a *universal* rule $\delta : \mathcal{Y}^{Mn} \to \mathcal{S}$. In particular, the test δ is not allowed to be a function of $\left(\{\mu_i\}_{i=1}^M, \pi\right)$.

For a universal test, the maximal error probability, which will be a function of the test and $\left(\{\mu_i\}_{i=1}^M, \pi\right)$, is

$$e\left(\delta, \left(\{\mu_i\}_{i=1}^M, \pi\right)\right) \triangleq \max_{S \in \mathcal{S}} \sum_{y^{Mn}:\ \delta(y^{Mn}) \neq S} p_S(y^{Mn}), \tag{3.55}$$

and the corresponding error exponent is defined as

$$\alpha\left(\delta, \left(\{\mu_i\}_{i=1}^{M}, \pi\right)\right) \triangleq \lim_{n\to\infty} -\frac{1}{n} \log e\left(\delta, \left(\{\mu_i\}_{i=1}^{M}, \pi\right)\right).$$

A universal test δ is termed *universally exponentially consistent* if for every $\mu_i,\ i = 1, \ldots, M,\ \mu_i \neq \pi$, it holds that

$$\alpha\left(\delta, \left(\{\mu_i\}_{i=1}^{M}, \pi\right)\right) > 0.$$

3.3.1 Necessary Condition for Existence of Universally Exponentially Consistent Test

Our first result concerns the necessary condition for the existence of a universally exponentially consistent test when the outliers can be distinctly distributed in an arbitrary manner. In our model for this section, the assumption of a known number of outliers is in fact critical, as a lack thereof would make it impossible to construct a universally exponentially consistent test even when there are *always some* outliers.

Theorem 7. *When the outliers can be distinctly distributed, with the number of outliers being unknown, there cannot exist a universally exponentially consistent test, even when the typical distribution is known and when the null hypothesis is excluded, i.e., there is at least one outlier regardless of the hypothesis.*

Proof. Without loss of generality, we can consider the following two hypotheses. The first hypothesis has S_1 as the set of outliers, and the second hypothesis has S_2, where $S_1 \subset S_2$. It suffices to prove that even when π and $\{\mu_i\}_{i\in S_1}$ are known, there cannot exist a universally exponentially consistent test in differentiating such two hypotheses.

For any test $\delta : \mathcal{Y}^{Mn} \to \{1, 2\}$, let $\delta = 1$ denote a decision in favor of the hypothesis with S_1 being the outliers, and 2 the hypothesis with S_2. We first show that in order to distinguish between S_1 and S_2, the empirical distributions of all the sequences $\gamma_1, \ldots, \gamma_M$, π and $\{\mu_i\}_{i\in S_1}$ are sufficient statistics for the error exponent. In particular, we now show that given any test, there is another test that achieves the same error exponent with its decision being made based *only on* the empirical distributions of all M sequences, π and $\{\mu_i\}_{i\in S_1}$. To this end, for feasible empirical distributions (for certain n) $\gamma_1, \ldots, \gamma_M$, let us denote the set of all M sequences conforming to these empirical distributions by $T_{(\gamma_1,\ldots,\gamma_M)}$. Among these observation sequences, let us denote the set of M sequences for which δ decides for S_1 by $T^{1,\pi}_{(\gamma_1,\ldots,\gamma_M)}$, which may depend on π and $\{\mu_i\}_{i\in S_1}$. Now consider another test δ' which decides on one of the two hypotheses based only on $\gamma_1, \ldots, \gamma_M$, π and $\{\mu_i\}_{i\in S_1}$. Specifically, this new

test is such that for all M sequences with empirical distributions $\gamma_1, \ldots, \gamma_M$, it decides for S_1 if $|T^{1,\pi}_{(\gamma_1,\ldots,\gamma_M)}| \geq \frac{1}{2}|T_{(\gamma_1,\ldots,\gamma_M)}|$, and for S_2 otherwise. It follows from this construction of δ' that for any $\{\mu_i\}_{i=1}^M$ and π,

$$\begin{aligned}&\max\left(\mathbb{P}_{S_1}\left\{\delta' \neq S_1\right\}, \mathbb{P}_{S_2}\left\{\delta' \neq S_2\right\}\right)\\ &\quad\leq\ 2\max\left(\mathbb{P}_{S_1}\left\{\delta \neq S_1\right\}, \mathbb{P}_{S_2}\left\{\delta \neq S_2\right\}\right),\end{aligned}$$

where $\mathbb{P}_{S_1}$ and $\mathbb{P}_{S_2}$ are the distributions under the hypothesis with S_1 and S_2 being the set of outliers, respectively. Consequently, the error exponent achievable by δ' is the same as that achievable by δ for any $\{\mu_i\}_{i=1}^M$ and π, where $\mu_i \neq \pi, i = 1, \ldots, M$.

We now consider tests that only depend on the empirical distributions of all the sequences $\gamma_1, \ldots, .\gamma_M$, π and $\{\mu_i\}_{i=1}^M$. Let assume that for any fixed π and $\{\mu_i\}_{i\in S_1}$, there exists $\epsilon = \epsilon\left(\pi, \{\mu_i\}_{i\in S_1}\right) > 0$ such that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1\left\{\delta \neq 1\right\} \ > \ \epsilon, \tag{3.56}$$

where $\mathbb{P}_1$ is the distribution under the hypothesis with S_1 being the outliers. It now follows from (3.56) and Lemma 1 that the set A of all M tuples $\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \in \mathcal{Y}^{Mn}$ whose empirical distributions $(\gamma_1, \ldots, \gamma_M)$ lie in the following set

$$E \triangleq \left\{(q_1, \ldots, q_M) : \sum_{i\in S_1} D\left(q_i \| \mu_i\right) + \sum_{j\notin S_1} D\left(q_i \| \pi\right) \leq \frac{\epsilon}{2}\right\} \tag{3.57}$$

must be such that

$$A \subseteq \left\{\delta = 1\right\}. \tag{3.58}$$

By applying Lemma 1 again, but now with respect to the hypothesis with S_2 being the outliers, we get that

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_2\left\{\delta \neq 2\right\}\\ &\leq \min_{(q_1,\ldots,q_M)\in E} \sum_{i\in S_1} D\left(q_i \| \mu_i\right) + \sum_{j\in S_2\setminus S_1} D\left(q_j \| \mu_j\right)\\ &\qquad\qquad + \sum_{k\notin S_2} D\left(q_k \| \pi\right),\end{aligned} \tag{3.59}$$

where $\mathbb{P}_2$ is the distribution under the hypothesis with S_2 being the outliers. Since ϵ is independent of $\{\mu_j\}_{j\in S_2\setminus S_1}$, we can pick $\{\mu_j\}_{j\in S_2\setminus S_1}$ to be such that $\sum_{j\in S_2\setminus S_1} D\left(\mu_j \| \pi\right) < \frac{\epsilon}{2}$. It

now follows from the definition of E, (3.59) and Lemma 1 that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_2\left\{\delta \neq 2\right\} \;=\; 0,$$

which establishes the assertion. □

Remark 3. The negative result in Theorem 7 should not be considered as overly pessimistic. It should be viewed as a theoretical result that holds *only when* each of the outliers can be arbitrarily distributed. In practice, there will likely be modeling constraints that would confine the set of all possible tuples of the distributions of all outliers. An extreme case of such constraints is when all the outliers are forced to be identically distributed, which is when universally exponential consistency is indeed attained (cf. Theorem 10) if the null hypothesis is excluded. An interesting future research direction would be to characterize the "least" stringent joint constraint on the distributions of the outliers that still allows us to construct universally exponentially consistent tests.

For the rest of this section, we restrict our attention to the case in which the number of outliers, denoted by $K \geq 1$, is known at the outset, i.e., $|S| = K$, for every $S \in \mathcal{S}$. Unlike the models in Sections 3.1 and 3.2 where the outlier sequence is always distributed according to a fixed distribution $\mu \neq \pi$ regardless of its identity $i = 1, \ldots, M$, in our model for this section, the distributions of different outlier sequences $\mu_i,\ i \in S$, *can* vary across the indices of the sequences, $i \in S$.

To contrast with the universal setting, the next result characterizes the optimal error exponent for the maximal error probability when both μ and π are known.

Proposition 8. *For every fixed number of outliers $K \geq 1$, when all the $\mu_i,\ i = 1, \ldots, M$, and π are known, the optimal error exponent is equal to*

$$\min_{1\leq i<j\leq M}\ C\left(\mu_i\left(y\right)\pi\left(y'\right), \pi\left(y\right)\mu_j\left(y'\right)\right). \tag{3.60}$$

When all outlier sequences are identically distributed, i.e., $\mu_i = \mu \neq \pi,\ i = 1, \ldots, M$, this optimal error exponent is independent of M and is equal to

$$2B\left(\mu, \pi\right). \tag{3.61}$$

Proof. The proposition follows from a well-known result in detection and estimation in the context of the multihypothesis testing problem[36]. In particular, the optimal error exponent for testing M hypotheses with i.i.d. observations with respect to $p_1, p_2, \ldots, p_M$ is characterized as $\min\limits_{1\leq i<j\leq M} C\left(p_i, p_j\right)$.

When all the $\{\mu_i\}_{i=1}^M$ and π are known, the underlying outlier hypothesis testing problem is just a multihypothesis testing problem based on i.i.d. vector observations (with M independent components) and consequently, the optimal error exponent can be computed as

$$\begin{aligned}
&\min_{S\neq S'} C\Big(\prod_{i\in S}\mu_i(y_i)\prod_{j\notin S}\pi(y_j), \prod_{i\in S'}\mu_i(y_i)\prod_{j\notin S'}\pi(y_j)\Big)\\
&= \min_{S\neq S'} C\Big(\prod_{i\in S\setminus S'}\mu_i(y_i)\prod_{j\in S'\setminus S}\pi(y_j), \prod_{i\in S\setminus S'}\pi(y_i)\prod_{j\in S'\setminus S}\mu_j(y_j)\Big)\\
&= \min_{S\neq S'}\max_{s\in[0,1]} -\log\left[\sum_{\substack{y_i,\ i\in S\setminus S'\\ y_j,\ j\in S'\setminus S}}\Big(\prod_{i\in S\setminus S'}\mu_i(y_i)^{1-s}\pi(y_i)^s\prod_{j\in S'\setminus S}\pi(y_j)^{1-s}\mu_j(y_j)^s\Big)\right] \qquad (3.62)\\
&= \min_{1\le i<j\le M}\ \max_{s\in[0,1]} -\log\left[\sum_{y_i,y_j}\Big(\mu_i(y_i)^{1-s}\pi(y_i)^s\pi(y_j)^{1-s}\mu_j(y_j)^s\Big)\right] \qquad (3.63)\\
&= \min_{1\le i<j\le M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right),
\end{aligned}$$

where the equality in (3.63) follows by virtue of fact that the outer minimum in (3.62) is attained among the pairs of S, S', with the largest number of sequences in their intersections: $K-1$.

When all the outliers are identically distributed, i.e., $\mu_i = \mu$, $i = 1, \ldots, M$, this optimal error exponent can be further simplified to be

$$\begin{aligned}
&\min_{1\le i<j\le M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right)\\
&= C\left(\mu(y)\pi(y'), \pi(y)\mu(y')\right) \;=\; 2B(\mu,\pi). \qquad (3.64)
\end{aligned}$$

□

3.3.2 Generalized Likelihood Test

We now give a summary of the GL test for the current models with a known number of outliers for both the setting when only π is known and for the completely universal setting.

Conditioned on the outlier subset being $S \in \mathcal{S}$, the likelihood of y^{Mn} is a function of the outlier indices, and the typical and outlier distributions (cf. (3.54)), i.e.,

$$p_S\left(y^{Mn}\right) \;=\; L\left(y^{Mn}, \{\mu_i\}_{i\in S}, \pi\right). \qquad (3.65)$$

When only π is known, we compute the generalized likelihood of y^{Mn} by replacing μ_i in

(3.65) with its ML estimate $\hat{\mu}_i \triangleq \gamma_i$, $i \in S$, as

$$\hat{p}_S^{\text{typ}}\left(y^{Mn}\right) \;=\; L\left(y^{Mn}, \{\hat{\mu}_i\}_{i\in S},\, \pi\right). \tag{3.66}$$

Similarly, for the completely universal setting, we compute the generalized likelihood of y^{Mn} by replacing the μ_i and π in (3.65) with their ML estimates $\hat{\mu}_i \triangleq \gamma_i$, $i \in S$, and $\hat{\pi}_S \triangleq \frac{\sum_{k\notin S}\gamma_k}{M-K}$, as

$$\hat{p}_S^{\text{univ}}\left(y^{Mn}\right) \;=\; L\left(y^{Mn}, \{\hat{\mu}_i\}_{i\in S},\, \hat{\pi}_S\right). \tag{3.67}$$

The test then selects the hypothesis under which the generalized likelihood is maximized (ties are broken arbitrarily), i.e.,

$$\delta\big(y^{Mn}\big) \;=\; \underset{S\subset\{1,\ldots,M\},\ |S|=K}{\operatorname{argmax}} \hat{p}_S^{\text{typ}} \tag{3.68}$$

for the setting when only π is known, and

$$\delta\big(y^{Mn}\big) \;=\; \underset{S\subset\{1,\ldots,M\},\ |S|=K}{\operatorname{argmax}} \hat{p}_S^{\text{univ}} \tag{3.69}$$

for the completely universal setting, respectively. It is straightforward to show using (2.4) that when only π is known, the GL test in (3.68) is equivalent to

$$\delta\big(y^{Mn}\big) \;=\; \underset{S\subset\{1,\ldots,M\},\ |S|=K}{\operatorname{argmin}} \sum_{j\notin S} D(\gamma_j\|\pi), \tag{3.70}$$

and when neither π nor $\{\mu_i\}_{i=1}^M$ is known, the test in (3.69) is equivalent to

$$\delta\big(y^{Mn}\big) \;=\; \underset{S\subset\{1,\ldots,M\},\ |S|=K}{\operatorname{argmin}} \sum_{j\notin S} D\left(\gamma_j \,\middle\|\, \tfrac{\sum_{k\notin S}\gamma_k}{M-K}\right). \tag{3.71}$$

3.3.3 Performance of Generalized Likelihood Test

Theorem 9. *For every fixed number of outliers $K \geq 1$, when only π is known but none of $\mu_i,\ i = 1,\ldots,M$ is known, the error exponent achievable by the GL test in (3.66) is equal to*

$$\min_{1\leq i\leq M}\ 2B\left(\mu_i, \pi\right). \tag{3.72}$$

When all outlier sequences are identically distributed, i.e., $\mu_i = \mu, i = 1, \ldots, M$, this achievable error exponent is equal to

$$2B(\mu, \pi), \tag{3.73}$$

which, from Proposition 8, is the optimal error exponent when μ is also known.

Proof. For each $S \subset \mathcal{S}$, denote the test statistic in (3.70) as

$$U_S^{\text{typ}} \triangleq \sum_{j \notin S} D(\gamma_j \| \pi). \tag{3.74}$$

Consider the test δ in (3.68) and (3.70). It follows from the fact that for every $S \in \mathcal{S}$,

$$\mathbb{P}_S\{\delta \neq S\} \;=\; \mathbb{P}_S\left\{\bigcup_{S' \neq S} \left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\}\right\}$$

that

$$\begin{aligned}
&\max_{S \neq S'} \mathbb{P}_S\left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\} \\
&\leq \max_{S \in \mathcal{S}} \mathbb{P}_S\{\delta \neq S\} \\
&\leq \max_{S \in \mathcal{S}} \sum_{S' \neq S} \mathbb{P}_S\left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\} \\
&\leq (|\mathcal{S}| - 1) \max_{S \neq S'} \mathbb{P}_S\left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\}.
\end{aligned} \tag{3.75}$$

Next, we get from (3.74) that for any $S \neq S' \in \mathcal{S}$,

$$\mathbb{P}_S\left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\} = \mathbb{P}_S\left\{\sum_{i \notin S} D(\gamma_i \| \pi) \geq \sum_{i \notin S'} D(\gamma_i \| \pi)\right\}.$$

Applying Lemma 1 with $J = M,\ p_i = \mu_i,\ i \in S,\ p_j = \pi,\ j \notin S$, and

$$E = \left\{(q_1, \ldots, q_M) :\ \sum_{i \notin S} D(q_i \| \pi) \geq \sum_{i \notin S'} D(q_i \| \pi)\right\}, \tag{3.76}$$

we get that the exponent for $\mathbb{P}_S\left\{U_S^{\text{typ}} \geq U_{S'}^{\text{typ}}\right\}$ is given by the value of the following

optimization problem:

$$\min_{\{q_i\}_{i\in S\backslash S'},\ \{q_j\}_{j\in S'\backslash S}} \sum_{i\in S\backslash S'} D\left(q_i\|\mu_i\right) + \sum_{j\in S'\backslash S} D\left(q_j\|\pi\right), \tag{3.77}$$

where the minimum above is over the set of $\{q_i\}_{i\in S\backslash S'}$, $\{q_j\}_{j\in S'\backslash S}$, such that

$$\sum_{j\in S'\backslash S} D(q_j\|\pi) \ \geq \ \sum_{i\in S\backslash S'} D(q_i\|\pi).$$

We now show that the optimum value in (3.77) is equal to $\sum_{i\in S\backslash S'} 2B\left(\mu_i,\pi\right)$. First, we show that the latter is a lower bound for the former. Substituting the constraint in (3.77) into the objective function, we get that the value of (3.77) is lower bounded by

$$\min_{\{q_i\}_{i\in S\backslash S'}} \sum_{i\in S\backslash S'} D\left(q_i\|\mu_i\right) + D\left(q_i\|\pi\right) = \sum_{i\in S\backslash S'} 2B\left(\mu_i,\pi\right), \tag{3.78}$$

where the equality follows from Lemma 2. Second, note that $|S\backslash S'|$ is always equal to $|S'\backslash S|$, and, hence, we can make a suitable correspondence between elements of $S\backslash S'$ to those of $S'\backslash S$. The converse implication now follows by assigning for every $i\in S\backslash S'$, and the corresponding $j\in S'\backslash S$, $q_i = q_j = \frac{\mu_i(y)^{1/2}\pi(y)^{1/2}}{\sum_{y'\in\mathcal{Y}} \mu_i(y')^{1/2}\pi(y')^{1/2}}$, and note that this assignment trivially satisfies the constraint in (3.77) and gives rise to the objective function being equal to $\sum_{i\in S\backslash S'} 2B\left(q_i,\pi\right)$.

Lastly, it follows from (3.75) that

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log\left(\max_{S\in\mathcal{S}} \mathbb{P}_S\left\{\delta\neq S\right\}\right)\\ &= \min_{S\neq S'} \sum_{i\in S\backslash S'} 2B\left(\mu_i,\pi\right) \ = \ \min_{1\leq i\leq M} \ 2B\left(\mu_i,\pi\right).\end{aligned}$$

When $\mu_i=\mu$, $i=1,\ldots,M$,

$$\min_{1\leq i\leq M} \ 2B\left(\mu_i,\pi\right) = 2B(\mu,\pi).$$

□

Remark 4. Since the tester in Proposition 8 is more capable (with the typical and outlier distributions known) than that in Theorem 9, the optimal error exponent in (3.60) must be no smaller than that in (3.72). This is verified simply by noting that for every $i,j,\ 1\leq i<$

$j \leq M$, we get from (2.2) that

$$\begin{aligned}
&C\left(\mu_i\left(y\right)\pi\left(y'\right),\pi\left(y\right)\mu_j\left(y'\right)\right)\\
&=\max_{s\in[0,1]}-\log\left[\sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}}\left(\mu_i\left(y\right)\pi\left(y'\right)\right)^s\left(\pi\left(y\right)\mu_j\left(y'\right)\right)^{1-s}\right]\\
&\geq B\left(\mu_i,\pi\right)+B\left(\mu_j,\pi\right)\\
&\geq \min\left(2B\left(\mu_i,\pi\right),2B\left(\mu_j,\pi\right)\right). \qquad (3.79)
\end{aligned}$$

As in Section 3.1, for the current models, a test δ is *universally exponentially consistent* if for every $\mu_i,\ i=1,\ldots,M,\ \mu_i\neq\pi$, it holds that $\alpha\left(\delta,\left(\{\mu_i\}_{i=1}^M,\pi\right)\right)>0$.

Theorem 10. *For every fixed number of outliers $1\leq K<\frac{M}{2}$, the GL test δ in (3.69) is universally exponentially consistent. Furthermore, for every $\{\mu_i\}_{i=1}^M,\pi,\ \mu_i\neq\pi,\ i=1,\ldots,M$, it holds that*

$$\begin{aligned}
&\alpha\left(\delta,\left(\{\mu\}_{i=1}^M,\pi\right)\right)\\
&=\min_{\substack{S,S'\subset\{1,\ldots,M\}\\|S|=|S'|=K}}\ \min_{q_1,\ldots,q_M}\Big(\sum_{i\in S}D\left(q_i\|\mu_i\right)+\sum_{j\notin S}D\left(q_j\|\pi\right)\Big), \qquad (3.80)
\end{aligned}$$

where the inner minimum above is over the set of $(q_1,\ldots,q_M)$ such that

$$\sum_{i\notin S}D\left(q_i\,\Big\|\,\tfrac{\sum_{k\notin S}q_k}{M-K}\right)\geq\sum_{i\notin S'}D\left(q_i\,\Big\|\,\tfrac{\sum_{k\notin S'}q_k}{M-K}\right). \qquad (3.81)$$

Proof. For each $S\subset\mathcal{S}$, denote the test statistic in (3.71) as

$$U_S^{\text{univ}}\triangleq\sum_{j\notin S}D\left(\gamma_j\,\|\,\tfrac{\sum_{k\notin S}\gamma_k}{M-K}\right).$$

Consider the test δ specified by (3.69) and (3.71). It now follows in the manner similar to (3.75) that

$$\begin{aligned}
&\max_{S\neq S'}\ \mathbb{P}_S\left\{U_S^{\text{univ}}\geq U_{S'}^{\text{univ}}\right\}\\
&\leq\ \max_{S\in\mathcal{S}}\mathbb{P}_S\left\{\delta\neq S\right\}\\
&\leq\ \max_{S\in\mathcal{S}}\sum_{S'\neq S}\mathbb{P}_S\left\{U_S^{\text{univ}}\geq U_{S'}^{\text{univ}}\right\}\\
&\leq\ \left(|\mathcal{S}|-1\right)\max_{S\neq S'}\mathbb{P}_S\left\{U_S^{\text{univ}}\geq U_{S'}^{\text{univ}}\right\}. \qquad (3.82)
\end{aligned}$$

The assertion (3.80) now follows from (3.82) upon noting that the application of Lemma 1 with $J = M,\ p_i = \mu_i,\ i \in S,\ p_j = \pi,\ j \notin S$, and

$$E \ = \ \Bigg\{ (q_1, \ldots, q_M) : \ \sum_{i \notin S} D\left(q_i \,\Big\|\, \tfrac{\sum_{k \notin S} q_k}{M-K} \right) \\ \geq \ \sum_{i \notin S'} D\left(q_i \,\Big\|\, \tfrac{\sum_{k \notin S'} q_k}{M-K} \right) \Bigg\}, \tag{3.83}$$

gives that the exponent for $\mathbb{P}_S \left\{ U_S^{\text{univ}} \geq U_{S'}^{\text{univ}} \right\}$ is equal to the value of the following optimization problem:

$$\min_{q_1, \ldots, q_M} \sum_{i \in S} D\left(q_i \| \mu_i \right) + \sum_{j \notin S} D\left(q_j \| \pi \right), \tag{3.84}$$

where the minimum is over the set of $\{q_1, \ldots, q_M\}$ such that

$$\sum_{i \notin S} D\left(q_i \,\Big\|\, \tfrac{\sum_{k \notin S} q_k}{M-K} \right) \ \geq \ \sum_{i \notin S'} D\left(q_i \,\Big\|\, \tfrac{\sum_{k \notin S'} q_k}{M-K} \right).$$

Lastly, the assertion of universally exponential consistency of the GL test in (3.69) and (3.71) follows from the compactness of the the feasible set of (3.84), continuity of the objective function in (3.84), and the fact that the objective function of (3.84) can only be zero at a collection $(q_i = \mu_i, i \in S,\ q_j = \pi, j \notin S)$, which is not in the constraint set. □

Note that universally exponential consistency does not imply that

$$\lim_{M \to \infty} \ \alpha\left(\delta, \left(\{\mu_i\}_{i=1}^M, \pi \right) \right) \ > \ 0. \tag{3.85}$$

Furthermore, it follows from Proposition 8 that (3.85) is not possible if $\left(\{\mu_i\}_{i=1}^M, \pi \right)$ satisfies that

$$\lim_{M \to \infty} \ \min_{1 \leq i < j \leq M} \ C\left(\mu_i(y)\, \pi(y'), \pi(y)\, \mu_j(y') \right) \ = \ 0. \tag{3.86}$$

Consequently, a natural question that arises is whether there exists a test that achieves a positive limiting error exponent as M approaches infinity whenever the optimal error exponent does not vanish with M, i.e., its achievable error exponent satisfies (3.85) whenever (3.86) *does not* hold. Such a test is said to be *asymptotically exponentially consistent.*

Theorem 11. *For every $M \geq 3$, and every fixed number of outliers $1 \leq K < \frac{M}{2}$, the error*

exponent achievable by the GL test in (3.69) is lower bounded by

$$\min_{q\,\in\,\mathcal{P}(\mathcal{Y})}\ \min_{i=1,\ldots,M}\ 2\,B(\mu_i\,,\,q)\,, \tag{3.87}$$

where the outer minimum above is over the set of q such that

$$D(q\|\pi) \leq \frac{1}{M-K}\Big(\min_{1\leq i<j\leq M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right)\ +\ KC_\pi\Big),$$

and $C_\pi \triangleq -\log\Big(\min\limits_{y\in\mathcal{Y}}\pi(y)\Big) < \infty$.

Furthermore, as $M\to\infty$, the error exponent achievable by the test in (3.69) converges as

$$\lim_{M\to\infty}\alpha\left(\delta, \left(\{\mu_i\}_{i=1}^{M}\,,\pi\right)\right) = \lim_{M\to\infty}\ \min_{i=1,\ldots,M} 2B(\mu_i,\pi), \tag{3.88}$$

which from (3.72) of Theorem 9 is also the limit of the achievable error exponent of the test in (3.68) using the knowledge of the typical distribution. The limiting error exponent on the right side of (3.88) is always positive whenever (3.86) does not hold.

When all outlier sequences are identically distributed, i.e., $\mu_i = \mu \neq \pi,\ i = 1,\ldots,M$, the test in (3.69) achieves the optimal error exponent asymptotically as the number of sequences approaches infinity, i.e.,

$$\lim_{M\to\infty}\alpha\left(\delta, (\mu,\pi)\right) = 2B\left(\mu,\pi\right). \tag{3.89}$$

Proof. First let denote the minimizing S and S' in the outer minimum of (3.80) by $S^\star$ and $S'^\star$ respectively, and the minimizing tuple $q_1,\ldots,q_M$ in the inner minimum of (3.80) by $q_1^*,\ldots,q_M^*$. Then, we get that the achievable error exponent in (3.80) is lower bounded as

$$\begin{aligned}
&\geq\ \sum_{i\in S^\star} D\left(q_i^\star\|\mu_i\right) + \sum_{j\notin S^\star} D\left(q_j^\star\|\pi\right)\\
&\quad - \sum_{j\notin S^\star} D\left(q_j^\star\,\Big\|\,\tfrac{\sum_{k\notin S^\star} q_k^\star}{M-K}\right) + \sum_{j\notin S'^\star} D\left(q_j^\star\,\Big\|\,\tfrac{\sum_{k\notin S'^\star} q_k^\star}{M-K}\right)\\
&=\ \sum_{i\in S^\star} D\left(q_i^\star\,\|\,\mu_i\right) + \sum_{j\notin S'^\star} D\left(q_j^\star\,\Big\|\,\tfrac{\sum_{k\notin S'^\star} q_k^\star}{M-K}\right)\\
&\quad + (M-T)\,D\left(\tfrac{\sum_{k\notin S^\star} q_k^\star}{M-K}\,\Big\|\,\pi\right)\\
&\geq\ D\left(q_t^\star\,\|\,\mu_t\right) + D\left(q_t^\star\,\Big\|\,\tfrac{\sum_{k\notin S'^\star} q_k^\star}{M-K}\right)\\
&\geq\ 2B\left(\mu_t\,,\,\tfrac{\sum_{k\notin S'^\star} q_k^\star}{M-K}\right),
\end{aligned} \tag{3.90}$$

where t is an arbitrarily chosen element in $S^\star \backslash S'^\star$.

On the other hand, it follows from Proposition 8 that

$$\begin{aligned}
&\min_{1\le i<j\le M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) \\
&\geq \sum_{i\in S^\star} D\left(q_i^\star \,\|\, \mu_i\right) + \sum_{j\notin S^\star} D\left(q_j^\star \,\|\, \pi\right) \\
&\geq \sum_{j\notin S^\star \cup S'^\star} D\left(q_j^\star \,\|\, \pi\right) \\
&\geq \left(M-K-|S^\star \backslash S'^\star|\right) D\left(\frac{\sum_{j\notin S^\star\cup S'^\star} q_j^\star}{(M-K-|S^\star\backslash S'^\star|)} \,\middle\|\, \pi\right). \qquad (3.91)
\end{aligned}$$

It now follows from (3.91) that

$$\begin{aligned}
&\left(M-K\right) D\left(\frac{\sum_{k\notin S'^\star} q_k^\star}{M-K} \,\middle\|\, \pi\right) \\
&\leq \left(M-K-|S^\star\backslash S'^\star|\right) D\left(\frac{\sum_{j\notin S^\star\cup S'^\star} q_j^\star}{(M-K-|S^\star\backslash S'^\star|)} \,\middle\|\, \pi\right) \\
&\quad + \left(|S^\star\backslash S'^\star|\right) D\left(\frac{\sum_{i\in S^\star\backslash S'^\star} q_i^\star}{|S^\star\backslash S'^\star|} \,\middle\|\, \pi\right) \\
&\leq \min_{1\le i<j\le M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) + |S^\star\backslash S'^\star| C_\pi \\
&\leq \min_{1\le i<j\le M} C\left(\mu_i(y)\pi(y'), \pi(y)\mu_j(y')\right) + K C_\pi. \qquad (3.92)
\end{aligned}$$

The lower bound in (3.87) now follows from (3.90) and (3.92).

The assertion (3.88) now follows from (3.87), Proposition 8 and the continuity of $B\left(\mu, q\right)$ and $D\left(q\|\pi\right)$ in the argument q. The assertion (3.89) follows as a special case of (3.88).

It is now left only to prove the asymptotically exponential consistency of the test. Having proved (3.88), this assertion now follows upon noting that for every $i, j,\ 1 \le i < j \le M$, it holds that

$$\begin{aligned}
&C\left(\mu_i\left(y\right)\pi\left(y'\right), \pi\left(y\right)\mu_j\left(y'\right)\right) \\
&\leq 2B\left(\mu_i\left(y\right)\pi\left(y'\right), \pi\left(y\right)\mu_j\left(y'\right)\right) \\
&= -2\log\left(\sum_{y,y'\in\mathcal{Y}\times\mathcal{Y}} \left(\mu_i\left(y\right)\pi\left(y'\right)\right)^{\frac{1}{2}} \left(\pi\left(y\right)\mu_j\left(y'\right)\right)^{\frac{1}{2}}\right) \\
&= 2B\left(\mu_i, \pi\right) + 2B\left(\mu_j, \pi\right),
\end{aligned}$$

where the first inequality above follows from Lemma 3. □

3.4 Multiple Identically Distributed Outliers

In this section, we look at the setting where there is uncertainty in the number of outliers, i.e., not all hypotheses in $\mathcal{S}$ have the same number of outliers. It is also assumed that for a fixed number of outliers $k = 0, 1, 2, \ldots$, $\mathcal{S}$ either contains *all* hypotheses with k outliers, or *none* of them. Considering the result in Theorem 7, it is assumed throughout this section that all outliers are identically distributed.

3.4.1 Generalized Likelihood Test

Now *with the assumption of identically distributed outliers being taken strictly*, we compute the generalized likelihood of y^{Mn} by replacing the $\mu_i, i \in S$, and π in (3.65) with their ML estimates $\hat{\mu}_S = \hat{\mu}_i \triangleq \frac{\sum_{k\in S}\gamma_k}{|S|}$, and $\hat{\pi}_S \triangleq \frac{\sum_{k\notin S}\gamma_k}{M-|S|}$, as

$$\hat{p}_S^{\text{univ}}\left(y^{Mn}\right) \;=\; L\left(y^{Mn}, \hat{\mu}_S, \hat{\pi}_S\right). \tag{3.93}$$

The test then selects the hypothesis under which the generalized likelihood in (3.93) is maximized (ties are broken arbitrarily), i.e.,

$$\delta(y^{Mn}) \;=\; \underset{S\in\mathcal{S}}{\text{argmax}}\; \hat{p}_S^{\text{univ}}\left(y^{Mn}\right). \tag{3.94}$$

It is straightforward to show using (2.4) that the GL test in (3.94) is equivalent to

$$\delta\big(y^{Mn}\big) \;=\; \underset{S\in\mathcal{S}}{\text{argmin}} \sum_{i\in S} D\big(\gamma_i \big\| \tfrac{\sum_{k\in S}\gamma_k}{K}\big) \;+ \sum_{j\notin S} D\big(\gamma_j \big\| \tfrac{\sum_{k\notin S}\gamma_k}{M-K}\big). \tag{3.95}$$

3.4.2 Performance of Proposed Test

Theorem 12. *When there are at most K, $1 \le K < M/2$, number of outliers in each hypothesis, and all the outlier sequences are identically distributed, the GL test in (3.94) is universally exponentially consistent for every hypothesis set excluding the null hypothesis. On the other hand, when the hypothesis set contains the null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known.*

Proof. We first prove that for every hypothesis set excluding the null hypothesis, the GL test in (3.94) is universally exponentially consistent.

For each $S \subset \mathcal{S}$, denote the test statistic in (3.95) as

$$\bar{U}_S^{\text{univ}} \triangleq \sum_{i \in S} D\big(\gamma_i \big\| \tfrac{\sum_{k \in S} \gamma_k}{K}\big) + \sum_{j \notin S} D\big(\gamma_j \big\| \tfrac{\sum_{k \notin S} \gamma_k}{M-K}\big).$$

Following the same argument leading to (3.75), it suffices to show that for any $S, S' \in \mathcal{S}, S' \neq S$,

$$\lim_{n \to \infty} -\frac{1}{n} \log \left(\mathbb{P}_S \left\{ \bar{U}_S^{\text{univ}} \ \geq \ \bar{U}_{S'}^{\text{univ}} \right\} \right) \ > \ 0. \tag{3.96}$$

Applying Lemma 1 with $J = M,\ p_i = \mu,\ i \in S,\ p_j = \pi,\ j \notin S$, and

$$\begin{aligned} E_{(S,S')} = \Big\{ (q_1, \ldots, q_M) : \ & \sum_{i \in S} D\left(q_i \,\middle\|\, \tfrac{\sum_{k \in S} q_k}{K} \right) \\ & + \sum_{j \notin S} D\left(q_j \middle\| \tfrac{\sum_{k \notin S} q_k}{M-K} \right) \geq \sum_{i \in S'} D\left(q_i \,\middle\|\, \tfrac{\sum_{k \in S'} q_k}{K} \right) \\ & + \sum_{j \notin S'} D\left(q_j \,\middle\|\, \tfrac{\sum_{k \notin S'} q_k}{M-K} \right) \Big\}, \end{aligned}$$

we get that the exponent for $\mathbb{P}_S \left\{ \bar{U}_S^{\text{univ}} \ \geq \ \bar{U}_{S'}^{\text{univ}} \right\}$ is given by the value of the following optimization problem:

$$\min_{\{q_1, q_2, \ldots, q_M\} \in E_{(S,S')}} \ \sum_{i \in S} D(q_i \| \mu) + \sum_{j \notin S} D(q_j \| \pi). \tag{3.97}$$

The solution to $\sum\limits_{i \in S} D\left(q_i \| \mu\right) + \sum\limits_{j \notin S} D\left(q_j \| \pi\right) = 0$ is uniquely given by $q_i = \mu$ for $i \in S$, $q_j = \pi$ for $j \notin S$. Because $|S| < M/2$, $|S'| < M/2$, there is no $S, S' \in \mathcal{S}$, such that $S = \{1, 2, \ldots, M\} \setminus S'$. Let $q_i = \mu$ for $i \in S$, $q_j = \pi$ for $j \notin S$, it then follows that

$$\begin{aligned} 0 \ &= \ \sum_{i \in S} D\left(q_i \,\middle\|\, \tfrac{\sum_{k \in S} q_k}{K} \right) + \sum_{j \notin S} D\left(q_j \,\middle\|\, \tfrac{\sum_{k \notin S} q_k}{M-K} \right) \\ &< \sum_{i \in S'} D\left(q_i \,\middle\|\, \tfrac{\sum_{k \in S'} q_k}{K} \right) + \sum_{j \notin S'} D\left(q_j \,\middle\|\, \tfrac{\sum_{k \notin S'} q_k}{M-K} \right) \end{aligned}$$

for any $S, S' \in \mathcal{S}$, $S' \neq S$. In other words, the objective function in (3.97) is strictly positive at any feasible $(q_1, q_2, \ldots, q_M)$. By the continuity of the objective function in (3.97) and the fact that $E_{(S,S')}$ is compact for any $S, S' \in \mathcal{S}$, it holds that the value of the optimization function in (3.97) is strictly positive for every pair of $S, S' \in \mathcal{S}$, $S \neq S'$. This establishes the exponential consistency of the GL test in (3.94).

Next to prove the second assertion, it suffices to prove that even when the typical distribution is known, there cannot exist a universally exponentially consistent test in differentiating the null hypothesis from any other hypothesis with a positive number of outliers. To this end, let $S \subset \{1, 2, \ldots, M\}$, $|S| \geq 1$ denote an arbitrary set of outliers. To distinguish between the null hypothesis and S, a test is done based on a decision rule $\delta : \mathcal{Y}^{Mn} \to \{0, 1\}$, where 0 corresponds to the null hypothesis and 1 the hypothesis with S being the outliers. It should be noted that δ can only be a function of π and the observations $\mathcal{Y}^{Mn}$.

We first show that in order to distinguish between the null hypothesis and S, the empirical distributions of all the sequences $\gamma_1, \ldots, \gamma_M$ and π are sufficient statistics for the error exponent. In particular, we now show that given any test, there is another test that achieves the same error exponent with its decision being made based *only on* the empirical distributions of all M sequences and π. To this end, for feasible empirical distributions (for certain n) $\gamma_1, \ldots, \gamma_M$, let us denote the set of all M sequences conforming to these empirical distributions by $T_{(\gamma_1,\ldots,\gamma_M)}$. Among these observation sequences, let us denote the set of M sequences for which δ decides for the null hypothesis by $T^{0,\pi}_{(\gamma_1,\ldots,\gamma_M)}$, which may depend on π. Now consider another test δ' which decides on one of the two hypotheses based only on $\gamma_1, \ldots, \gamma_M$ and π. Specifically, this new test is such that for all M sequences with empirical distributions $\gamma_1, \ldots, \gamma_M$, it decides for the null hypothesis if $|T^{0,\pi}_{(\gamma_1,\ldots,\gamma_M)}| \geq \frac{1}{2}|T_{(\gamma_1,\ldots,\gamma_M)}|$, and for S otherwise. It follows from this construction of δ' that for any μ and π,

$$\begin{aligned} &\max\left(\mathbb{P}_0\left\{\delta' \neq 0\right\}, \mathbb{P}_1\left\{\delta' \neq 1\right\}\right) \\ &\leq\ 2\max\left(\mathbb{P}_0\left\{\delta \neq 0\right\}, \mathbb{P}_1\left\{\delta \neq 1\right\}\right), \end{aligned}$$

where $\mathbb{P}_0, \mathbb{P}_1$ are the distributions under the null hypothesis, and under the hypothesis with S being the outliers, respectively. Consequently, the error exponent achievable by δ' is the same as that achievable by δ for any μ, π, $\mu \neq \pi$.

Having shown that the empirical distributions of the M sequences and π are sufficient statistics, it suffices to consider tests that depend only on $\gamma_1, \ldots, \gamma_M$, and π. In particular, for any such δ, let assume that for any π, there exists $\epsilon = \epsilon(\pi) > 0$ such that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_0\left\{\delta \neq 0\right\}\ >\ \epsilon. \tag{3.98}$$

Let E be the set of empirical distributions

$$E \triangleq \left\{(q_1, \ldots, q_M) : \sum_{i\in S} D\left(q_i \| \pi\right) + \sum_{j\notin S} D\left(q_j \| \pi\right) \leq \frac{\epsilon}{2}\right\}.$$

For an arbitrary element $(q_1, \ldots, q_M) \in E$, consider the set A of all M tuples $\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \in \mathcal{Y}^{Mn}$ conforming to the empirical distributions $(q_1, \ldots, q_M)$. It then follows from Lemma 1 that

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_0\Big\{\left(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}\right) \in A\}\\ &= \sum_{i\in S} D\left(q_i\|\pi\right) + \sum_{j\notin S} D\left(q_j\|\pi\right) \;\leq\; \frac{\epsilon}{2}.\end{aligned}$$

It now follows from (3.98) that

$$E \subseteq \{\delta = 0\}.$$

By applying Lemma 1 again, but now with respect to the hypothesis with S being the outliers, we get that

$$\begin{aligned}&\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1\left\{\delta \neq 1\right\}\\ &\leq \min_{(q_1,\ldots,q_M)\in E} \;\sum_{i\in S} D\left(q_i\|\mu\right) \;+\; \sum_{j\notin S} D\left(q_j\|\pi\right).\end{aligned} \tag{3.99}$$

Since ϵ is independent of μ, and μ can be chosen arbitrarily close to π, we can pick μ to be such that $\sum_{i\in S} D\left(\mu\|\pi\right) < \frac{\epsilon}{2}$. It now follows from the definition of E, (3.99) and Lemma 1 that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}_1\left\{\delta \neq 1\right\} \;=\; 0,$$

which establishes the assertion, since if (3.98) did not hold, the error exponent for δ would also have been zero. □

Remark 5. When the null hypothesis is present, we can make a suitable modification to the test in (3.94) similar to (3.41) to get a universal test that achieves a positive exponent for every conditional error probability, conditioned on any non-null hypothesis, and additional consistency under the null hypothesis.

3.5 Optimal Test When Only μ Is Known

Now we address the issue raised in Remark 1. In particular, when only μ is known, instead of using the corresponding version of the GL test in Section 3.1.1, we adopt the following

test $\tilde{\delta}$:

$$\tilde{\delta}(y^{Mn}) \;=\; \arg\min_{i=1,\ldots,M} D(\gamma_i\|\mu), \tag{3.100}$$

where γ_i denotes the empirical distribution of $\boldsymbol{y}^{(i)}, i = 1, \ldots, M$, and the ties in (3.100) are broken arbitrarily.

It now follows from (3.100) that

$$\mathbb{P}_1\{\tilde{\delta} \neq 1\} \;\leq\; (M-1)\mathbb{P}_1\big\{D(\gamma_1\|\mu) \;\geq\; D(\gamma_2\|\mu)\big\}.$$

Applying Lemma 1 with $J = 2$, $p_1 = \mu$, $p_2 = \pi$, and

$$E \;=\; \big\{\,(q_1, q_2) \;:\; D\,(q_1\|\mu) \;\geq\; D\,(q_2\|\mu)\,\big\},$$

we get that the exponent for $\mathbb{P}_1\{\tilde{\delta} \neq 1\}$ is given by the value of the following optimization problem:

$$\begin{aligned}
&\min_{\substack{q_1,q_2 \in \mathcal{P}(\mathcal{Y}) \\ D(q_1\|\mu) \geq D(q_2\|\mu)}} D\,(q_1\|\mu) + D\,(q_2\|\pi) \\
&\qquad \geq \;\min_{q_2}\; D\,(q_2\|\mu) + D\,(q_2\|\pi)\,, \\
&\qquad =\;\; 2B(\mu, \pi),
\end{aligned}$$

where the inequality follows by substituting the constraint into the objective function and the equality follows from Lemma 2.

3.6 Conclusion

In this chapter, we formulated and studied the problem of outlier hypothesis testing in the fixed sample size setting. Our main contribution was in proving that GL tests yield exponentially decaying probability of error with the number of observations under various universal settings. In particular, for the case with exactly one outlier, the GL test was shown to be universally exponentially consistent. We also provided a characterization of the error exponent achievable by the GL test for each $M \geq 3$. Surprisingly the GL test is not only universally exponentially consistent, but also asymptotically optimal as the number of sequences goes to infinity. Specifically, as M goes to infinity, the error exponent achievable by the GL test converges to the absolutely optimal error exponent when both the outlier and typical distri-

butions are known. When it is also possible that there is no outlier among the sequences, a suitable modification of the GL test was shown to achieve exponential consistency under each non-null hypothesis, and consistency under the null hypothesis universally. We then extended our models to cover the case with more than one outlier. For models with a known number of outliers, the distributions of the outliers could be distinct as long as each of them differs from the typical distribution. The GL test was shown to be universally exponentially consistent. Furthermore, we characterized the limiting error exponent achieved by such a test, and established its universally asymptotically exponential consistency. When the number of outliers is not known, it was shown that the assumption of the outliers being identically distributed and the exclusion of the null hypothesis were both essential for existence of universally exponentially consistent test. In particular, for models with an unknown number of identically distributed outliers, the GL test is universally exponentially consistent when the null hypothesis is excluded. When the null hypothesis is included, a slight modification of the GL test was shown to achieve a positive error exponent under every non-null hypothesis, and also consistency under the null hypothesis universally.

CHAPTER 4

SEQUENTIAL SETTING

In Chapter 3, we studied universal outlier hypothesis testing in a fixed sample size setting. The main finding therein was that the generalized likelihood (GL) test is far more efficient for universal outlier hypothesis testing than for the other inference problems, such as homogeneity testing and classification [7, 10, 11]. In particular, the GL test was shown to be *universally exponentially consistent* for outlier hypothesis testing, whereas it is impossible to achieve universally exponential consistency for homogeneity testing or classification without training data [10, 11]. We also showed that the GL test is *asymptotically optimal* in the limit of the large number of sequences. In this chapter, we generalize the scope of these previous findings to the sequential setting.

Sequential hypothesis testing has a rich history going back to the seminal work of Wald [37]. A majority of the results on sequential hypothesis testing have been for the case where the conditional distributions of the observations under the hypotheses are completely known (see, e.g., [26, 27, 37–40]). For the case where the distribution of the observations is not completely specified, there have been a number of results for composite sequential hypothesis testing with parametric families of distributions. There are two general approaches for constructing sequential tests for such parametric settings, one based on a weighted (or mixture) likelihood function for each hypothesis (see, e.g., [41]), and the other based on a maximum (generalized) likelihood function for each hypothesis (see, e.g., [42]). There have also been a limited number of papers on non-parametric approaches to sequential hypothesis testing where the functional form of the distribution is unknown, but it is known, for example, that the conditional distributions under the various hypotheses are rigid translations of each other (see, e.g., [43]). Sequential outlier hypothesis testing is closely related to the so called *slippage problem* studied in the sequential setting (see, e.g., [44]). In the slippage problem, N populations are identically distributed except possibly for one. The goal is to decide whether or not one of the populations has "slipped", and if so, which one. However, prior work on the slippage problem has concerned the situation when the typical and "slipped" distributions are tightly coupled, for example, when they are mean-shifted versions of each other. In universal sequential outlier hypothesis testing, we have no information regarding the outlier distribution, or we

have no information regarding both the outlier and typical distributions. In particular, the outlier and typical distributions can be arbitrarily close to each other. In addition, we have no training data to learn the unknown distributions before the test is performed. To the best of our knowledge, there has been no prior work on sequential outlier hypothesis testing in such a fully non-parametric setting that we study in this dissertation. On the other hand, we make the simplifying assumption that each instantaneous observation takes value in a finite common (known) alphabet. Under this assumption, we show that it is possible to construct an efficient universal test that will be proven to be universally consistent, and to sometimes be asymptotically optimal universally or in the limit as the number of sequences goes to infinity. The proposed universal test has the flavor of the repeated significance test [28, 29], where the test stops when the GL for the most likely hypothesis is larger than that for all the competing hypotheses by a time-dependent threshold, if that event happens before a predetermined deadline.

Sections 4.1 and 4.2 concern models with at most one outlier and up to $K > 1$ identically distributed outliers, respectively. We discuss the extension to the model with multiple distinct outliers in Section 4.3.

4.1 At Most One Outlier

In this section, we consider models where there is at most one outlier among the M sequences. We assume that the outlier distribution is independent of the identity of the outlier. In particular, the observations in an sequence are distributed (i.i.d.) according to the outlier distribution $\mu \in \mathcal{P}(\mathcal{Y})$. When the i-th sequence is the outlier, the joint distribution of the *first* n observations is given by the same expression as in (3.1). Under the hypothesis with no outlier, namely, the *null* hypothesis, all sequences are commonly distributed according to the typical distribution. The joint distribution of the first n observations is given in (3.2).

A sequential test for the outlier consists of a stopping rule and a final decision rule. The stopping rule defines a random (Markov) time, denoted by N, which is the number of observations that are taken until a final decision is made. At the stopping time $N = n$, a decision is made based on a decision rule $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$. The overall goal in the sequential testing is to achieve a certain level of accuracy for the final decision using the fewest number of observations on average.

We consider the sequential outlier hypothesis testing problem in two settings: the setting where only π is known, and the completely universal setting where neither μ nor π is known. Consequently, a universal test is not allowed to be a function of μ, and of μ or π, in the

respective settings.

Similar to the fixed sample size setting, the accuracy of a sequential test is gauged using the maximal error probability $P_{\max}$, which is a function of both the test and (μ, π) and is defined as

$$P_{\max} \triangleq \max_{i=0,1,\ldots,M} \mathbb{P}_i \left\{ \delta\left(\boldsymbol{Y}^N\right) \neq i \right\},$$

where $\mathbb{P}_i,\ i = 0, 1, \ldots, M$, are the probabilities under the null hypothesis and the non-null hypotheses when the i-th sequence, $i = 1, \ldots, M$, is the outlier. We say that a sequence of tests is *universally consistent* if the maximal error probability converges to zero for any $\mu, \pi, \mu \neq \pi$. Further, we say that it is *universally exponentially consistent* if the exponent for the maximal error probability with respect to the expected stopping time under each hypothesis is strictly positive, i.e., there exists $\alpha_i > 0$ such that for any $\mu, \pi, \mu \neq \pi$ as $P_{\max} \to 0$,

$$\mathbb{E}_i [N] \ \leq \ \frac{-\log P_{\max}}{\alpha_i} (1 + o(1)), \tag{4.1}$$

where $o(1)$ denotes a term that vanishes as $P_{\max} \to 0$.

We first consider the setting where both the typical and outlier distributions are known. In this non-universal setting, the Multihypothesis Sequential Probability Ratio Test (MSPRT) was shown to be asymptotically optimal in the limit as the error probability goes to zero [27]. For a given threshold $T > 1$ and with $\hat{i}\left(\boldsymbol{y}^n\right) \triangleq \operatorname*{argmax}_{i=0,1,\ldots,M} p_i\left(\boldsymbol{y}^n\right)$, denoting the instantaneous maximum likelihood (ML) estimate of the hypothesis at time n, the stopping time N^* and the final decision rule δ^* of the MSPRT are defined as follows:

$$N^* \ = \ \operatorname*{argmin}_{n \geq 1} \left[\frac{p_{\hat{i}}\left(\boldsymbol{Y}^n\right)}{\max\limits_{j \neq \hat{i}} p_j\left(\boldsymbol{Y}^n\right)} \ > \ T \right], \tag{4.2}$$

$$\delta^* \ = \ \hat{i}\left(\boldsymbol{Y}^{N^*}\right). \tag{4.3}$$

The following proposition (cf.[26, 27]) characterizes the asymptotic optimality of the MSPRT when the distributions of the observations are known.

Proposition 13. *As the threshold T in (4.2) approaches infinity, the MSPRT in (4.2) and (4.3) satisfies*

$$P_{\max} \ \leq \ O\left(\frac{1}{T}\right).$$

In addition, for each $i = 1, \ldots, M$, *as* $T \to \infty$,

$$\mathbb{E}_i\left[N^*\right] \;=\; \frac{\log T}{D\left(\mu\|\pi\right)}(1+o(1)) \;=\; \frac{-\log P_{\max}}{D\left(\mu\|\pi\right)}(1+o(1)),$$

and

$$\mathbb{E}_0\left[N^*\right] \;=\; \frac{\log T}{D\left(\pi\|\mu\right)}(1+o(1)) \;=\; \frac{-\log P_{\max}}{D\left(\pi\|\mu\right)}(1+o(1)).$$

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests (N, δ) *with vanishing maximal error probability, it holds for every* $i = 1, \ldots, M$, *that*

$$\mathbb{E}_i[N] \;\geq\; \frac{-\log P_{\max}}{D\left(\mu\|\pi\right)}(1+o(1)), \tag{4.4}$$

and that

$$\mathbb{E}_0[N] \;\geq\; \frac{-\log P_{\max}}{D\left(\pi\|\mu\right)}(1+o(1)). \tag{4.5}$$

Now we consider the universal settings where the outlier distribution is unknown, and where neither the outlier nor typical distribution is known. For the fixed sample size problem with at most one outlier, it was proved in Chapter 3 that a universally exponentially consistent test cannot exist. Therefore, we proposed a test therein that fulfilled a lesser objective of attaining universally exponential consistency under all the non-null hypotheses, while satisfying *only* universal consistency under the null hypothesis. We now describe a universal sequential test satisfying a similar objective tailored to the sequential setting.

4.1.1 Proposed Universal Test

Our universal test has stopping and final decision rules similar to those of the MSPRT in (4.2) and (4.3), but with the unknown likelihood functions $p_i(\boldsymbol{y}^n)$, $i = 1, \ldots, M$, being replaced with the appropriate GL functions. Specifically, when only π is known, the GL of $\boldsymbol{y}^n$ can be computed as in (3.5). When neither π nor μ is known, the GL of $\boldsymbol{y}^n$ is given by (3.6). Another key idea in the test is the adoption of a time-dependent threshold similar to that in [28, 29].

When only π is known and with $\hat{\imath} \;\triangleq\; \underset{i=1,\ldots,M}{\operatorname{argmax}}\; \hat{p}_i^{\,\mathrm{typ}}\left(\boldsymbol{y}^n\right) = \underset{i=1,\ldots,M}{\operatorname{argmax}} D(\gamma_i\|\pi)$, denoting the instantaneous estimate of the non-null hypothesis (using the GL) at time n, consider the

following (stopping) time:

$$\begin{aligned}\tilde{N} &\triangleq \operatorname*{argmin}_{n\geq 1}\left[\frac{\hat{p}_{\hat{i}}^{\mathrm{typ}}\left(\boldsymbol{y}^n\right)}{\max\limits_{j\neq\hat{i}}\hat{p}_j^{\mathrm{typ}}\left(\boldsymbol{y}^n\right)} \;>\; T(n+1)^{M|\mathcal{Y}|}\right] \\ &= \operatorname*{argmin}_{n\geq 1}\left[\min_{j\neq\hat{i}} n\big(D(\gamma_{\hat{i}}\|\pi)-D(\gamma_j\|\pi)\big) > \log T + M|\mathcal{Y}|\log(n+1)\right]. \qquad (4.6)\end{aligned}$$

Our test stops at this time or at $\lfloor T\log T\rfloor$, depending on which one is smaller, i.e.,

$$N^* \;=\; \min\left(\tilde{N}, \lfloor T\log T\rfloor\right), \qquad (4.7)$$

and correspondingly, the final decision is made according to

$$\delta^* \;=\; \begin{cases}\hat{i}\left(\boldsymbol{Y}^{N^*}\right) & \text{if } \tilde{N}\leq T\log T;\\ 0 & \text{if } \tilde{N} > T\log T.\end{cases} \qquad (4.8)$$

Similarly, when neither μ nor π is known, the test can be described by the following stopping and final decision rules:

$$N^* \;=\; \min\left(\tilde{N}, \lfloor T\log T\rfloor\right), \qquad (4.9)$$

$$\delta^* \;=\; \begin{cases}\hat{i}\left(\boldsymbol{Y}^{N^*}\right) & \text{if } \tilde{N}\leq T\log T;\\ 0 & \text{if } \tilde{N} > T\log T,\end{cases} \qquad (4.10)$$

where

$$\tilde{N} \triangleq \operatorname*{argmin}_{n\geq 1}\left[\min_{j\neq\hat{i}}\; n\Big[\sum_{k\neq j} D\left(\gamma_k\Big\|\tfrac{\sum_{\ell\neq j}\gamma_\ell}{M-1}\right) - \sum_{k\neq\hat{i}} D\left(\gamma_k\Big\|\tfrac{\sum_{\ell\neq\hat{i}}\gamma_\ell}{M-1}\right)\Big] > \log T + M|\mathcal{Y}|\log(n+1)\right], \qquad (4.11)$$

and $\hat{i} = \hat{i}\left(\boldsymbol{y}^n\right) \triangleq \operatorname*{argmin}_{i=1,\ldots,M}\sum_{k\neq i} D\left(\gamma_k\Big\|\frac{\sum_{\ell\neq i}\gamma_\ell}{M-1}\right)$. This proposed universal test has the flavor of the repeated significance test [28, 29], where the test stops when the GL for the most likely hypothesis is larger than those for all the competing hypotheses by a time-dependent threshold, if that event happens before a predetermined deadline.

4.1.2 Performance of Proposed Test

Theorem 14. *When only π is known, for every M and any $\mu \neq \pi$, the proposed test in (4.6), (4.7), (4.8) is universally consistent, and yields for every T that*

$$P_{\max} \;\leq\; O\left(\frac{1}{T}\right), \tag{4.12}$$

where the constant in the term $O\left(\frac{1}{T}\right)$ in (4.12) depends only on M, μ, π, but not on T. In addition, for each $i = 1, \ldots, M$, as $T \to \infty$,

$$\mathbb{E}_i\left[N^*\right] \;=\; \frac{\log T}{D(\mu\|\pi)}(1+o(1)) \;=\; \frac{-\log P_{\max}}{D(\mu\|\pi)}(1+o(1)). \tag{4.13}$$

First define for each $i = 1, \ldots, M$,

$$\tilde{N}_i \;\triangleq\; \underset{n\geq 1}{\operatorname{argmin}} \left[\min_{j\neq i} n\left(D\left(\gamma_i\|\pi\right) - D\left(\gamma_j\|\pi\right)\right) > \log T + M|\mathcal{Y}|\log(n+1)\right]. \tag{4.14}$$

The proof relies on the following two lemmas.

Lemma 4. *Under every non-null hypothesis $i = 1, \ldots, M$, it holds that*

$$\mathbb{P}_i\{\tilde{N}_i \geq n\} \;\leq\; MTn^{(M+2)|\mathcal{Y}|}e^{-(n-1)2B(\mu,\pi)}.$$

Proof. We get by the definition of $\tilde{N}_i$ in (4.14) that

$$\begin{aligned}
&\mathbb{P}_i\{\tilde{N}_i \geq n\} \\
&\leq \sum_{j\neq i} \mathbb{P}_i\Big\{(n-1)\Big[D(\gamma_i^{(n-1)}\|\pi) - D(\gamma_j^{(n-1)}\|\pi)\Big] \;\leq\; \log T + M|\mathcal{Y}|\log n\Big\} \\
&\leq \sum_{j\neq i} \mathbb{P}_i\left\{D(\gamma_i\|\mu) + D\left(\gamma_j\|\pi\right) \;\geq\; -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}|\log n\right) + \left(D(\gamma_i\|\mu) + D(\gamma_i\|\pi)\right)\right\} \\
&\leq \sum_{j\neq i} \mathbb{P}_i\left\{D\left(\gamma_i\|\mu\right) + D\left(\gamma_j\|\pi\right) \;\geq\; -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}|\log n\right) + 2B(\mu,\pi)\right\},
\end{aligned} \tag{4.15}$$

where the last inequality follows from Lemma 2. Continuing from (4.15) by using (2.6) upon noting the independence of the i-th and j-th sequences and that the numbers of feasible empirical distributions γ_i, γ_j, each from sequences of length $(n-1)$, are both upper bounded

by $n^{|\mathcal{Y}|}$ (cf. Chapter 2), we get that

$$\begin{aligned}\mathbb{P}_i\{\tilde{N}_i \geq n\} &\leq M\left(Tn^{M|\mathcal{Y}|}\right) n^{2|\mathcal{Y}|} e^{-(n-1)2B(\mu,\pi)} \\ &\leq MTn^{(M+2)|\mathcal{Y}|} e^{-(n-1)2B(\mu,\pi)}.\end{aligned} \tag{4.16}$$

□

Lemma 5. *Under each non-null hypothesis $i = 1, \ldots, M$,*

$$\lim_{T\to\infty} \mathbb{E}_i\left[\left|\frac{\tilde{N}_i}{\log T} - \frac{1}{D\left(\mu\|\pi\right)}\right|\right] = 0. \tag{4.17}$$

Proof. First observe that under hypothesis $i = 1, \ldots, M$, we obtain by the strong law of large numbers that for every $y \in \mathcal{Y}$, $\frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\left\{Y_k^{(i)} = y\right\}$ converges to $\mu(y)$ a.s. Consequently, it follows that $\gamma_i \to \mu$ a.s. Similarly under hypothesis i and for every $j \neq i$, $\gamma_j \to \pi$ a.s.

For each $i = 1, \ldots, M$, and any fixed T, it holds by Lemma 4 that $\tilde{N}_i$ is finite a.s. under $\mathbb{P}_i$, i.e.,

$$\mathbb{P}_i\{\tilde{N}_i \geq n\} \to 0 \quad \text{as} \quad n \to \infty. \tag{4.18}$$

It then follows from this a.s. finiteness and the definition of $\tilde{N}_i$ in (4.14) that with probability 1 under $\mathbb{P}_i$,

$$\min_{j\neq i}\left(D\big(\gamma_i^{(\tilde{N}_i)}\|\pi\big) - D\big(\gamma_j^{(\tilde{N}_i)}\|\pi\big)\right) > \frac{\log T + M|\mathcal{Y}|\log(\tilde{N}_i+1)}{\tilde{N}_i}; \tag{4.19}$$

$$\min_{j\neq i}\left(D\big(\gamma_i^{(\tilde{N}_i-1)}\|\pi\big) - D\big(\gamma_j^{(\tilde{N}_i-1)}\|\pi\big)\right) \leq \frac{\log T + M|\mathcal{Y}|\log \tilde{N}_i}{\tilde{N}_i - 1}. \tag{4.20}$$

Next, by observing that for any distribution q, $D(q\|\pi) \leq \log\left(\frac{1}{\min_y \pi(y)}\right) < \infty$ (π has a full support), we get from (4.19) that

$$\begin{aligned}\mathbb{P}_i\{\tilde{N}_i \leq n\} &\leq \mathbb{P}_i\left\{\tilde{N}_i D\left(\gamma_i^{(\tilde{N}_i)}\Big\|\pi\right) > \log T;\ \tilde{N}_i \leq n\right\} \\ &\leq \mathbb{P}_i\left\{n\log\left(\frac{1}{\min\limits_y \pi(y)}\right) > \log T\right\} \\ &= 0, \quad \text{for every } n < \frac{\log T}{\log\left(\frac{1}{\min\limits_y \pi(y)}\right)},\end{aligned}$$

thereby yielding that $\tilde{N}_i \to \infty$, as $T \to \infty$ a.s. under $\mathbb{P}_i$. Consequently, we conclude from the continuity of $D(\cdot\|\pi)$ and the a.s. convergences of $\gamma_i^{(n)}$ to μ and $\gamma_j^{(n)}$, $j \neq i$, to π that under $\mathbb{P}_i$,

$$\min_{j\neq i}\Big(D\big(\gamma_i^{(\tilde{N}_i)}\|\pi\big) - D\big(\gamma_j^{(\tilde{N}_i)}\|\pi\big)\Big), \min_{j\neq i}\Big(D\big(\gamma_i^{(\tilde{N}_i-1)}\|\pi\big) - D\big(\gamma_j^{(\tilde{N}_i-1)}\|\pi\big)\Big) \;\to\; D(\mu\|\pi),$$

a.s., as $T \to \infty$. It now follows from this, (4.19) and (4.20) that under $\mathbb{P}_i$, $\frac{\tilde{N}_i}{\log T}$ converges a.s. and, hence, in probability to $\frac{1}{D(\mu\|\pi)}$.

To go from the convergence in probability to convergence in mean, it now suffices to prove that the sequence of random variables $\frac{\tilde{N}_i}{\log T}$ is uniformly integrable as $T \to \infty$. To this end, for any $\nu > 0$ sufficiently large, we upper bound the following quantity using Lemma 4 as follows:

$$\begin{aligned}
&\mathbb{E}_i\left[\frac{\tilde{N}_i}{\log T}\,\mathbb{I}_{\left\{\frac{\tilde{N}_i}{\log T}\geq\nu\right\}}\right]\\
&\leq\; \mathbb{E}_i\left[\;\frac{\Big(\tilde{N}_i - \lfloor\nu\log T\rfloor + \nu\log T\Big)}{\log T}\,\mathbb{I}_{\left\{\tilde{N}_i\geq\lfloor\nu\log T\rfloor\right\}}\;\right]\\
&\leq\; \frac{1}{\log T}\mathbb{E}_i\Big[\,\Big(\tilde{N}_i - \lfloor\nu\log T\rfloor\Big)\;\mathbb{I}_{\left\{\tilde{N}_i-\lfloor\nu\log T\rfloor\;\geq\;0\right\}}\Big] \;+\; \frac{\nu\log T}{\log T}\mathbb{P}_i\left\{\tilde{N}_i\geq\lfloor\nu\log T\rfloor\right\}\\
&=\; \frac{1}{\log T}\sum_{\ell=1}^{\infty}\mathbb{P}_i\left\{\tilde{N}_i\geq\lfloor\nu\log T\rfloor+\ell\right\} \;+\; \nu\mathbb{P}_i\left\{\tilde{N}_i\geq\lfloor\nu\log T\rfloor\right\}\\
&\leq\; \frac{MT}{\log T}\sum_{\ell=1}^{\infty}e^{-(\nu\log T+\ell-2)2B(\mu,\pi)}\,(\lfloor\nu\log T\rfloor+\ell)^{(M+2)|\mathcal{Y}|}\\
&\qquad\qquad +\;\nu MTe^{-(\nu\log T-2)2B(\mu,\pi)}\,(\lfloor\nu\log T\rfloor)^{(M+2)|\mathcal{Y}|}\,. \qquad (4.21)
\end{aligned}$$

Continuing from (4.21), it then follows that for all T sufficiently large so that $\lfloor\nu\log T\rfloor \geq 1$,

$$\begin{aligned}
&\mathbb{E}_i\left[\frac{\tilde{N}_i}{\log T}\,\mathbb{I}_{\left\{\frac{\tilde{N}_i}{\log T}\geq\nu\right\}}\right]\\
&\leq \frac{MT}{\log T}\sum_{\ell=1}^{\infty}e^{-(\nu\log T+\ell-2)2B(\mu,\pi)}\,(2\lfloor\nu\log T\rfloor\ell)^{(M+2)|\mathcal{Y}|}\\
&\quad +\;\nu MTe^{-(\nu\log T-2)2B(\mu,\pi)}\,(\lfloor\nu\log T\rfloor)^{(M+2)|\mathcal{Y}|}\,.\\
&=\; \frac{MT}{\log T}\,(2\lfloor\nu\log T\rfloor)^{(M+2)|\mathcal{Y}|}\,e^{-2\nu B(\mu,\pi)\log T}\times\left(e^{4B(\mu,\pi)}\sum_{\ell=1}^{\infty}e^{-2B(\mu,\pi)\ell}\;\ell^{(M+2)|\mathcal{Y}|}\right)\\
&\qquad +\;\nu MT\,(\lfloor\nu\log T\rfloor)^{(M+2)|\mathcal{Y}|}\,e^{-2\nu B(\mu,\pi)\log T}\times e^{4B(\mu,\pi)}, \qquad (4.22)
\end{aligned}$$

which vanishes as $T \to \infty$, for any $\nu > \frac{1}{2B(\mu,\pi)}$, thereby establishing the uniform integrability and, hence, (4.17). □

Proof. We start by proving (4.12). It follows from the description of the test in (4.6), (4.7) and (4.8) that for any $i, j = 1, \ldots, M,\ i \neq j$,

$$\begin{aligned}
\mathbb{P}_i\{\delta^* = j\} &\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{N^* = \tilde{N} = n,\ \delta^* = j\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{n\left(D\left(\gamma_j\|\pi\right) - D\left(\gamma_i\|\pi\right)\right) > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{nD\left(\gamma_j\|\pi\right) > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \frac{1}{T}\sum_{n=1}^{\infty} (n+1)^{-(M-1)|\mathcal{Y}|} \qquad (4.23) \\
&\leq \frac{C'(|\mathcal{Y}|, M)}{T}, \qquad (4.24)
\end{aligned}$$

where (4.23) follows from (2.6) and the polynomial upper bound on the number of empirical distributions.

In addition, for each $i = 1, \ldots, M$,

$$\begin{aligned}
\mathbb{P}_i\{\delta^* = 0\} &= \mathbb{P}_i\left\{\tilde{N} > T\log T\right\} \\
&\leq \frac{\mathbb{E}_i[\tilde{N}]}{T\log T} \qquad (4.25) \\
&\leq \frac{\mathbb{E}_i[\tilde{N}_i]}{T\log T} \qquad (4.26) \\
&\leq \frac{C'\left(\mu, \pi, |\mathcal{Y}|, M\right)}{T}, \qquad (4.27)
\end{aligned}$$

where (4.25), (4.26), and (4.27) are from the Markov inequality, the fact that for each $i = 1, \ldots, M$, $\tilde{N} \leq \tilde{N}_i$ with probability 1, and Lemma 5, respectively.

Next, it follows from the definition of $\tilde{N}$ in (4.6), and that of N^* in (4.7) that

$$\begin{aligned}
\mathbb{P}_0\left\{\delta^* \neq 0\right\} &= \mathbb{P}_0\left\{N^* = \tilde{N}\right\} \\
&= \mathbb{P}_0\left\{\tilde{N} \leq T\log T\right\} \\
&\leq \mathbb{P}_0\left\{\tilde{N} \text{ is finite}\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}_0\left\{\tilde{N} = n\right\} \\
&\leq \sum_{n=1}^{\infty}\sum_{i=1}^{M} \mathbb{P}_0\left\{nD\left(\gamma_i\|\pi\right) > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \frac{M}{T}\sum_{n=1}^{\infty} (n+1)^{-(M-1)|\mathcal{Y}|} \\
&\leq \frac{C'\left(|\mathcal{Y}|, M\right)}{T}. \qquad (4.28)
\end{aligned}$$

The combination of (4.24), (4.27), and (4.28) constitutes (4.12).

The first equality in (4.13) now follows from that for each $i = 1, \ldots, M$, the limit in probability of $\frac{N^*}{\log T}$ (under $\mathbb{P}_i$) is the same as that of $\frac{\tilde{N}_i}{\log T}$, which is $\frac{1}{D(\mu\|\pi)}$ (cf. (4.17)) by virtue of the fact that (cf. (4.24) and (4.27)) for every $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}_i\left\{\left|\frac{N^*}{\log T} - \frac{1}{D\left(\mu\|\pi\right)}\right| > \epsilon\right\} &= \mathbb{P}_i\left\{\left|\frac{N^*}{\log T} - \frac{1}{D\left(\mu\|\pi\right)}\right| > \epsilon,\ \delta = i\right\} + \mathbb{P}_i\left\{\delta \neq i\right\} \\
&= \mathbb{P}_i\left\{\left|\frac{\tilde{N}_i}{\log T} - \frac{1}{D\left(\mu\|\pi\right)}\right| > \epsilon,\ \delta = i\right\} + \mathbb{P}_i\left\{\delta \neq i\right\},
\end{aligned}$$

and the uniform integrability of $\frac{N^*}{\log T}$, which, in turn, follows from $N^* \leq \tilde{N}_i$ with probability 1, and the uniform integrability of $\frac{\tilde{N}_i}{\log T}$, as in the proof of Lemma 5. □

Remark 1. *While attaining universal consistency under the null hypothesis, the proposed test in (4.6), (4.7) and (4.8) not only achieves universally exponential consistency under all non-null hypotheses, but also yields the optimal asymptote for the expected stopping time under each of those hypotheses universally (cf. (4.4)).*

Theorem 15. *When neither μ nor π is known, for every M and any $\mu \neq \pi$, the proposed test in (4.9), (4.10) and (4.11), is universally consistent, and yields for every T that*

$$P_{\max} \leq O\left(\frac{1}{T}\right). \qquad (4.29)$$

In addition, for each $i = 1, \ldots, M$, as $T \to \infty$,

$$\begin{aligned}\mathbb{E}_i[N^*] &= \frac{\log T}{D\left(\mu \| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right) + (M-2)D\left(\pi \| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right)}(1+o(1)) \\ &\leq \frac{-\log P_{\max}}{D\left(\mu \| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right) + (M-2)D\left(\pi \| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right)}(1+o(1)). \qquad (4.30)\end{aligned}$$

First define for each $i = 1, \ldots, M$,

$$\tilde{N}_i \triangleq \operatorname*{argmin}_{n \geq 1} \left[\min_{j \neq i} \; n \left[\sum_{k \neq j} D\left(\gamma_k \| \tfrac{\sum_{\ell \neq j} \gamma_\ell}{M-1}\right) - \sum_{k \neq i} D\left(\gamma_k \| \tfrac{\sum_{\ell \neq i} \gamma_\ell}{M-1}\right) \right] > \log T + M|\mathcal{Y}| \log(n+1) \right]. \qquad (4.31)$$

The proof relies on the following two lemmas.

Lemma 6. *Under every non-null hypothesis $i = 1, \ldots, M$, and every $n \geq 1$, it holds that*

$$\mathbb{P}_i\{\tilde{N}_i \geq n\} \leq \left(M^2 - 1\right) T n^{2M|\mathcal{Y}|} e^{-(n-1)b}, \qquad (4.32)$$

where b is a function of μ, π that is always positive.

Proof. It follows from the definition of $\tilde{N}_i$ in (4.31) that

$$\begin{aligned}&\mathbb{P}_i\left\{\tilde{N}_i \geq n\right\} \\ &\leq \sum_{j \neq i} \mathbb{P}_i \Bigg\{ (n-1) \left[\sum_{k \neq j} D\left(\gamma_k^{(n-1)} \| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(n-1)}}{M-1}\right) - \sum_{k \neq i} D\left(\gamma_k^{(n-1)} \| \tfrac{\sum_{\ell \neq i} \gamma_\ell^{(n-1)}}{M-1}\right) \right] \\ &\qquad\qquad \leq \log T + M|\mathcal{Y}| \log n \Bigg\} \\ &= \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D\left(\gamma_k \| \tfrac{\sum_{\ell \neq i} \gamma_\ell}{M-1}\right) \geq -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}| \log n\right) + \sum_{k \neq j} D\left(\gamma_k \| \tfrac{\sum_{\ell \neq j} \gamma_\ell}{M-1}\right) \right\} \\ &\leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D\left(\gamma_k \| \pi\right) \geq -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}| \log n\right) + \sum_{k \neq j} D\left(\gamma_k \| \tfrac{\sum_{\ell \neq j} \gamma_\ell}{M-1}\right) \right\}, \qquad (4.33)\end{aligned}$$

where (4.33) follows from the sum centroid inequality (2.7) with $\mathcal{C} = \{\gamma_k | k = 1, \ldots, M, \; k \neq i\}$, and $q = \pi$.

Continuing from (4.33), we have

$$
\begin{aligned}
&\mathbb{P}_i\left\{\tilde{N}_i \geq n\right\} \\
&\leq \sum_{j\neq i}\mathbb{P}_i\left\{\begin{array}{l}\sum\limits_{k\neq i} D\left(\gamma_k\middle\|\pi\right) \geq -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}|\log n\right) + \sum\limits_{k\neq j} D\left(\gamma_k\middle\|\frac{\sum_{\ell\neq j}\gamma_l}{M-1}\right) \\ D\left(\gamma_i\|\mu\right)\leq\epsilon, \text{ and } D\left(\gamma_j\|\pi\right)\leq\epsilon, \text{ for all } j\neq i\end{array}\right\} \\
&\quad + \sum_{j\neq i}\mathbb{P}_i\left\{D\left(\gamma_i\|\mu\right)>\epsilon, \text{ or } D\left(\gamma_j\|\pi\right)>\epsilon, \text{ for some } j\neq i\right\} \\
&\leq \sum_{j\neq i}\mathbb{P}_i\left\{\begin{array}{l}\sum\limits_{k\neq i} D\left(\gamma_k\middle\|\pi\right) \geq -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}|\log n\right) + \sum\limits_{k\neq j} D\left(\gamma_k\middle\|\frac{\sum_{\ell\neq j}\gamma_l}{M-1}\right) \\ D\left(\gamma_i\|\mu\right)\leq\epsilon, \text{ and } D\left(\gamma_j\|\pi\right)\leq\epsilon, \text{ for all } j\neq i\end{array}\right\} \qquad (4.34) \\
&\quad + \left(M-1\right)Mn^{|\mathcal{Y}|}e^{-(n-1)\epsilon},
\end{aligned}
$$

where (4.34) is by (2.6). Note that $\sum\limits_{k\neq j} D\left(\gamma_k\middle\|\frac{\sum_{\ell\neq j}\gamma_\ell}{M-1}\right)$ in (4.34) is zero only if for all $k\neq j$, $\gamma_k=\gamma$ for some γ. This cannot happen if the ϵ in (4.34) is chosen to be sufficiently small, because it also holds for the event in (4.34) that $\mu\neq\pi$, $D\left(\gamma_i\|\mu\right)\leq\epsilon$, and for any $j\neq i$ that $D\left(\gamma_j\|\pi\right)\leq\epsilon$. We then conclude that when the ϵ is chosen to be sufficiently small (as a function of μ, π), it follows that $\sum\limits_{k\neq j} D\left(\gamma_k\middle\|\frac{\sum_{\ell\neq j}\gamma_\ell}{M-1}\right)\geq a(\mu,\pi)>0$. Continuing from (4.34) with the ϵ chosen sufficiently small and upon noting that the number of feasible empirical distributions $\gamma_k^{(n-1)}$ for each $k\neq i$, is upper bounded by $n^{|\mathcal{Y}|}$, we obtain

$$
\begin{aligned}
&\mathbb{P}_i\left\{\tilde{N}_i \geq n\right\} \\
&\leq \sum_{j\neq i}\mathbb{P}_i\left\{\begin{array}{l}\sum\limits_{k\neq i} D\left(\gamma_k\middle\|\pi\right) \geq -\frac{1}{n-1}\left(\log T + M|\mathcal{Y}|\log n\right) + a\left(\mu,\pi\right) \\ \qquad D\left(\gamma_i\|\mu\right)\leq\epsilon, \text{ and } D\left(\gamma_j\|\pi\right)\leq\epsilon, \text{ for all } j\neq i\end{array}\right\} \\
&\quad + \left(M-1\right)Mn^{|\mathcal{Y}|}e^{-(n-1)\epsilon} \\
&\leq \left(M-1\right)\left(Tn^{M|\mathcal{Y}|}e^{-(n-1)a(\mu,\pi)}\right)n^{(M-1)|\mathcal{Y}|} + \left(M-1\right)Mn^{|\mathcal{Y}|}e^{-(n-1)\epsilon} \\
&\leq \left(M^2-1\right)Tn^{2M|\mathcal{Y}|}e^{-(n-1)\min(a(\mu,\pi),\epsilon)} \\
&= \left(M^2-1\right)Tn^{2M|\mathcal{Y}|}e^{-(n-1)b}. \qquad (4.35)
\end{aligned}
$$

□

Lemma 7. *Under each non-null hypothesis $i=1,\ldots,M$,*

$$
\lim_{T\to\infty}\mathbb{E}_i\left[\left|\frac{\tilde{N}_i}{\log T} - \frac{1}{D\left(\mu\middle\|\frac{1}{M-1}\mu+\frac{M-2}{M-1}\pi\right)+(M-2)D\left(\pi\middle\|\frac{1}{M-1}\mu+\frac{M-2}{M-1}\pi\right)}\right|\right] = 0. \qquad (4.36)
$$

Proof. Under hypothesis $i = 1, \ldots, M$, the strong law of large numbers yields that as $n \to \infty$, $\gamma_i \to \mu$ a.s., and $\gamma_j \to \pi$ a.s. for every $j \neq i$. Hence, we get from the continuity of the relative entropy in both its arguments (jointly) [30] that under $\mathbb{P}_i$,

$$\min_{j \neq i} \left[\sum_{k \neq j} D\left(\gamma_k^{(n)} \Big\| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(n)}}{M-1} \right) - \sum_{k \neq i} D\left(\gamma_k^{(n)} \Big\| \tfrac{\sum_{\ell \neq i} \gamma_\ell^{(n)}}{M-1} \right) \right] \stackrel{\text{a.s.}}{\to} D\big(\mu \big\| \tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\big) + (M-2) D\big(\pi \big\| \tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\big), \tag{4.37}$$

as $n \to \infty$.

By Lemma 6, we see that $\tilde{N}_i$ is finite a.s. under $\mathbb{P}_i$, $i = 1, \ldots, M$. It then follows from this a.s. finiteness and the definition of $\tilde{N}_i$ in (4.31) that with probability 1 under $\mathbb{P}_i$,

$$\min_{j \neq i} \left[\sum_{k \neq j} D\left(\gamma_k^{(\tilde{N}_i)} \Big\| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(\tilde{N}_i)}}{M-1} \right) - \sum_{k \neq i} D\left(\gamma_k^{(\tilde{N}_i)} \Big\| \tfrac{\sum_{\ell \neq i} \gamma_\ell^{(\tilde{N}_i)}}{M-1} \right) \right] > \frac{\log T + M|\mathcal{Y}| \log(\tilde{N}_i + 1)}{\tilde{N}_i}; \tag{4.38}$$

$$\min_{j \neq i} \left[\sum_{k \neq j} D\left(\gamma_k^{(\tilde{N}_i - 1)} \Big\| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(\tilde{N}_i - 1)}}{M-1} \right) - \sum_{k \neq i} D\left(\gamma_k^{(\tilde{N}_i - 1)} \Big\| \tfrac{\sum_{\ell \neq i} \gamma_\ell^{(\tilde{N}_i - 1)}}{M-1} \right) \right] \leq \frac{\log T + M|\mathcal{Y}| \log \tilde{N}_i}{\tilde{N}_i - 1}. \tag{4.39}$$

The a.s. convergence of $\frac{\tilde{N}_i}{\log T}$ to $\frac{1}{D\left(\mu \big\| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right) + (M-2) D\left(\pi \big\| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right)}$ under hypothesis i will follow from (4.37), (4.38) and (4.39) by an argument based on sample-paths similar to that in the proof of Lemma 5 if we establish that under hypothesis i, $\tilde{N}_i \to \infty$, a.s., as $T \to \infty$. To this end, we note that for any $j \neq i$, $\sum\limits_{k \neq j} D\left(\gamma_k^{(\tilde{N}_i)} \Big\| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(\tilde{N}_i)}}{M-1} \right) \leq M \log(M-1)$, and, hence, we get from (4.38) that for any $n \geq 1$ and any $j \neq i$,

$$\begin{aligned} \mathbb{P}_i\left\{ \tilde{N}_i \leq n \right\} &\leq \mathbb{P}_i\left\{ \tilde{N}^i \left[\sum_{k \neq j} D\left(\gamma_k^{(\tilde{N}_i)} \Big\| \tfrac{\sum_{\ell \neq j} \gamma_\ell^{(\tilde{N}_i)}}{M-1} \right) \right] > \log T;\ \tilde{N}_i \leq n \right\} \\ &\leq \mathbb{P}_i\left\{ nM \log(M-1) > \log T \right\} \\ &= 0, \quad \text{for every } n < \frac{\log T}{M \log(M-1)}, \end{aligned}$$

thereby yielding that $\tilde{N}_i \to \infty$ a.s. as $T \to \infty$, and, hence, the aforementioned a.s. convergence of $\frac{\tilde{N}_i}{\log T}$ to $\frac{1}{D\left(\mu \big\| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right) + (M-2) D\left(\pi \big\| \frac{1}{M-1}\mu + \frac{M-2}{M-1}\pi\right)}$ under hypothesis i.

The main claim (4.36) follows by using the exponential tail bound for $\tilde{N}_i$ in Lemma 6

to establish the uniform integrability of the sequence $\frac{\tilde{N}_i}{\log T}$ as in the argument leading to (4.22). □

Proof. We start by proving (4.29). First, note that for any $i, j = 1, \ldots, M,\ i \neq j$,

$$
\begin{aligned}
&\mathbb{P}_i\left\{\delta^* = j\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{N^* = \tilde{N} = n,\ \delta^* = j\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{n\left[\sum_{k\neq i} D\left(\gamma_k \middle\| \tfrac{\sum_{\ell\neq i}\gamma_\ell}{M-1}\right) - \sum_{k\neq j} D\left(\gamma_k \middle\| \tfrac{\sum_{\ell\neq j}\gamma_\ell}{M-1}\right)\right] > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{n\sum_{k\neq i} D\left(\gamma_k \middle\| \tfrac{\sum_{\ell\neq i}\gamma_\ell}{M-1}\right) > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_i\left\{n\sum_{k\neq i} D\left(\gamma_k \middle\| \pi\right) > \log T + M|\mathcal{Y}|\log(n+1)\right\} \qquad (4.40) \\
&\leq \frac{1}{T}\sum_{n=1}^{\infty} (n+1)^{-|\mathcal{Y}|} \qquad (4.41) \\
&= \frac{C'(|\mathcal{Y}|)}{T}, \qquad (4.42)
\end{aligned}
$$

where (4.40) follows from (2.7) and (4.41) follows from (2.6) and the polynomial upper bound on the number of empirical distributions.

In addition, for each $i = 1, \ldots, M$, it follows from (4.36), the Markov inequality and the fact that $\tilde{N} \leq \tilde{N}_i$ with probability 1, that

$$
\mathbb{P}_i\left\{\delta^* = 0\right\} = \mathbb{P}_i\left\{\tilde{N} > T\log T\right\} \leq \frac{C'\left(\mu, \pi, |\mathcal{Y}|, M\right)}{T}. \qquad (4.43)
$$

Next, it follows from the definitions of $\tilde{N}, N^*$ in (4.11), (4.9) and (2.7) that

$$\begin{aligned}
\mathbb{P}_0\left\{\delta^* \neq 0\right\} &= \mathbb{P}_0\left\{N^* = \tilde{N}\right\} \\
&= \mathbb{P}_0\left\{\tilde{N} \leq T\log T\right\} \\
&\leq \mathbb{P}_0\left\{\tilde{N} \text{ is finite}\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}_0\left\{\tilde{N} = n\right\} \\
&\leq \sum_{n=1}^{\infty}\sum_{i=1}^{M} \mathbb{P}_0\left\{n\Big[\sum_{k\neq i} D\left(\gamma_k\Big\|\tfrac{\sum_{\ell\neq i}\gamma_\ell}{M-1}\right)\Big] > \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \sum_{n=1}^{\infty}\sum_{i=1}^{M} \mathbb{P}_0\left\{n\sum_{k\neq i} D\left(\gamma_k\big\|\pi\right) \;>\; \log T + M|\mathcal{Y}|\log(n+1)\right\} \\
&\leq \frac{M}{T}\sum_{n=1}^{\infty}(n+1)^{-|\mathcal{Y}|} \\
&\leq \frac{C'\left(|\mathcal{Y}|, M\right)}{T}. \qquad (4.44)
\end{aligned}$$

The combination of (4.42), (4.43), and (4.44) constitutes (4.29).

The proof of (4.30) follows similar steps as in the proof of (4.13): first, under $\mathbb{P}_i$, $i = 1,\ldots,M$, the limit in probability of $\frac{N^*}{\log T}$ is identical to that of $\frac{\tilde{N}_i}{\log T}$ (cf. (4.42), (4.43)); and second, the uniform integrability of $\frac{N^*}{\log T}$ follows from the uniform integrability of $\frac{\tilde{N}_i}{\log T}$, which was already established, by virtue of the fact that for each $i = 1,\ldots,M$, $N^* \leq \tilde{N}_i$ with probability 1. □

Remark 2. *Applying (2.8) with $L = M-2, p = \mu$, and $\bar{p} = \pi$, we get that*

$$D\left(\mu\big\|\tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\right) + (M-2)D\left(\pi\big\|\tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\right) \;\leq\; D(\mu\|\pi). \qquad (4.45)$$

This is consistent with (4.30) and (4.4). It also follows from (4.45) that

$$\lim_{M\to\infty} D\left(\mu\big\|\tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\right) + (M-2)D\left(\pi\big\|\tfrac{1}{M-1}\mu + \tfrac{M-2}{M-1}\pi\right) \;=\; D(\mu\|\pi),$$

i.e., as $M\to\infty$, the asymptotic performance of the test in (4.9), (4.10) and (4.11) under each non-null hypothesis (cf. (4.30)) when neither μ nor π is known, approaches the (optimal) asymptotic performance of the test in (4.6), (4.7) and (4.8) under each of those hypotheses (cf. (4.13) and (4.4)) when only π is known.

Remark 3. *The particular functional forms of the time-dependent thresholds in (4.6) and*

(4.11), and of the deterministic time horizons in (4.7) and (4.9) are chosen solely for the simplicity of exposition. In fact, it follows from our proofs that the results in Theorems 14 and 15 continue to hold when the stopping time takes a more general form as follows. Consider

$$\tilde{N} \triangleq \underset{n\geq 1}{argmin} \left[\frac{\hat{p}_{\hat{i}}\left(\boldsymbol{y}^n\right)}{\max\limits_{j\neq\hat{i}} \hat{p}_j\left(\boldsymbol{y}^n\right)} > C\left(T(n+1)^{M|\mathcal{Y}|}\right) \right], \tag{4.46}$$

where for each $i = 1, \ldots, M$, $\hat{p}_i = \hat{p}_i^{typ}$ *for the setting where* π *is known, and* $\hat{p}_i = \hat{p}_i^{univ}$ *for the completely universal setting, and* $\log C$ *is a constant offset to the time-dependent threshold that does not depend on* T*. The test stops at this time or* $\lfloor f(T) \rfloor$*, depending on which one is smaller, i.e.,*

$$N^* = \min\left(\tilde{N}, \lfloor f(T) \rfloor\right),$$

and correspondingly, the final decision is made according to

$$\delta^* = \begin{cases} \hat{i}\left(\boldsymbol{Y}^{N^*}\right) & \text{if } \tilde{N} \leq f(T); \\ 0 & \text{if } \tilde{N} > f(T), \end{cases}$$

where $f(T)$ *is any function increasing at least as fast as* $T \log T$*.*

4.2 Multiple Identically Distributed Outliers

We now generalize our results in Section 4.1 to models with multiple outliers. It was shown in Chapter 3 that for the fixed sample size setting, the assumption of the outliers being identically distributed is essential for the existence of a test that is universally exponentially consistent (under all the non-null hypotheses) when the number of outliers is not completely specified (anything from 1 to K). Therefore, we start by considering the setting where there are *up to* K outliers among the M sequences with $K < \frac{M}{2}$, and that all the outliers are identically distributed according to μ. In Section 4.3, we shall look at the extension with possibly distinctly distributed outliers but with their total number being known.

The test for the outliers is done based on a *universal* rule $\delta\left(\boldsymbol{Y}^N\right) \in \mathcal{S}$, where $\mathcal{S}$ denotes the set of all subsets of $\{1, \ldots, M\}$ of size at most K (including the empty subset), and N

is a stopping time. The maximal error probability will now be defined as

$$P_{\max} \triangleq \max_{S \in \mathcal{S}} \mathbb{P}_S \left\{ \delta \left(\boldsymbol{Y}^N \right) \neq S \right\}. \tag{4.47}$$

A sequence of tests is *universally consistent* if the maximal error probability converges to zero for any $\mu, \pi, \mu \neq \pi$. The notion of *universally exponential consistency* can be defined in the same manner as that in (4.1).

As in Section 4.1, for the setting with both the typical and outlier distributions being known and with $\hat{S}\left(\boldsymbol{y}^n\right) \triangleq \operatorname*{argmax}_{S \in \mathcal{S}} p_S\left(\boldsymbol{y}^n\right)$, the MSPRT with the stopping and final decision rules being

$$N^* = \operatorname*{argmin}_{n \geq 1} \left[\frac{p_{\hat{S}}\left(\boldsymbol{Y}^n\right)}{\max\limits_{S \neq \hat{S}} p_S\left(\boldsymbol{Y}^n\right)} > T \right], \tag{4.48}$$

$$\delta^* = \hat{S}\left(\boldsymbol{Y}^{N^*}\right), \tag{4.49}$$

is asymptotically optimal (cf.[26, 27]).

Proposition 16. *As the threshold T in (4.48) approaches infinity, the MSPRT in (4.48) and (4.49) satisfies*

$$P_{\max} \leq O\left(\frac{1}{T}\right).$$

In addition, as $T \to \infty$, for each $S \in \mathcal{S}, |S| = K$,

$$\mathbb{E}_S\left[N^*\right] = \frac{\log T}{D\left(\mu \| \pi\right)}(1 + o(1)) = \frac{-\log P_{\max}}{D\left(\mu \| \pi\right)}(1 + o(1));$$

for each $S \in \mathcal{S}, 1 \leq |S| < K$,

$$\mathbb{E}_S\left[N^*\right] = \frac{\log T}{\min\left(D\left(\mu \| \pi\right), D\left(\pi \| \mu\right)\right)}(1 + o(1)) = \frac{-\log P_{\max}}{\min\left(D\left(\mu \| \pi\right), D\left(\pi \| \mu\right)\right)}(1 + o(1));$$

and

$$\mathbb{E}_0\left[N^*\right] = \frac{\log T}{D\left(\pi \| \mu\right)}(1 + o(1)) = \frac{-\log P_{\max}}{D\left(\pi \| \mu\right)}(1 + o(1)).$$

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of

tests (N, δ) *with vanishing maximal error probability, it holds for each* $S \in \mathcal{S}, |S| = K$*, that*

$$\mathbb{E}_S[N] \;\geq\; \frac{-\log P_{\max}}{D\left(\mu \| \pi\right)}(1 + o(1));$$

for each $S \in \mathcal{S}, 1 \leq |S| < K$*, that*

$$\mathbb{E}_S[N] \;\geq\; \frac{-\log P_{\max}}{\min\left\{D\left(\mu \| \pi\right), D\left(\pi \| \mu\right)\right\}}(1 + o(1));$$

and that

$$\mathbb{E}_0[N] \;\geq\; \frac{-\log P_{\max}}{D\left(\pi \| \mu\right)}(1 + o(1)).$$

4.2.1 Proposed Universal Test

When only π is known, we compute the GL of $\boldsymbol{y}^n$ under each non-null hypothesis corresponding to a non-empty subset $S \subset \{1, \ldots, M\}$ by replacing the unknown μ in (3.54) with its ML estimate $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$. Similarly, when neither π nor μ is known, we compute the GL of $\boldsymbol{y}^n$ under each non-null hypothesis corresponding to a non-empty $S \in \mathcal{S}$ by replacing the unknown μ and π in (3.54) with their ML estimates $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$, and $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} \gamma_j}{M - |S|}$, respectively.

When only π is known and with

$$\hat{S}\left(\boldsymbol{y}^n\right) \;\triangleq\; \underset{S \in \mathcal{S}, S \neq \emptyset}{\operatorname{argmax}}\; \hat{p}_S^{\text{typ}}\left(\boldsymbol{y}^n\right) = \underset{S \in \mathcal{S}, S \neq \emptyset}{\operatorname{argmin}} \Big[\sum_{i \in S} D\Big(\gamma_i \| \tfrac{\sum_{k \in S} \gamma_k}{|S|}\Big) + \sum_{j \notin S} D(\gamma_j \| \pi) \Big]$$

denoting the instantaneous estimate of the non-null hypothesis (using the GL) at time n, our proposed universal test can be described by the following stopping and final decision rules:

$$N^* \;=\; \min\left(\tilde{N}, \lfloor T \log T \rfloor\right), \tag{4.50}$$

$$\delta^* \;=\; \begin{cases} \hat{S}\left(\boldsymbol{Y}^{N^*}\right) & \text{if } \tilde{N} \leq T \log T \\ 0 & \text{if } \tilde{N} > T \log T, \end{cases} \tag{4.51}$$

where

$$\tilde{N} \triangleq \underset{n\geq 1}{\operatorname{argmin}} \left[\min_{\substack{S'\neq\hat{S}\\ S'\neq\emptyset}} n \left[\sum_{i\in S'} D\left(\gamma_i \middle\| \tfrac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j\middle\|\pi\right) - \sum_{i\in\hat{S}} D\left(\gamma_i\middle\|\tfrac{\sum_{k\in\hat{S}}\gamma_k}{|\hat{S}|}\right) - \sum_{j\notin\hat{S}} D\left(\gamma_j\middle\|\pi\right) \right] > \log T + (M+1)|\mathcal{Y}|\log(n+1) \right]. \quad (4.52)$$

Similarly, when neither μ nor π is known, the test can be written as in (4.50) and (4.51) but with

$$\hat{S}\left(\boldsymbol{y}^n\right) \triangleq \underset{S\in\mathcal{S},S\neq\emptyset}{\operatorname{argmax}} \hat{p}_S^{\text{univ}}\left(\boldsymbol{y}^n\right) = \underset{S\in\mathcal{S},S\neq\emptyset}{\operatorname{argmin}} \Big[\sum_{i\in S} D\Big(\gamma_i\|\tfrac{\sum_{k\in S}\gamma_k}{|S|}\Big) + \sum_{j\notin S} D\Big(\gamma_j\|\tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\Big)\Big], \quad (4.53)$$

and

$$\tilde{N} \triangleq \underset{n\geq 1}{\operatorname{argmin}} \left[\min_{\substack{S'\neq\hat{S}\\ S'\neq\emptyset}} n \left[\sum_{i\in S'} D\left(\gamma_i \middle\| \tfrac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j\middle\|\tfrac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) - \sum_{i\in\hat{S}} D\left(\gamma_i\middle\|\tfrac{\sum_{k\in\hat{S}}\gamma_k}{|\hat{S}|}\right) - \sum_{j\notin\hat{S}} D\left(\gamma_j\middle\|\tfrac{\sum_{k\notin\hat{S}}\gamma_k}{M-|\hat{S}|}\right) \right] > \log T + (M+1)|\mathcal{Y}|\log(n+1) \right]. \quad (4.54)$$

4.2.2 Performance of Proposed Test

Theorem 17. *When only π is known, the test in (4.50), (4.51), (4.52) is universally consistent, and yields for every T that*

$$P_{\max} \leq O\left(\frac{1}{T}\right). \quad (4.55)$$

In addition, for each non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$, as $T \to \infty$,

$$\mathbb{E}_S[N^*] \;=\; \frac{\log T}{\alpha_S}(1+o(1)) \tag{4.56}$$

$$\leq \begin{cases} \frac{-\log P_{\max}}{D(\mu\|\pi)}(1+o(1)), & |S|=K; \\ \frac{-\log P_{\max}}{\min(D(\mu\|\pi),\eta_S(\mu\|\pi))}(1+o(1)), & 1\leq |S|<K, \end{cases} \tag{4.57}$$

where

$$\alpha_S \;\triangleq\; \min_{\substack{S'\neq S \\ S'\neq\emptyset}} \Bigg[|S\cap S'| D\left(\mu\Big\| \tfrac{|S\cap S'|\mu+|S'\backslash S|\pi}{|S'|}\right) + \big|S\backslash S'\big| D(\mu\|\pi) + \big|S'\backslash S\big| D\left(\pi\Big\| \tfrac{|S\cap S'|\mu+|S'\backslash S|\pi}{|S'|}\right) \Bigg] \;>\; 0, \tag{4.58}$$

and

$$\eta_S(\mu\|\pi) \;\triangleq\; |S| D\left(\mu\Big\|\frac{|S|\mu+\pi}{|S|+1}\right) + D\left(\pi\Big\|\frac{|S|\mu+\pi}{|S|+1}\right). \tag{4.59}$$

First define for each $S \in \mathcal{S}, S \neq \emptyset$,

$$\tilde{N}_S \;\triangleq\; \operatorname*{argmin}_{n\geq 1} \left(\min_{\substack{S'\neq S \\ S'\neq\emptyset}} n \left[\sum_{i\in S'} D\left(\gamma_i\Big\|\tfrac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j\|\pi\right) - \sum_{i\in S} D\left(\gamma_i\Big\|\tfrac{\sum_{k\in S}\gamma_k}{|S|}\right) - \sum_{j\notin S} D\left(\gamma_j\|\pi\right) \right] > \log T + (M+1)|\mathcal{Y}|\log(n+1) \right). \tag{4.60}$$

The proof relies on the following two lemmas.

Lemma 8. *Under every non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$, and each $n \geq 1$, it holds that*

$$\mathbb{P}_S\{\tilde{N}_S \geq n\} \;\leq\; (M+1)\, M^K T n^{(2M+1)|\mathcal{Y}|} e^{-(n-1)b}, \tag{4.61}$$

where $b > 0$ is a function only of μ and π.

Proof. It follows by the definition of $\tilde{N}_S$ in (4.60) and (2.7) that

$$
\begin{aligned}
&\mathbb{P}_S\{\tilde{N}_S \geq n\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} (n-1)\left[\sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right) \right. \\ \quad \left. - \sum\limits_{i\in S} D\left(\gamma_i \Big\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) - \sum\limits_{j\notin S} D\left(\gamma_j \Big\| \pi\right)\right] \leq \log T + (M+1)|\mathcal{Y}| \log n \end{array} \right\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum\limits_{i\in S} D\left(\gamma_i \Big\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum\limits_{j\notin S} D\left(\gamma_j \| \pi\right) \\ \quad \geq \ -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}| \log n\right) + \sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right) \end{array} \right\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum\limits_{i\in S} D\left(\gamma_i \| \mu\right) + \sum\limits_{j\notin S} D\left(\gamma_j \| \pi\right) \\ \quad \geq \ -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}| \log n\right) + \sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right) \end{array} \right\} \qquad (4.62) \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum\limits_{i\in S} D\left(\gamma_i \| \mu\right) + \sum\limits_{j\notin S} D\left(\gamma_j \Big\| \pi\right) \\ \quad \geq -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}| \log n\right) + \sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right) \\ D(\gamma_i\|\mu) \ \leq \ \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j\|\pi) \ \leq \ \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + \sum_{S' \neq S} \mathbb{P}_S\Big\{ D(\gamma_i\|\mu) \ > \ \epsilon \text{ for some } i \in S, \text{ or } D(\gamma_j\|\pi) \ > \ \epsilon \text{ for some } j \notin S \Big\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum\limits_{i\in S} D\left(\gamma_i \| \mu\right) + \sum\limits_{j\notin S} D\left(\gamma_j \| \pi\right) \\ \quad \geq -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}| \log n\right) + \sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right) \\ D(\gamma_i\|\mu) \ \leq \ \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j\|\pi) \ \leq \ \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + M^K M n^{|\mathcal{Y}|} e^{-(n-1)\epsilon}. \qquad (4.63)
\end{aligned}
$$

Note that the term $\sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \| \pi\right)$ is zero only if $\gamma_i = \gamma$ for all $i \in S'$, for some γ and $\gamma_j = \pi$ for all $j \notin S'$. As in the event whose probability is concerned in (4.63), it also holds that $D(\gamma_i\|\mu) \leq \epsilon$ for all $i \in S$, and $D(\gamma_j\|\pi) \leq \epsilon$ for all $j \notin S$, attaining this zero value cannot happen if ϵ in (4.63) is chosen sufficiently small, since $S' \neq S, \ S' \neq \emptyset$. We conclude that when ϵ is chosen to be sufficiently small (as a function of (μ, π)), it holds therein that $\sum\limits_{i\in S'} D\left(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum\limits_{j\notin S'} D\left(\gamma_j \Big\| \pi\right) \ > \ a(\mu,\pi) \ > \ 0$. Continuing from (4.63)

with the ϵ chosen sufficiently small, we get that

$$\begin{aligned}
&\mathbb{P}_S\{\tilde{N}_S \geq n\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum\limits_{i \in S} D(\gamma_i \| \mu) + \sum\limits_{j \notin S} D(\gamma_j \| \pi) \ \geq \ -\frac{1}{n-1} \left(\log T + (M+1)|\mathcal{Y}| \log n \right) + a\big(\mu, \pi\big) \\ D(\gamma_i \| \mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j \| \pi) \leq \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon} \\
&\leq M^K T n^{(2M+1)|\mathcal{Y}|} e^{-a(n-1)} + M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon} \\
&\leq (M+1)\, M^K T n^{(2M+1)|\mathcal{Y}|} e^{-(n-1)\min(a,\epsilon)}. \qquad (4.64)
\end{aligned}$$

□

Lemma 9. *Under each non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$,*

$$\lim_{T \to \infty} \mathbb{E}_S \left[\left| \frac{\tilde{N}_S}{\log T} - \frac{1}{\alpha_S} \right| \right] \ = \ 0, \qquad (4.65)$$

where α_S is defined in (4.58).

Proof. Under hypothesis $S \in \mathcal{S}, S \neq \emptyset$, the strong law of large numbers yields that as $n \to \infty$, $\gamma_i^{(n)} \to \mu$ a.s. for every $i \in S$, and $\gamma_j^{(n)} \to \pi$ a.s. for every $j \notin S$, hence, we obtain that under $\mathbb{P}_S$,

$$\begin{aligned}
&\sum_{i \in S'} D\left(\gamma_i^{(n)} \Big\| \tfrac{\sum_{k \in S'} \gamma_k^{(n)}}{|S'|} \right) + \sum_{j \notin S'} D\left(\gamma_j^{(n)} \| \pi \right) \ \overset{\text{a.s.}}{\to} \\
&\qquad |S \cap S'| D\left(\mu \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S| \pi}{|S'|} \right) + |S \backslash S'| D(\mu \| \pi) + |S' \backslash S| D\left(\pi \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S| \pi}{|S'|} \right), \qquad (4.66)
\end{aligned}$$

as $n \to \infty$. Taking minimum over $S' \neq S$ on both sides of (4.66), we see that under $\mathbb{P}_S$,

$$\min_{S' \neq S} \sum_{i \in S'} D\left(\gamma_i \Big\| \tfrac{\sum_{k \in S'} \gamma_k}{|S'|} \right) + \sum_{j \notin S'} D\left(\gamma_j \| \pi \right) \ \to \ \alpha_S \quad \text{a.s.},$$

as $n \to \infty$, where α_S is defined in (4.58).

By Lemma 8, we see that $\tilde{N}_S$ is finite a.s. under $\mathbb{P}_S$, $S \in \mathcal{S}, S \neq \emptyset$. It then follows from

this a.s. finiteness and the definition of $\tilde{N}_S$ in (4.60) that with probability 1 under $\mathbb{P}_S$,

$$\min_{S' \neq S} \left[\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S)} \Big\| \frac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S)}}{|S'|} \right) + \sum_{j \notin S'} D\left(\gamma_j^{(\tilde{N}_S)} \Big\| \pi \right) \right.$$
$$\left. - \sum_{i \in S} D\left(\gamma_i^{(\tilde{N}_S)} \Big\| \frac{\sum_{k \in S} \gamma_k^{(\tilde{N}_S)}}{|S|} \right) - \sum_{j \notin S} D\left(\gamma_j^{(\tilde{N}_S)} \Big\| \pi \right) \right] > \frac{\log T + (M+1)|\mathcal{Y}| \log(\tilde{N}_S + 1)}{\tilde{N}_S}; \tag{4.67}$$

$$\min_{S' \neq S} \left[\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S - 1)} \Big\| \frac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S - 1)}}{|S'|} \right) + \sum_{j \notin S'} D\left(\gamma_j^{(\tilde{N}_S - 1)} \Big\| \pi \right) \right.$$
$$\left. - \sum_{i \in S} D\left(\gamma_i^{(\tilde{N}_S - 1)} \Big\| \frac{\sum_{k \in S} \gamma_k^{(\tilde{N}_S - 1)}}{|S|} \right) - \sum_{j \notin S} D\left(\gamma_j^{(\tilde{N}_S - 1)} \Big\| \pi \right) \right] \leq \frac{\log T + (M+1)|\mathcal{Y}| \log \tilde{N}_S}{\tilde{N}_S - 1}. \tag{4.68}$$

The claim in (4.65) now follows from (4.66), (4.67) and (4.68) if we can establish that under $\mathbb{P}_S$, $\tilde{N}_S \to \infty$, a.s., as $T \to \infty$, and the uniform integrability of the sequence $\frac{\tilde{N}_S}{\log T}$. To this end, we have from (4.67) that for any $n \geq 1$, and any $S' \neq S$, $S' \neq \emptyset$,

$$\begin{aligned}
&\mathbb{P}\{\tilde{N}_S \leq n\} \\
&\leq \mathbb{P}_S \Big\{ \tilde{N}_S \Big[\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S)} \Big\| \frac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S)}}{|S'|} \right) + \sum_{k \notin S'} D\left(\gamma_j^{(\tilde{N}_S)} \Big\| \pi \right) > \log T;\ \tilde{N}_S \leq n \Big\} \\
&\leq \mathbb{P}_S \Big\{ n \Big(M \max \left(\log M, \log \left(\frac{1}{\min_y \pi(y)} \right) \right) \Big) > \log T \Big\} \\
&= 0, \quad \text{for every } n < \frac{\log T}{M \max \left(\log M, \log \left(\frac{1}{\min_y \pi(y)} \right) \right)},
\end{aligned} \tag{4.69}$$

thereby yielding that $\tilde{N}_S \to \infty$ a.s. as $T \to \infty$. Using the exponential tail bound in Lemma 8 to establish the uniform integrability of the sequence $\frac{\tilde{N}_S}{\log T}$ similarly as in the previous proofs (details skipped), we obtain (4.65). □

Proof. We now proceed to prove (4.55). First, note that for any $S, S' \in \mathcal{S}$, $S \neq S', S, S' \neq \emptyset$,

we get from (2.7) that

$$
\begin{aligned}
&\mathbb{P}_S\{\delta^* = S'\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\{N^* = n, \delta^* = S'\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S \left\{ \begin{array}{l} n\Big[\sum\limits_{i\in S} D\left(\gamma_i \Big\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum\limits_{j\notin S} D\left(\gamma_j\|\pi\right) - \sum\limits_{i\in S'} D\Big(\gamma_i \Big\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\Big) + \sum\limits_{j\notin S'} D\left(\gamma_j\|\pi\right)\Big] \\ \quad > \log T + (M+1)|\mathcal{Y}|\log(n+1) \end{array} \right\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\Big\{ n\Big[\sum_{i\in S} D\left(\gamma_i \Big\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum_{j\notin S} D\left(\gamma_j\|\pi\right)\Big] > \log T + (M+1)|\mathcal{Y}|\log(n+1)\Big\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\Big\{ \sum_{i\in S} D\left(\gamma_i\|\mu\right) + \sum_{j\notin S} D\left(\gamma_j\|\pi\right) > \frac{1}{n}\left(\log T + (M+1)|\mathcal{Y}|\log\left(n+1\right)\right)\Big\} \qquad (4.70) \\
&\leq \sum_{n=1}^{\infty} \frac{1}{T}(n+1)^{-|\mathcal{Y}|} \\
&= \frac{1}{T}C'(|\mathcal{Y}|). \qquad (4.71)
\end{aligned}
$$

In addition, for each $S \in \mathcal{S}, S \neq \emptyset$, we obtain from (4.65), the Markov inequality, and that $\tilde{N} \leq \tilde{N}_S$ with probability 1, that

$$
\mathbb{P}_S\{\delta^* = 0\} \;=\; \mathbb{P}_S\{\tilde{N} > T\log T\} \;\leq\; \frac{C'(\mu,\pi,|\mathcal{Y}|,M,K)}{T}. \qquad (4.72)
$$

Next, it follows from the definitions of $\tilde{N}$ and N^* in (4.52), (4.50) and (2.7) that

$$
\begin{aligned}
&\mathbb{P}_0\{\delta^* \neq 0\} \\
&= \mathbb{P}_0\{N^* = \tilde{N}\} \\
&= \mathbb{P}_0\{\tilde{N} \leq T\log T\} \\
&\leq \mathbb{P}_0\{\tilde{N} \text{ is finite}\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} = n\} \\
&\leq \sum_{n=1}^{\infty} \sum_{S} \mathbb{P}_0 \left\{ \begin{array}{l} n\left[\sum\limits_{i\in S} D\left(\gamma_i \Big\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum\limits_{j\notin S} D\left(\gamma_j\|\pi\right)\right] \\ \quad > \log T + (M+1)|\mathcal{Y}|\log(n+1) \end{array} \right\}
\end{aligned}
$$

$$\leq \sum_{n=1}^{\infty} \sum_{S} \mathbb{P}_0 \Big\{ \sum_{i \in S} D\left(\gamma_i \| \pi\right) + \sum_{j \notin S} D\left(\gamma_j \| \pi\right) > \frac{1}{n} \left(\log T + (M+1)|\mathcal{Y}| \log\left(n+1\right)\right) \Big\} \tag{4.73}$$

$$\leq \frac{1}{T} M^K \sum_{n=1}^{\infty} (n+1)^{-|\mathcal{Y}|}$$

$$= \frac{1}{T} C'(|\mathcal{Y}|, M, K). \tag{4.74}$$

The claim in (4.55) now follows from (4.71), (4.72) and (4.74).

The claim in (4.56) follows as previously from (4.65), (4.71), (4.72) and from the fact that for each $S \in \mathcal{S},\ S \neq \emptyset,\ N^* \leq \tilde{N}_S$ with probability 1.

It is now left to prove (4.57). First observe that when $|S| = K$, it holds for any $S' \in \mathcal{S}, S' \neq S, S' \neq \emptyset$, that $|S \backslash S'| \geq 1$. Consequently, we get when $|S| = K$ that

$$\alpha_S = \min_{\substack{S' \neq S \\ S' \neq \emptyset}} \Big[|S \cap S'| D\left(\mu \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S \backslash S'\right| D(\mu \| \pi) + \left|S' \backslash S\right| D\left(\pi \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) \Big] = D(\mu \| \pi), \tag{4.75}$$

where the last equality above is attained by selecting S' to be S but one (any) element.

We next consider the case with $1 \leq |S| < K$. Then, for any $S' \in \mathcal{S},\ S' \neq \emptyset$, such that $S \backslash S' \neq \emptyset$, the term inside the minimum on the right side of (4.75) is still lower bounded by $D(\mu \| \pi)$. Now for any other $S' \neq S,\ S' \neq \emptyset$, such that $S' \supset S$, it follows that

$$\begin{aligned} & |S \cap S'| D\left(\mu \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S \backslash S'\right| D(\mu \| \pi) + \left|S' \backslash S\right| D\left(\pi \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) \\ & \geq |S| D\left(\mu \Big\| \tfrac{|S|\mu + |S' \backslash S|\pi}{|S'|}\right) + D\left(\pi \Big\| \tfrac{|S|\mu + |S' \backslash S|\pi}{|S'|}\right) \\ & \geq \min_{p \in \mathcal{P}(\mathcal{Y})} |S| D(\mu \| p) + D(\pi \| p) \\ & = |S| D\left(\mu \Big\| \frac{|S|\mu + \pi}{|S| + 1}\right) + D\left(\pi \Big\| \frac{|S|\mu + \pi}{|S| + 1}\right) = \eta_S(\mu \| \pi), \end{aligned} \tag{4.76}$$

where (4.76) follows from (2.7) with $\mathcal{C}$ therein comprising $|S|$ copies of μ and one π. The combination of (4.75) and (4.76) constitutes the claim in (4.57). □

Theorem 18. *When neither μ nor π is known, the universal test in (4.50), (4.51) and (4.54) is universally consistent, and yields for every T that*

$$P_{\max} \leq O\left(\frac{1}{T}\right). \tag{4.77}$$

In addition, for each non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$, as $T \to \infty$,

$$\mathbb{E}_S[N^*] = \frac{\log T}{\overline{\alpha}_S}(1+o(1)) \tag{4.78}$$

$$\leq \begin{cases} \frac{-\log P_{\max}}{\overline{\eta}(\mu\|\pi)}(1+o(1)), & |S| = K; \\ \frac{-\log P_{\max}}{\min(\overline{\eta}(\mu\|\pi), \eta_S(\mu\|\pi))}(1+o(1)), & 1 \leq |S| < K, \end{cases} \tag{4.79}$$

where

$$\overline{\alpha}_S \triangleq \min_{\substack{S' \neq S \\ S' \neq \emptyset}} \Bigg[|S \cap S'| D\left(\mu \Big\| \tfrac{|S\cap S'|\mu + |S'\setminus S|\pi}{|S'|}\right) + |S \setminus S'| D\left(\mu \Big\| \tfrac{|S\setminus S'|\mu + |S^c \cap S'^c|\pi}{M-|S'|}\right) + |S' \setminus S| D\left(\pi \Big\| \tfrac{|S\cap S'|\mu + |S'\setminus S|\pi}{|S'|}\right) + |S^c \cap S'^c| D\left(\pi \Big\| \tfrac{|S\setminus S'|\mu + |S^c\cap S'^c|\pi}{M-|S'|}\right) \Bigg] > 0, \tag{4.80}$$

and

$$\overline{\eta}_S(\mu\|\pi) \triangleq D\left(\mu \Big\| \frac{\mu + (M-K-|S|)\pi}{M-K-|S|+1}\right) + (M-K-|S|) D\left(\pi \Big\| \frac{\mu + (M-K-|S|)\pi}{M-K-|S|+1}\right), \tag{4.81}$$

and $\eta_S(\mu\|\pi)$ is as in (4.59).

First define for each $S \in \mathcal{S}, S \neq \emptyset$

$$\tilde{N}_S \triangleq \operatorname*{argmin}_{n \geq 1} \left(\min_{\substack{S' \neq S \\ S' \neq \emptyset}} n \left[\sum_{i \in S'} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) - \sum_{i\in S} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S}\gamma_k}{|S|}\right) - \sum_{j\notin S} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\right) \right] > \log T + (M+1)|\mathcal{Y}| \log(n+1). \right) \tag{4.82}$$

The proof relies on the following two lemmas.

Lemma 10. *Under every non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$, and each $n \geq 1$,*

$$\mathbb{P}_S\{\tilde{N}_S \geq n\} \leq (M+1) M^K T n^{(2M+1)|\mathcal{Y}|} e^{-(n-1)b}, \tag{4.83}$$

where $b > 0$ is a function only of μ and π.

Proof. We get from (2.7) and (4.82) that

$$
\begin{aligned}
&\mathbb{P}_S\{\tilde{N}_S \geq n\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} (n-1)\left[\sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right)\right. \\ \left. - \sum_{i\in S} D\left(\gamma_i \middle\| \frac{\sum_{k\in S}\gamma_k}{|S|}\right) - \sum_{j\notin S} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S}\gamma_k}{M-|S|}\right)\right] \leq \log T + (M+1)|\mathcal{Y}|\log n \end{array} \right\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum_{i\in S} D\left(\gamma_i \| \mu\right) + \sum_{j\notin S} D\left(\gamma_j \| \pi\right) \geq -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}|\log n\right) \\ \qquad\qquad + \sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) \end{array} \right\} \qquad (4.84) \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum_{i\in S} D\left(\gamma_i \| \mu\right) + \sum_{j\notin S} D\left(\gamma_j \middle\| \pi\right) \geq -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}|\log n\right) \\ \qquad + \sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) \\ D(\gamma_i\|\mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j\|\pi) \leq \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + \sum_{S' \neq S} \mathbb{P}_S \Big\{ D(\gamma_i\|\mu) > \epsilon \text{ for some } i \in S, \text{ or } D(\gamma_j\|\pi) > \epsilon \text{ for some } j \notin S \Big\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum_{i\in S} D\left(\gamma_i \| \mu\right) + \sum_{j\notin S} D\left(\gamma_j \middle\| \pi\right) \geq -\frac{1}{n-1}\left(\log T + (M+1)|\mathcal{Y}|\log n\right) \\ \qquad + \sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) \\ D(\gamma_i\|\mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j\|\pi) \leq \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + M^K M n^{|\mathcal{Y}|} e^{-(n-1)\epsilon}. \qquad (4.85)
\end{aligned}
$$

Similar to the previous proofs, since $S' \neq S,\ S' \neq \emptyset$, $\sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right)$ is zero only if $\gamma_i = \gamma$ for all $i \in S'$, for some γ and $\gamma_j = \gamma'$ for all $j \notin S'$, for some γ'. For sufficiently small ϵ, in the event whose probability is concerned in (4.85), attaining this zero value cannot happen, because it also holds that $D(\gamma_i\|\mu) \leq \epsilon$ for all $i \in S$, and $D(\gamma_j\|\pi) \leq \epsilon$ for all $j \notin S$. We conclude that when ϵ is chosen to be sufficiently small (as a function of (μ, π)), it holds that $\left[\sum_{i\in S'} D\left(\gamma_i \middle\| \frac{\sum_{k\in S'}\gamma_k}{|S'|}\right) + \sum_{j\notin S'} D\left(\gamma_j \middle\| \frac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right)\right] \geq a(\mu,\pi) > 0.$

Continuing from (4.85) with ϵ chosen sufficiently small, we get that

$$\begin{aligned}
&\mathbb{P}_S\{\tilde{N}_S \geq n\} \\
&\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \begin{array}{l} \sum_{i \in S} D(\gamma_i \| \mu) + \sum_{j \notin S} D(\gamma_j \| \pi) \ \geq \ -\frac{1}{n-1} \left(\log T + (M+1)|\mathcal{Y}| \log n\right) + a\big(\mu, \pi\big) \\ D\left(\gamma_i \| \mu\right) \leq \epsilon \text{ for every } i \in S, \text{ and } D\left(\gamma_j \| \pi\right) \leq \epsilon \text{ for every } j \notin S \end{array} \right\} \\
&\quad + M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon} \\
&\leq \ M^K T n^{(2M+1)|\mathcal{Y}|} e^{-a(n-1)} + M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon} \\
&\leq \ (M+1) M^K T n^{(2M+1)|\mathcal{Y}|} e^{-(n-1)\min(a,\epsilon)}. \qquad (4.86)
\end{aligned}$$

□

Lemma 11. *Under each non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$,*

$$\lim_{T \to \infty} \mathbb{E}_S \left[\left| \frac{\tilde{N}_S}{\log T} - \frac{1}{\overline{\alpha}_S} \right| \right] \ = \ 0, \qquad (4.87)$$

where $\overline{\alpha}_S$ is defined in (4.80).

Proof. Since under hypothesis $S \in \mathcal{S}, S \neq \emptyset$, $\gamma_i \to \mu$ a.s., for every $i \in S$, and $\gamma_j \to \pi$ a.s., for every $j \notin S$, we get that under $\mathbb{P}_S$,

$$\begin{aligned}
&\sum_{i \in S'} D\left(\gamma_i^{(n)} \Big\| \tfrac{\sum_{k \in S'} \gamma_k^{(n)}}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j^{(n)} \Big\| \tfrac{\sum_{k \notin S'} \gamma_k^{(n)}}{M - |S'|}\right) \ \overset{\text{a.s.}}{\to} \\
&\qquad \big|S \cap S'\big| D\left(\mu \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \big|S \backslash S'\big| D\left(\mu \Big\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right) \\
&\qquad + \big|S' \backslash S\big| D\left(\pi \Big\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \big|S^c \cap S'^c\big| D\left(\pi \Big\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right), \qquad (4.88)
\end{aligned}$$

as $n \to \infty$. Taking minimum over $S' \in \mathcal{S}$ on both sides of (4.88), we get that under $\mathbb{P}_S$,

$$\min_{S' \neq S} \sum_{i \in S'} D\left(\gamma_i \Big\| \tfrac{\sum_{k \in S'} \gamma_k}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j \Big\| \tfrac{\sum_{k \notin S'} \gamma_k}{M - |S'|}\right) \ \to \ \overline{\alpha}_S \quad \text{a.s.}$$

as $n \to \infty$.

By Lemma 10, we get that $\tilde{N}_S$ is finite a.s. under $\mathbb{P}_S$, $S \in \mathcal{S}, S \neq \emptyset$. It then follows from

this a.s. finiteness and the definition of $\tilde{N}_S$ in (4.82) that with probability 1 under $\mathbb{P}_S$,

$$\min_{S' \neq S} \left[\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S)} \| \tfrac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S)}}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j^{(\tilde{N}_S)} \| \tfrac{\sum_{k \notin S'} \gamma_k^{(\tilde{N}_S)}}{M-|S'|}\right) \right.$$
$$\left. - \sum_{i \in S} D\left(\gamma_i^{(\tilde{N}_S)} \| \tfrac{\sum_{k \in S} \gamma_k^{(\tilde{N}_S)}}{|S|}\right) - \sum_{j \notin S} D\left(\gamma_j^{(\tilde{N}_S)} \| \tfrac{\sum_{k \notin S} \gamma_k^{(\tilde{N}_S)}}{M-|S|}\right) \right]$$
$$> \frac{\log T + (M+1)|\mathcal{Y}| \log(\tilde{N}_S + 1)}{\tilde{N}_S}; \qquad (4.89)$$

$$\min_{S' \neq S} \left[\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S - 1)} \| \tfrac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S - 1)}}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j^{(\tilde{N}_S - 1)} \| \tfrac{\sum_{k \notin S'} \gamma_k^{(\tilde{N}_S - 1)}}{M-|S'|}\right) \right.$$
$$\left. - \sum_{i \in S} D\left(\gamma_i^{(\tilde{N}_S - 1)} \| \tfrac{\sum_{k \in S} \gamma_k^{(\tilde{N}_S - 1)}}{|S|}\right) - \sum_{j \notin S} D\left(\gamma_j^{(\tilde{N}_S - 1)} \| \tfrac{\sum_{k \notin S} \gamma_k^{(\tilde{N}_S - 1)}}{M-|S|}\right) \right]$$
$$\leq \frac{\log T + (M+1)|\mathcal{Y}| \log \tilde{N}_S}{\tilde{N}_S - 1}. \qquad (4.90)$$

The a.s. convergence of $\frac{\tilde{N}_S}{\log T}$ to $\frac{1}{\bar{\alpha}_S}$ follows from (4.88), (4.89) and (4.90) if we can establish that under each hypothesis $S \in \mathcal{S}$, $\tilde{N}_S \to \infty$, a.s. This can be established similarly as in the previous proofs upon noting for any $S' \neq S,\ S' \neq \emptyset$,

$$\sum_{i \in S'} D\left(\gamma_i^{(\tilde{N}_S)} \| \tfrac{\sum_{k \in S'} \gamma_k^{(\tilde{N}_S)}}{|S'|}\right) + \sum_{j \notin S'} D\left(\gamma_j^{(\tilde{N}_S)} \| \tfrac{\sum_{k \notin S'} \gamma_k^{(\tilde{N}_S)}}{M-|S'|}\right) \ \leq \ M \log M.$$

The proof of (4.87) follows as previously from using Lemma 10 to prove the uniform integrability of the sequence $\frac{\tilde{N}_S}{\log T}$. □

Proof. We now proceed to prove (4.77). First, note that for any $S, S' \in \mathcal{S}, S \neq S',\ S, S' \neq \emptyset$, we get from (2.7) that

$$\mathbb{P}_S\{\delta^* = S'\}$$
$$\leq \sum_{n=1}^{\infty} \mathbb{P}_S\{N^* = \tilde{N} = n,\ \delta^* = S'\}$$

$$
\begin{aligned}
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\Bigg\{ n\Bigg[\sum_{i\in S} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum_{j\notin S} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\right) \\
&\qquad\qquad - \sum_{i\in S'} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S'}\gamma_k}{|S'|}\right) - \sum_{j\notin S'} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right)\Bigg] > \log T + (M+1)|\mathcal{Y}|\log(n+1)\Bigg\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\Bigg\{ n\Bigg[\sum_{i\in S} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum_{j\notin S} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\right)\Bigg] > \log T + (M+1)|\mathcal{Y}|\log(n+1)\Bigg\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_S\Bigg\{ \sum_{i\in S} D\left(\gamma_i \| \mu\right) + \sum_{j\notin S} D\left(\gamma_j \| \pi\right) > \frac{1}{n}\left(\log T + (M+1)|\mathcal{Y}|\log\left(n+1\right)\right)\Bigg\} \\
&\leq \frac{1}{T}\sum_{n=1}^{\infty} (n+1)^{-|\mathcal{Y}|} \\
&= \frac{1}{T} C'(|\mathcal{Y}|). \qquad (4.91)
\end{aligned}
$$

Also, for each $S \in \mathcal{S}, S \neq \emptyset$, we get from (4.87), the Markov inequality, and that $\tilde{N} \leq \tilde{N}_S$ with probability 1, that

$$
\mathbb{P}_S\{\delta^* = 0\} \;=\; \mathbb{P}_S\{\tilde{N} > T\log T\} \;\leq\; \frac{C'(\mu,\pi,|\mathcal{Y}|,M,K)}{T}. \qquad (4.92)
$$

Next, it follows from the definitions of $N^*, \tilde{N}$ in (4.50), (4.54), and (2.7) that

$$
\begin{aligned}
&\mathbb{P}_0\{\delta^* \neq 0\} \\
&= \mathbb{P}_0\{N^* = \tilde{N}\} \\
&= \mathbb{P}_0\{\tilde{N} \leq T\log T\} \\
&\leq \mathbb{P}_0\{\tilde{N} \text{ is finite}\} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} = n\} \\
&\leq \sum_{n=1}^{\infty}\sum_{S} \mathbb{P}_0\Bigg\{ n\Bigg[\sum_{i\in S} D\left(\gamma_i \Big\| \tfrac{\sum_{k\in S}\gamma_k}{|S|}\right) + \sum_{j\notin S} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\right) \geq \\
&\qquad\qquad\qquad\qquad\qquad\qquad \log T + (M+1)|\mathcal{Y}|\log(n+1)\Bigg\} \\
&\leq \sum_{n=1}^{\infty}\sum_{S} \mathbb{P}_0\Bigg\{ \sum_{i\in S} D\left(\gamma_i \| \pi\right) + \sum_{j\notin S} D\left(\gamma_j \| \pi\right) \geq \frac{1}{n}\left(\log T + (M+1)|\mathcal{Y}|\log\left(n+1\right)\right)\Bigg\}
\end{aligned} \qquad (4.93)
$$

$$\leq \frac{1}{T} M^K \sum_{n=1}^{\infty} (n+1)^{-|\mathcal{Y}|}$$
$$= \frac{1}{T} C'(|\mathcal{Y}|, M, K). \tag{4.94}$$

The combination of (4.91), (4.92) and (4.94) constitutes (4.77).

The claim in (4.78) follows as previously from (4.87), (4.91), (4.92) and from the fact that for each $S \in \mathcal{S}, S \neq \emptyset,\ N^* \leq \tilde{N}_S$ with probability 1.

It is now left to prove (4.79). First observe that when $|S| = K$, it holds for any $S' \in \mathcal{S}, S' \neq S, S' \neq \emptyset$, that $|S \backslash S'| \geq 1$, and $|S^c \cap S'^c| \geq M - K - |S|$. It then follows that

$$\begin{aligned}
\overline{\alpha}_S = \min_{\substack{S' \neq S \\ S' \neq \emptyset}} \Bigg[& \left|S \cap S'\right| D\left(\mu \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S \backslash S'\right| D\left(\mu \middle\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right) \\
& + \left|S' \backslash S\right| D\left(\pi \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S^c \cap S'^c\right| D\left(\pi \middle\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right) \Bigg] \qquad (4.95) \\
\geq\ & D\left(\mu \middle\| \tfrac{|S \backslash S'|\mu + |S'^c \cap S^c|\pi}{M - |S'|}\right) + (M - K - |S|)\, D\left(\pi \middle\| \tfrac{|S \backslash S'|\mu + |S'^c \cap S^c|\pi}{M - |S'|}\right) \\
\geq\ & \min_{p \in \mathcal{P}(\mathcal{Y})} D(\mu \| p) + (M - K - |S|)\, D(\pi \| p) \\
=\ & D\left(\mu \middle\| \frac{\mu + (M - K - |S|)\,\pi}{M - K - |S| + 1}\right) + (M - K - |S|)\, D\left(\pi \middle\| \frac{\mu + (M - K - |S|)\,\pi}{M - K - |S| + 1}\right) \\
=\ & \overline{\eta}_S(\mu \| \pi), \qquad (4.96)
\end{aligned}$$

where (4.96) follows from (2.7) with $\mathcal{C}$ therein comprising $M - K - |S|$ copies of π and one μ.

We continue for the case with $1 \leq |S| < K$. For any $S' \in \mathcal{S}, S' \neq \emptyset$, such that $S \backslash S' \neq \emptyset$, the term inside the minimum on the right side of (4.95) is still lower bounded by $\overline{\eta}(\mu \| \pi)$. Now for any other $S' \neq S$, such that $S' \supset S$, we obtain that

$$\begin{aligned}
& \Bigg[\left|S \cap S'\right| D\left(\mu \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S \backslash S'\right| D\left(\mu \middle\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right) \\
& \quad + \left|S' \backslash S\right| D\left(\pi \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + \left|S^c \cap S'^c\right| D\left(\pi \middle\| \tfrac{|S \backslash S'|\mu + |S^c \cap S'^c|\pi}{M - |S'|}\right) \Bigg] \\
& \quad \geq\ |S| D\left(\mu \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) + D\left(\pi \middle\| \tfrac{|S \cap S'|\mu + |S' \backslash S|\pi}{|S'|}\right) \\
& \quad \geq\ \min_{p \in \mathcal{P}(\mathcal{Y})} |S| D(\mu \| p) + D(\pi \| p) \\
& \quad = |S| D\left(\mu \middle\| \frac{|S|\mu + \pi}{|S| + 1}\right) + D\left(\pi \middle\| \frac{|S|\mu + \pi}{|S| + 1}\right) \ =\ \eta_S(\mu \| \pi). \qquad (4.97)
\end{aligned}$$

We conclude from (4.96) and (4.97) that (4.79) holds. □

Remark 4. *It follows from (4.81) and (2.8) that as $M \to \infty$ (while K is kept fixed),*

$$\overline{\eta}_S(\mu, \pi) \;\to\; D(\mu \| \pi), \tag{4.98}$$

i.e., the asymptotic performance for the test in (4.50), (4.51) and (4.54) when neither μ nor π is known (cf. (4.79)) converges to that for the test in (4.50), (4.51) and (4.52) when π is known (cf. (4.57)) as $M \to \infty$.

Remark 5. *Similar to Remark 3, the results in Theorems 17 and 18 continue to hold when the deterministic time horizon $T \log T$ in (4.50) is replaced with a more general form $f(T)$ as long as $f(T)$ increases at least as fast as $T \log T$, and a constant offset $\log C$ is added to the time-dependent thresholds in (4.52) and (4.54) on the right side of the inequalities.*

4.3 Multiple Distinctly Distributed Outliers

We now extend the results to the setting with multiple distinctly distributed outliers. In the fixed sample size setting, we prove in Chapter 3 that when the outliers can be arbitrarily distinctly distributed, the assumption of the number of outliers being known is essential for the existence of a universally exponentially consistent test. Therefore, for the sequential setting, we assume that the number of outliers is known when they can be distinctly distributed outliers. Since the proofs of the results in this section are similar to those for the results in the previous sections, we present the proposed universal test and the results pertaining to its asymptotic performance without proofs.

Let $S \subset \{1, \ldots, M\}$, $|S| = K$, denote the set of K outliers. Each of the i-th outlier, $i \in S$, is distributed according to μ_i, which can be arbitrarily distinct from one another as long as each $\mu_i \neq \pi$. Conditioned on S being the set of outliers, the joint distribution of the first n observations under the hypothesis with the outlier subset being S is given in (3.54).

The test for the outliers is done based on a rule $\delta\left(\boldsymbol{Y}^N\right) \in \mathcal{S}_K$, for an appropriate stopping time N and where $\mathcal{S}_K$ will now denote the set of all subsets of $\{1, \ldots, M\}$ of size *exactly* K. Notice that unlike in the previous sections, the current model does not include the null hypothesis with no outlier. The maximal error probability is defined as previously in (4.47) but with the maximum being over $\mathcal{S}_K$ instead.

As previously, for the setting with both the typical and outlier distributions being known and with $\hat{S}(\boldsymbol{y}^n) \triangleq \underset{S \in \mathcal{S}_K}{\operatorname{argmax}}\; p_S(\boldsymbol{y}^n)$, the MSPRT has stopping and final decision rules as in (4.48) and (4.49), but with the joint distribution $p_S(\boldsymbol{y}^n)$ instead as in (3.54), and with the maximum in the denominator in (4.48) being over $\mathcal{S}_K \backslash \left\{\hat{S}\right\}$ instead. This MSPRT is

asymptotically optimal (cf.[26, 27]).

Proposition 19. *As the threshold T in (4.48) approaches infinity, the MSPRT in (4.48) and (4.49), with $p_S(\boldsymbol{y}^n)$ as in (3.54), $\hat{S}$ being computed over $\mathcal{S}_K$, and the maximum in the denominator in (4.48) being over $\mathcal{S}_K \backslash \left\{\hat{S}\right\}$) satisfies*

$$P_{\max} \;\leq\; O\left(\frac{1}{T}\right).$$

In addition, for each $S \in \mathcal{S}_K$, as $T \to \infty$,

$$\mathbb{E}_S[N^*] \;=\; \frac{\log T(1+o(1))}{\left(\min\limits_{i\in S} D(\mu_i\|\pi)\right)+\left(\min\limits_{j\notin S} D(\pi\|\mu_j)\right)} \;=\; \frac{-\log P_{\max}(1+o(1))}{\left(\min\limits_{i\in S} D(\mu_i\|\pi)\right)+\left(\min\limits_{j\notin S} D(\pi\|\mu_j)\right)}. \tag{4.99}$$

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests (N,δ) with vanishing maximal error probability, for each $S \in \mathcal{S}_K$,

$$\mathbb{E}_S[N] \;\geq\; \frac{-\log P_{\max}}{\left(\min\limits_{i\in S} D(\mu_i\|\pi)\right)+\left(\min\limits_{j\notin S} D(\pi\|\mu_j)\right)}(1+o(1)).$$

4.3.1 Proposed Universal Test

When only π is known, we can compute the corresponding GL of $\boldsymbol{y}^n$ under each hypothesis $S \in \mathcal{S}_K$ by replacing the unknown $\mu_i,\ i \in S$, in (3.54) with its ML estimate $\hat{\mu}_i \triangleq \gamma_i, i \in S$. In particular, with $\hat{S}(\boldsymbol{y}^n) = \underset{S\in\mathcal{S}_K}{\operatorname{argmin}} \sum\limits_{j\notin S} D(\gamma_j\|\pi)$ denoting the instantaneous estimate of the hypothesis (using the GL) at time n, the proposed universal test can be described by the following stopping and final decision rules:

$$N^* \;=\; \underset{n\geq 1}{\operatorname{argmin}} \left[\min_{\substack{S'\neq\hat{S}\\ S'\in\mathcal{S}_K}} n\left[\sum_{j\notin S'} D\left(\gamma_j\middle\|\pi\right) - \sum_{j\notin\hat{S}} D\left(\gamma_j\middle\|\pi\right)\right] > \log T + (M+1)|\mathcal{Y}|\log(n+1)\right]; \tag{4.100}$$

$$\delta^* \;=\; \hat{S}\left(\boldsymbol{Y}^{N^*}\right). \tag{4.101}$$

Similarly, when neither μ nor π is known, the test can be written as

$$N^* \;=\; \operatorname*{argmin}_{n\geq 1} \Bigg[\min_{\substack{S'\neq\hat{S}\\ S'\in\mathcal{S}_K}} n\Big[\sum_{j\notin S'} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin S'}\gamma_k}{M-|S'|}\right) - \sum_{j\notin \hat{S}} D\left(\gamma_j \Big\| \tfrac{\sum_{k\notin \hat{S}}\gamma_k}{M-|\hat{S}|}\right) \Big] \\ > \log T + (M+1)|\mathcal{Y}|\log(n+1)\Bigg]; \tag{4.102}$$

$$\delta^* \;=\; \hat{S}\left(\boldsymbol{Y}^{N^*}\right), \tag{4.103}$$

but with $\hat{S}\left(\boldsymbol{y}^n\right) = \operatorname*{argmin}_{S\in\mathcal{S}} \sum_{j\notin S} D\Big(\gamma_j \| \tfrac{\sum_{k\notin S}\gamma_k}{M-|S|}\Big)$. Note that since the null hypothesis is not present in this case, there is no need to truncate the stopping time by a predefined horizon as in (4.50).

4.3.2 Performance of Proposed Test

Using techniques as in the proofs of the results in the the previous sections, it is easy to verify that the proposed test achieves the following performance.

Theorem 20. *With the number of outliers K being known and when only π is known, the test in (4.100) and (4.101) is universally exponentially consistent, and yields for every T that*

$$P_{\max} \;\leq\; O\left(\frac{1}{T}\right).$$

In addition, for each non-null hypothesis $S\in\mathcal{S}_K$ as $T\to\infty$,

$$\mathbb{E}_S\left[N^*\right] \;=\; \frac{\log T}{\min\limits_{i\in S} D\left(\mu_i\|\pi\right)}(1+o(1)) \;\leq\; \frac{-\log P_{\max}}{\min\limits_{i\in S} D\left(\mu_i\|\pi\right)}(1+o(1)). \tag{4.104}$$

Theorem 21. *With the number of outliers K being known, but neither μ nor π being known, the test in (4.102) and (4.103) is universally exponentially consistent, and yields for every T that*

$$P_{\max} \;\leq\; O\left(\frac{1}{T}\right).$$

In addition, for each non-null hypothesis $S \in \mathcal{S}_K$ as $T \to \infty$,

$$\mathbb{E}_S[N^*] \leq \frac{-\log P_{\max}}{\min_{i \in S}\left(D\left(\mu_i \| \frac{\mu_i + (M-2K)\pi}{M-2K+1}\right) + (M-2K)D\left(\pi \| \frac{\mu_i + (M-2K)\pi}{M-2K+1}\right)\right)}(1+o(1)). \quad (4.105)$$

Remark 6. *As $M \to \infty$, the denominator of (4.105) converges to $\min_{i \in S} D(\mu_i \| \pi)$, which is the asymptotic performance of the universal test in (4.100) and (4.101) when π is known (cf. (4.104)). It is also clear from (3.8), (4.104) and (4.105) that our proposed test in (4.102) and (4.103) is asymptotically exponentially consistent, i.e., as $M \to \infty$, its limiting error exponent is positive whenever that of the MSPRT is.*

4.4 Numerical Results

We now provide some numerical results for an example with $|\mathcal{Y}| = 4$. We compare the performance of the sequential test with that of the fixed sample size (FSS) test studied in Chapter 3. In this example, we assume that there are at most two outliers among five sequences with the pair of outlier and typical distributions being $\mu = (0.4, 0.05, 0.5, 0.05)$ and $\pi = (0.07, 0.42, 0.1, 0.41)$. The plots in Figure 4.1 are for the case where the underlying hypothesis has one outlier, and those in Figure 4.2 for two outliers. Depending on the nature of the test, the horizontal axis corresponds to the average stopping time for the sequential testing, and the length of each sequence for the FSS test. In both figures, the vertical axis corresponds to the *conditional* error probabilities incurred by each test conditioned on the underlying hypothesis. It follows from the result in Chapter 3 that there cannot exist a universally exponentially consistent FSS test with respect to $P_{\max}$ primarily because the conditional error probability under the null hypothesis is the bottleneck. Hence, we consider only the two conditional error probabilities under hypotheses with one outlier and two outliers, respectively, with respect to which the FSS test is universally exponentially consistent. When comparing the FSS test to our sequential test, it is natural to compare the fixed sample size of the FSS test to the expected stopping time under a hypothesis for which the conditional error probability is considered. It should be noted that although our result concerning the achievable asymptote of the expected stopping time of the sequential test in (4.79) is with respect to the maximum error probability, $P_{\max}$, the same asymptote is still achievable when $P_{\max}$ is replaced by a conditional error probability. For the sequential test, the thresholds are chosen to be $T = \{1.3, 1.35, 1.4, \ldots, 2.55\}$, and the corresponding deterministic time horizon $f(T) = \{170, 175, 180, \ldots, 300\}$. The constant offset in (4.46) is set to be $C = 15.05$. For the FSS test, the lengths of each of the sequences are chosen such

that they are within the same range as the average stopping times of the sequential test.

As shown in both figures, the sequential test starts to outperform the FSS test when the average stopping time is sufficiently large. Replacing $P_{\max}$ with the corresponding conditional error probability, the result in (4.79) suggests that to achieve the same level of conditional error probability, the expected stopping time under a hypothesis with two outliers should be less than that under a hypothesis with one outlier. The simulation results in Figures 4.1 and 4.2 corroborate such theoretical findings. It is also interesting to note that in both figures, there is a drastic drop in the conditional error probability incurred by the sequential test when the average stopping time exceeds a certain value. The same phenomenon is not observed in the simulation results of the FSS test. The drop in the conditional error probability can be explained by that fact that the sequential test is more adaptive compared with the FSS test. The parameters of the sequential test, i.e., the threshold T, the corresponding deterministic time horizon $f(T)$, and the constant offset C, are chosen independent of the true hypothesis and the underlying distributions. As a result, these parameters may not be optimal for the true hypothesis, and the distributions associated with it. Despite the arbitrary choice of the test parameters, as the average stopping time increases with T, the sequential test quickly adapts to the true hypothesis and the unknown distributions, and yields a drastic improvement in its performance.

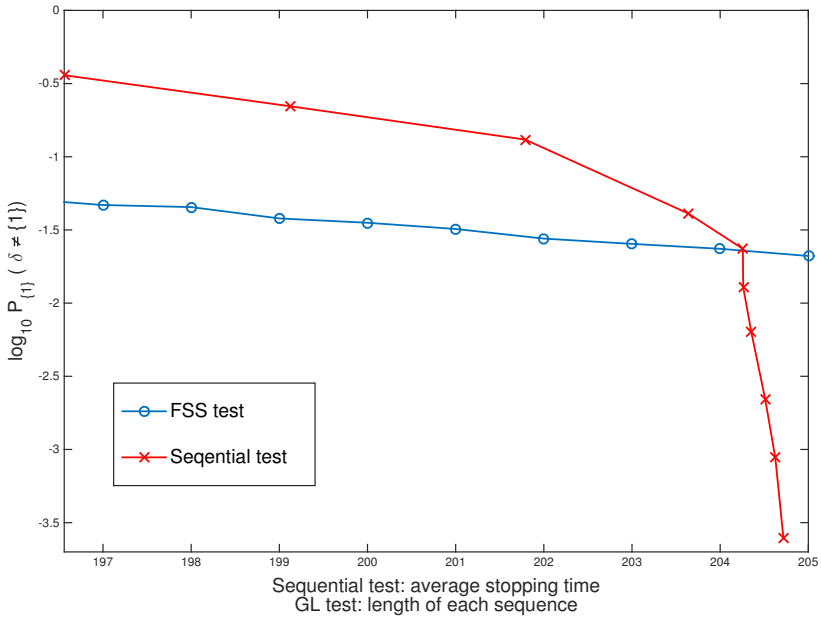

Figure 4.1: Performance of proposed FSS and sequential tests, with one outlier

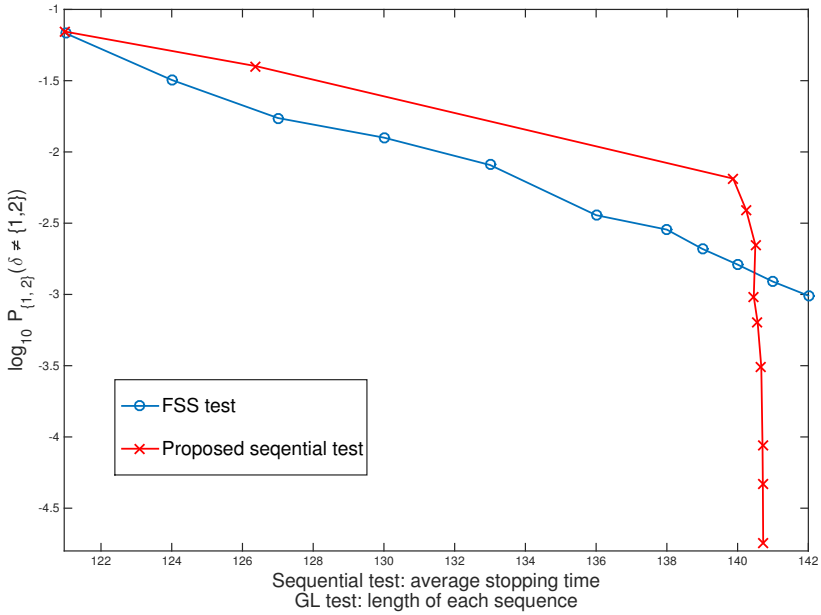

Figure 4.2: Performance of proposed FSS and sequential tests, with two outliers

4.5 Discussion

In practice, it would be of interest to determine how to set the value of the threshold T of the universal test to satisfy a predefined level of error probability. By carefully inspecting (4.24), (4.27) and (4.28) in the proof of Theorem 14, we see that although arbitrarily small probabilities of $\mathbb{P}_i\left\{\delta^* \neq i, \delta^* \neq 0\right\}$, $i = 1, \ldots, M$, $\mathbb{P}_0\left\{\delta^* \neq 0\right\}$ can be achieved with a suitably large T universally for every μ, the same cannot be achieved universally for the probability $\mathbb{P}_i\left\{\delta^* = 0\right\}$, $i = 1, \ldots, M$, unless we are given a lower bound on the distance between μ and π. This complication arises because of the nature of the universal setting under consideration, and is not a drawback of our test. Specifically, given any test, there will always be μ, π sufficiently close to each other, that will incur large error probability. To put it differently, in the considered universal setting, we need to be content with a sequence of tests indexed, say, by T rather than a single test, that will guarantee a certain level of error probability for a sufficiently large T. Note that throughout we have assumed that we are working with only one set of (test) data. Additional training data for the typical and outlier distributions, when available separately from the test data, could be used to facilitate setting an appropriate threshold value to be used for the test data.

CHAPTER 5

EXTENSION TO CONTINUOUS ALPHABETS

The theoretic results in Chapters 3 and 4 only hold when the underlying alphabet is finite. In this chapter, we generalize our results to the fixed sample size setting with continuous alphabets.

In a recent work, Zou et al. [45] proposed a kernel-based test for universal outlier hypothesis testing in the fixed sample size setting. Such a test is based on the mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [46], and it is applicable when the underlying distributions are continuous. The test is based on estimates of the *maximum mean discrepancy* (MMD) between the distributions underlying the observation sequences. The MMD test has the same structure as that of the fixed sample size GL tests in Chapter 3. Specifically, when there is exactly one outlier, the MMD test selects the sequence such that the estimate of the MMD between the underlying distribution of the selected sequence, and that of (possibly a mixture distribution) the other sequences, is maximized. The MMD test is appropriate for the universal setting because the MMD between two distributions can be estimated using the observations in a completely non-parametric manner. It is shown in [45] that the MMD test is universally consistent, and sometimes universally exponentially consistent for various models. However, it is not known whether the MMD test is optimal asymptotically as the number of sequences goes to infinity, and it is not clear how to generalize the MMD test to the sequential setting.

We now propose a test for continuous alphabets that is similar to the GL test for finite alphabets. The proposed test is based on a non-parametric estimate of the Kullback-Leibler (KL) divergence, and it is applicable to the universal setting. We show that such a test is universally consistent for outlier hypothesis testing. We also provide numerical results that compare the proposed test and the MMD test.

5.1 Divergence Estimator for Continuous Probability Measures

Let P and Q be continuous probability measures on a measurable space $(\Omega, \mathcal{F})$. We say that P is *absolutely continuous* with respect to Q, denoted as $P \ll Q$, if for any set $A \in \mathcal{F}$ such

that $P(A) = 0$, it also holds that $Q(A) = 0$. The KL divergence between P and Q is defined as

$$D(P\|Q) \triangleq \int_{\Omega} \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}$$

when $P \ll Q$, and $+\infty$ otherwise. For simplicity of the exposition, we assume that P and Q are absolutely continuous (w.r.t the Lebesgue measure) probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ that have $D(P\|Q) < \infty$.

If $P \ll Q$, the Radon-Nykodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ can be approximated by $\frac{\triangle P}{\triangle Q}$ as $\triangle Q$ diminishes, where $\triangle P$ (or $\triangle Q$) denotes the measure of a small segment in $\mathcal{B}_{\mathbb{R}}$. Next we discuss two estimators for the KL divergence of continuous probability measures, which all have as an intermediate step an estimate of the Radon-Nykodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$.

5.1.1 Naive Plug-in Estimator

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be i.i.d. observations drawn from P and Q, respectively. Let $\{I_i\}_{i=1}^T, I_i \subset \mathbb{R}$ be a collection of intervals, the union of which constitutes the whole real line, i.e., $\cup_{i=1,\ldots,T} I_i = \mathbb{R}$. Note that the location of the intervals in $\{I_i\}_{i=1}^T$ is *independent* of the observations $\mathbf{X}$ and $\mathbf{Y}$. Let k_i and l_i be the number of observations in $\boldsymbol{x}$ and $\boldsymbol{y}$ that take values in the interval $I_i, i = 1, \ldots, T$, respectively. For a fixed partition $\{I_i\}_{i=1}^T$, the corresponding empirical probability measure induced by the observations $\mathbf{X}$ and $\mathbf{Y}$ are

$$P_m(I_i) = \frac{k_i}{m}, \quad i = 1, \ldots, T, \tag{5.1}$$

and

$$Q_n(I_i) = \frac{l_i}{n}, \quad i = 1, \ldots, T, \tag{5.2}$$

respectively. The naive plug-in estimator for $D(P\|Q)$ is obtained by simply plugging P_m and Q_n into the expression for the KL divergence, i.e.,

$$\begin{aligned} \hat{D}_{\text{plug-in}}(\boldsymbol{x}\|\boldsymbol{y}) &= \sum_{i=1}^{T} P_m(I_i) \log \frac{P_m(I_i)}{Q_n(I_i)} \\ &= \sum_{i=1}^{T} \frac{k_i}{m} \log \frac{k_i/m}{l_i/n}. \end{aligned} \tag{5.3}$$

The term "naive" is used to contrast the above estimator with the following estimator where the partition is a function of the observations in $\mathbf{Y}$.

5.1.2 Estimator Based on Data-Dependent Partition

A drawback of the naive plug-in estimator is that $\hat{D}_{\text{plug-in}}(\mathbf{X}\|\mathbf{Y})$ may be infinite even though it holds that $P \ll Q$. We now introduce a estimator using a *data-dependent* partition that resolves this issue [47].

We see from an alternative definition of $D(P\|Q)$

$$D(P\|Q) = \int_{\mathbb{R}} \mathrm{d}Q \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q} \tag{5.4}$$

that the density of P can be estimated with respect to Q by $\frac{\triangle P}{\triangle Q}$, which is finite as long as $P \ll Q$. Then the resulting estimate of $D(P\|Q)$ is also guaranteed to be finite when $P \ll Q$.

Denote the *order statistics* of $\mathbf{Y}$ by $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$, which satisfies that $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$. We now partition the real line into empirically equiprobable intervals (except for possibly the last one) with respect to $\mathbf{Y}$. In particular, let

$$\{I_i^n\}_{i=1}^{T_n} = \left\{(-\infty, Y_{l_n}],\ (Y_{l_n}, Y_{2l_n}],\ \ldots,\ (Y_{l_n(T_n-1)}, \infty)\right\}, \tag{5.5}$$

where $l_n \in \mathbb{N} \leq n$ is the number of points in each intervals (except for possibly the last one), and $T_n = \lfloor n/l_n \rfloor$ is the number of such intervals. Let k_i denote the number of observations from P that belong to the interval I_i^n, $i = 1, \ldots, T_n$. We approximate $\frac{\mathrm{d}P}{\mathrm{d}Q}$ in each segment $I_i^n, i = 1, \ldots, T_n - 1$, by $\frac{k_i/m}{l_n/n}$, and in $I_{T_n}^n$ by $\frac{k_{T_n}/m}{l_n/n+\delta_n}$. Then the KL divergence between P and Q can be estimated as

$$\hat{D}_{m,n}(\boldsymbol{x}\|\boldsymbol{y}) = \sum_{i=1}^{T_n-1} \frac{k_i}{m} \log \frac{k_i/m}{l_n/n} + \frac{k_{T_n}/m}{l_n/n + \delta_n}, \tag{5.6}$$

where $\delta_n = (n - l_n T_n)$ is the correction term for the last segment $I_{T_n}^n$. In contrast to the naive plug-in estimator in (5.3), the density of P is now estimated with respect to Q (cf. (5.4)), which is guaranteed to be finite as long as $P \ll Q$.

It can be shown (cf. Theorem 1 in [47]) that the divergence estimator in (5.6) is strongly consistent.

Proposition 20. *Let P and Q be absolutely continuous probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Assume that the divergence between P and Q is finite. Let $\mathbf{X} = \{X_1, \ldots, X_m\}$ and*

$\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ *be i.i.d. observations drawn from* P *and* Q*, respectively. Let* l_n *and* T_n *be defined as in (5.5). If* $l_n, T_n \to \infty$ *as* $n \to \infty$*, then the estimator in (5.6) is strongly consistent, i.e., it holds that*

$$\hat{D}_{m,n}(\mathbf{X}\|\mathbf{Y}) \to D(P\|Q) \quad a.s. \tag{5.7}$$

as $m, n \to \infty$*.*

Remark 7. *It is suggested by the numerical results (cf. Section V in [47]) that the estimator based on the data-dependent partition outperforms the plug-in estimator in the case of small sample sizes, i.e., the former converges much faster to the true KL divergence as the sample size increases.*

5.2 Proposed Universal Test for Continuous Alphabets

Recall that for a finite alphabet, when there is exactly one outlier (cf. Chapter 3), the GL test can be equivalently written as (cf. (3.10))

$$\delta\big(y^{Mn}\big) = \underset{i=1,\ldots,M}{\operatorname{argmin}} \sum_{j\neq i} D\big(\gamma_j \big\| \tfrac{\sum_{k\neq i}\gamma_k}{M-1}\big). \tag{5.8}$$

The GL test can be interpreted alternatively as follows. It starts by estimating the distribution underlying each individual sequence. For each hypothesis, it then computes an estimate of the KL divergence between each typical sequence and that of the collection of all typical sequences. In particular, conditioned on sequence i being the outlier, $D\big(\gamma_j \big\| \frac{\sum_{k\neq i}\gamma_k}{M-1}\big)$ is indeed the naive plug-in estimate (for a finite alphabet) of the KL divergence between sequence $j, j \neq i$, and the collection of all typical sequences. Then the GL test decides on the hypothesis such that the sum of all such estimates is minimized. In other words, the GL test selects the hypothesis under which there is the least amount of "divergence" among all typical sequences.

A straightforward generalization of the GL test to settings with continuous alphabets is to replace the plug-in estimator for finite alphabets in (5.8) with an estimator appropriate for continuous alphabets, i.e., the estimator in (5.3) or (5.6).

Recall that $\boldsymbol{y}^{(i)} = \left(y_1^{(i)}, \ldots, y_n^{(i)}\right)$ denotes the i-th sequence. Let $y^{Mn} \setminus \boldsymbol{y}^{(i)}$ denote the collection of all sequences except for the i-th one, $i = 1, \ldots, M$. When there is exactly one

outlier, using the estimator based on data-dependent partition, the proposed test is

$$\delta\big(y^{Mn}\big) = \underset{i=1,\ldots,M}{\operatorname{argmin}} \sum_{j\neq i} \hat{D}_{n,(M-1)n}\big(\boldsymbol{y}^{(j)} \,\big\|\, y^{Mn} \setminus \boldsymbol{y}^{(i)}\big), \tag{5.9}$$

where $\hat{D}_{n,(M-1)n}\big(\boldsymbol{y}^{(j)} \,\big\|\, y^{Mn} \setminus \boldsymbol{y}^{(i)}\big)$ is defined in (5.6). For models with at most one outlier, the proposed test is (cf. (3.42))

$$\delta(y^{Mn}) = \begin{cases} \underset{i=1,\ldots,M}{\arg\min} \sum\limits_{k\neq i} \hat{D}\left(\boldsymbol{y}^{(k)} \,\middle\|\, y^{Mn} \setminus \boldsymbol{y}^{(l)}\right), & \text{if} \quad \max\limits_{j\neq j'} \Big[\sum\limits_{k\neq j} \hat{D}\left(\boldsymbol{y}^{(k)} \middle\| y^{Mn} \setminus \boldsymbol{y}^{(l)}\right) \\ & \quad - \sum\limits_{k\neq j'} \hat{D}\left(\boldsymbol{y}^{(k)} \middle\| y^{Mn} \setminus \boldsymbol{y}^{(l)}\right) \Big] > \lambda_n, \\ 0, & \text{otherwise}, \end{cases} \tag{5.10}$$

where $\lambda_n = \Theta(\frac{\log n}{n})$, and the ties in the first case of (5.10) are broken arbitrarily. Similarly, for models with multiple outliers, the proposed tests are obtained by replacing the plug-in estimates of the KL divergence in (3.71) and (3.95) with those based on data-dependent partition, respectively.

Remark 8. *An alternative test can be constructed using the naive plug-in estimator in (5.3) in place of the one based on a data-dependent partition. As we shall see in the simulation result, the test using a data-dependent partition outperforms the one using the naive plug-in estimator by a large margin on a synthetic data set for outlier hypothesis testing.*

5.3 Performance of Proposed Test

The following theorem establishes that our proposed test in (5.9) is universally consistent for outlier hypothesis testing when there is exactly one outlier.

Theorem 22. *Let both μ and π be absolutely continuous probability measures defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. When there is exactly one outlier, the proposed test in (5.9) is universally consistent for all $\mu, \pi, \mu \neq \pi$.*

Proof. For $i = 1, \ldots, M$, denote the test statistics in (5.9) as

$$U_i \triangleq \sum_{j\neq i} \hat{D}_{n,(M-1)n}\left(\mathbf{Y}^{(j)} \| \, Y^{Mn} \setminus \mathbf{Y}^{(i)}\right). \tag{5.11}$$

By the symmetry of the problem, it is clear that $\mathbb{P}_i\{\delta \neq i\}$ is the same for every $i = 1, \ldots, M$;

hence

$$\max_{i=1,\ldots,M} \mathbb{P}_i\{\delta \neq i\} = \mathbb{P}_1\{\delta \neq 1\}. \tag{5.12}$$

It now follows from

$$\mathbb{P}_1\{\delta \neq 1\} = \mathbb{P}_1\{\cup_{j\neq 1} U_1 \geq U_j\}$$

that

$$\mathbb{P}_1\{U_1 \geq U_2\} \leq \mathbb{P}_1\{\delta \neq 1\} \leq \sum_{j=2}^{M} \mathbb{P}_1\{U_1 \geq U_j\}. \tag{5.13}$$

Next, we get from the definition of the test in (5.9) that

$$\mathbb{P}_1\left\{U_1 \geq U_2\right\} = \mathbb{P}_1\left\{ \sum_{j\neq 1} \hat{D}_{n,(M-1)n}\left(\mathbf{Y}^{(j)} \parallel Y^{Mn} \setminus \mathbf{Y}^{(1)}\right) \geq \sum_{j\neq 2} \hat{D}_{n,(M-1)n}\left(\mathbf{Y}^{(j)} \parallel Y^{Mn} \setminus \mathbf{Y}^{(2)}\right) \right\}.$$

Conditioned on H_1 being the true hypothesis, the first sequence is distributed according to μ, and all other sequences are distributed according to π. It simply follows from (5.11), and the strong consistency of the estimator in (5.6) that under H_1,

$$U_1 \to 0 \quad \text{a.s.},$$

and

$$U_2 \to (M-2)D\left(\pi \,\middle\|\, \frac{\mu + (M-2)\pi}{M-1}\right) + D\left(\mu \,\middle\|\, \frac{\mu + (M-2)\pi}{M-1}\right) \quad \text{a.s.}$$

as $n \to \infty$. We then obtain that

$$\mathbb{P}_1\left\{U_1 \geq \epsilon\right\} \to 0, \tag{5.14}$$

and

$$\mathbb{P}_1\Bigg\{ \Bigg| U_2 - (M-2)D\left(\pi \,\Big\|\, \frac{\mu+(M-2)\pi}{M-1}\right) - D\left(\mu \,\Big\|\, \frac{\mu+(M-2)\pi}{M-1}\right) \Bigg| \geq \epsilon \Bigg\} \to 0 \tag{5.15}$$

for any $\epsilon > 0$ as $n \to \infty$. Now by choosing $\epsilon > 0$ to be sufficiently small, i.e., $0 < \epsilon < \frac{1}{2}\left[(M-2)D(\pi\| \frac{\mu+(M-2)\pi}{M-1}) + D(\mu \,\|\, \frac{\mu+(M-2)\pi}{M-1})\right]$, it holds by (5.14) and (5.15) that

$$\mathbb{P}_1\{U_1 \geq U_2\} \to 0 \tag{5.16}$$

as $n \to \infty$. Since M is fixed and finite, the universal consistency of the test in (5.9) follows from (5.12), (5.13) and (5.16). □

Remark 9. *Using the same arguments as in the proof of Theorem 22, we can show that the proposed test is universally consistent for models with at most one outlier, and for the models in Theorem 10 and Theorem 12 with multiple outliers.*

5.4 Numerical Results

We now compare the performance of the MMD based test in [45], the test based on data-dependent partition in (5.9), and the test using the naive plug-in estimator in (5.3), on a synthetic data set with a continuous alphabet. The number of quantization intervals is chosen to be $\sqrt{n}$ for the test based on data-dependent partition, and the test using the naive plug-in estimator. In this example, we consider a fixed sample size setting with exactly one outlier among $M = 5$ sequences. The outlier and typical observations are Gaussian random variables with different parameters. Specifically, we have $\mu = \mathcal{N}(0, 2)$ as the outlier distribution, and $\pi = \mathcal{N}(1, 2)$ as the typical distribution.

In Figure 5.1, the horizontal axis corresponds to the length of each sequence, and the vertical axis corresponds to the maximum error probability incurred by each test. As shown in Figure 5.1, the MMD test yields the best performance among all three tests. It is proved in [45] that the MMD test is universally exponentially consistent for models with exactly one outlier. Our simulation results corroborate such theoretical findings. Although not proved in this dissertation, the test based on data-dependent partition, and the one using naive plug-in estimator both seem to be exponentially consistent as suggested by the numerical results. And the test based on data-dependent partition outperforms the one using naive plug-in estimator by a large margin.

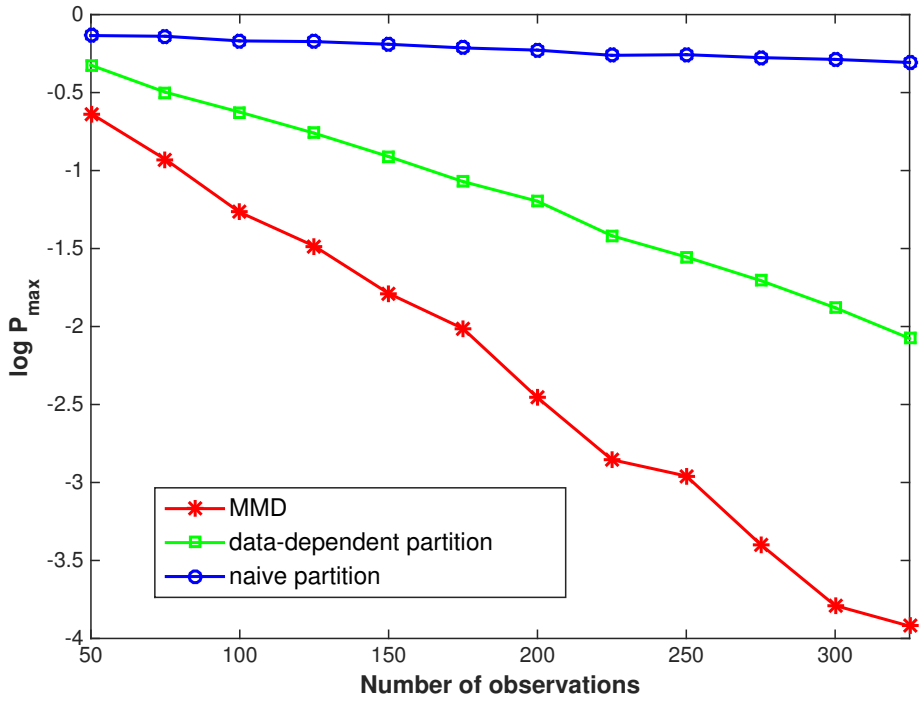

Figure 5.1: Comparison between different tests for continuous alphabets

CHAPTER 6

CONNECTION TO CLUSTER ANALYSIS

The goal of cluster analysis is to segment a collection of data objects into homogeneous subsets or "clusters", such that objects assigned to the same cluster are more closely related to one another than objects assigned to different clusters [12–15]. Cluster analysis is also concerned with exploring the data objects to determine if they can be meaningfully represented by a relatively small number of groups. Similar to classification, cluster analysis creates labeling of the objects with class (cluster) labels. The labels are derived from the data in cluster analysis, whereas for classification, unlabeled objects are assigned a class label using a model developed from training objects with known labels.

Of central importance to a majority of clustering algorithms is the notion of proximity, or sometimes referred to as similarity or dissimilarity, which is a quantitative measurement that characterizes how "close" two objects are [12–15]. In cluster analysis, an object is usually described by a set of measurements. The similarity between a pair of objects is often given by an appropriately chosen distance metric, which can be computed using such measurements. For instance, a popular choice of the distance metric is the Euclidean distance for continuous measurements, and Jaccard coefficient for binary measurements [48, 49]. The similarities (dissimilarities) between pairs of objects are summarized in a similarity (dissimilarity) matrix, where the (i, j)-th entry of the matrix is the similarity between the i-th and the j-th objects.

Generally speaking, the objective of a clustering algorithm is to minimize the *heterogeneity* within clusters while maximizing the *separation* between clusters [12–15]. The greater the similarity within clusters and the greater the differences between clusters, the better or more distinct the clustering assignment is. The heterogeneity of a particular cluster can be defined as the sum over all the dissimilarities between pairs of objects within such a cluster. The separation between clusters can be defined in a similar manner, i.e., it is the sum over all the dissimilarities between pairs of objects belonging different clusters. Having chosen appropriate measures of heterogeneity and separation, a clustering algorithm seeks to either minimize the heterogeneity within clusters, or to maximize the separation between clusters, or a combination of both.

6.1 Cluster Analysis Techniques

Some popular categories of clustering methods are summarized as follows. Parametric clustering methods such as the EM algorithm estimate a mixture density from the collection of observations. And observations generated by the same mixture component are assigned to the same cluster. In prototype-based methods such as K-means [13, 16, 17], a cluster is represented by its corresponding prototype, which is often a centroid, i.e., the average of all the data objects in the cluster. In situations where a centroid is not meaningful, e.g., when the data has categorical values, a popular choice of cluster prototype is a medoid [18]. Density-based methods define a cluster as a dense region of objects, which is surrounded by a region of low density [21, 22]. Density-based methods are most appropriate when noise is present in the observations. In graph-based cluster analysis, a graph is constructed where each node in the graph represents a data object, and edges are assigned using the similarity matrix of these data objects [23]. For instance, in contiguity-based graphs, an edge is assigned to a pair of nodes if the similarity between them is greater than a threshold [15]. No single definition of a cluster in graphs is universally accepted [23]. For example, in the loosest sense, a graph cluster can be defined as a connected component [23], and strictest definition is that each cluster should be a maximal clique [50].

In the following sections, we discuss two popular clustering algorithms that are relevant to this chapter. A widely used algorithm is the K-means algorithm [13, 16, 17] mentioned previously. Another modern clustering method is the so-called spectral clustering, which is a graph-based technique. Spectral clustering algorithms partition the data space by performing cluster analysis over the spectrum (eigenvectors) of the similarity matrix [51–53].

6.1.1 K-Means Clustering

Let $x_i, i = 1, \ldots, n$, be vector observations that take values in $\mathbb{R}^p$. Let the dissimilarity between a pair of observations x_i and $x_{i'}$ be given by the Euclidean distance

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2.$$

The number of clusters, denoted by K, is predetermined and satisfies $K < n$. Each cluster is uniquely indexed by an integer $k \in \{1, \ldots, K\}$. Let C be a many-to-one mapping where for each observation x_i, $k = C(i)$ is the cluster membership of the i-th observation. The

total heterogeneity of a particular cluster assignment C is

$$\begin{aligned} H(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}\|x_i - x_{i'}\|^2 \\ &= \sum_{k=1}^{K} n_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2, \end{aligned} \tag{6.1}$$

where $n_k \triangleq \sum_{i=1}^{n} I(C(i)=k)$ is the number of observations belonging to the k-th cluster, and $\bar{x}_k \triangleq \frac{1}{n_k}\sum_{C(i)=k} x_i$ is the center of the k-th cluster.

The optimal cluster assignment $C^\star$, which minimizes the total heterogeneity defined in (6.1), solves the following optimization problem

$$C^\star = \arg\min_{C} \sum_{k=1}^{K} n_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2. \tag{6.2}$$

Note that given the members associated with the k-th cluster, $k = 1, \ldots, K$, it also holds that

$$\bar{x}_k = \arg\min_{m} \sum_{C(i)=k}\|x_i - m\|^2.$$

As a result, the optimization problem in (6.2) can be equivalently written as

$$C^\star = \arg\min_{C, \{m_k\}_{k=1}^{K}} \sum_{k=1}^{K} n_k \sum_{C(i)=k}\|x_i - m_k\|^2. \tag{6.3}$$

The K-means algorithm is an iterative descent algorithm to solve for the cluster assignment $C^\star$ in (6.3). In particular, it iterates between the following two steps until the cluster assignment does not change.

1. For a given cluster assignment C, the total heterogeneity in (6.3) is minimized with respect to $\{m_1, \ldots, m_K\}$, which yields the cluster centers of the current cluster assignment C.

2. Given a set of cluster centers $\{m_1, \ldots, m_K\}$, (6.3) is minimized by assigning each observation to the most proximate cluster, i.e.,

$$C(i) = \arg\min_{1 \le k \le K}\|x_i - m_k\|^2.$$

6.1.2 Spectral Clustering

An effective approach to achieve an aggregation of vertices in a graph is through spectral theory. Spectral graphical theory has been applied to a number of problems including model reduction for complex Markov chain models [54], load balancing in parallel computation [55], and cluster analysis [52]. Spectral clustering refers to a family of graph-based algorithms, which achieve a partition over a data set by partitioning the vertices of the graph that is associated with such a data set. In particular, the partition is obtained by analyzing the eigen-structure of the adjacency matrix of the graph. As a result, spectral clustering captures the global information encoded in such a graph, and often outperforms traditional algorithms such as K-means or single linkage clustering.

Let $x_i, i = 1, \ldots, n$, be a set of observations. The similarity between each pair of observations x_i, x_j is denoted by $s_{i,j}, i \neq j, i, j = 1, \ldots, n$, which is either given, or that it can be computed using a certain similarity function. An example for such a similarity function is the Gaussian similarity function $s_{i,j} = \exp\{-\|x_i - x_j\|^2/(2\sigma^2)\}$, where the parameter σ controls the width of the neighborhood.

The similarity structure of the observations can be represented by a (weighted) *undirected graph* $G = (V, E)$, where V is the vertex set, and E the edge set. The vertex $v_i \in V$ represents the observation $x_i, i = 1, \ldots, n$. There are several popular methods to encode the similarity structure of the observations into a graph G. For example, in the ϵ-neighborhood method, we connect all pairs of nodes such that the pairwise similarity between them exceeds a certain thresholds ϵ, producing an unweighted graph. To construct fully connected graphs, we simply connect all pairs of nodes that have a positive similarity between them. Then we assign the edge $e_{i,j}$ that connects v_i and v_j with a weight equal to $s_{i,j}$, $i \neq j, i, j = 1, \ldots, n$. This method produces a weighted graph.

Given a (weighted) undirected graph G, we can construct the corresponding adjacency (similarity) matrix $W = (w_{i,j})_{i,j=1,\ldots,n}$. For both unweighted and weighted graphs, let $w_{i,j} = 0$ for all pairs of nodes that are not connected by an edge. For each v_i, v_j that are connected by an edge, $i \neq j, i, j = 1, \ldots, n$, we can set $w_{i,j} = 1$ if G is unweighted, and $w_{i,j} = s_{i,j}$ if G is weighted.

The sum of all edges connected to vertex v_i is called the *degree* of a vertex, i.e.,

$$d_i = \sum_{j=1}^{n} w_{i,j},$$

$i = 1, \ldots, n$. The *degree matrix* D is an $n \times n$ diagonal matrix with its diagonal entries being $d_1, d_2, \ldots, d_n$.

Most spectral clustering algorithms start with the notion of a *graph Laplacian* [56, 57], which is defined as

$$L = D - W. \tag{6.4}$$

The corresponding *normalized* symmetric graph Laplacian is given by

$$L_{\text{sym}} \triangleq D^{-1/2} L D^{-1/2}. \tag{6.5}$$

To measure the "size" of a subset $A \subset V$, we use

$$\text{vol}(A) \triangleq \sum_{i \in A} d_i.$$

For two subset of vertices A and B, $A, B \subset V$, the *cut* between the two sets is

$$W(A, B) \triangleq \sum_{i \in A, j \in B} w_{i,j}.$$

The goal of cluster analysis is to partition a collection of data objects to subgroups such that different groups are well "separated." Given the similarity matrix of a data set, a natural way to construct such a partition is to solve the so-called *mincut* problem, which for a given number of K subsets, one solves for the partition $A_1, A_2, \ldots, A_K$ such that

$$(A_1^\star, A_2^\star, \ldots, A_K^\star) = \underset{A_1, A_2, \ldots, A_K}{\arg\min} \frac{1}{2} \sum_{i=1}^{K} W(A_i, A_i^c),$$

where $A_i^c = V \setminus A_i$ is the complement of $A_i, i = 1, \ldots, K$.

In practice, the mincut approach may lead to clusters that are of imbalanced sizes [51]. To avoid this, an alternative objective is to minimize the so-called "Ncut", i.e.,

$$\text{Ncut}\,(A_1, A_2, \ldots, A_K) \triangleq \frac{1}{2} \sum_{i=1}^{K} \frac{W(A_i, A_i^c)}{\text{vol}(A_i)}. \tag{6.6}$$

Unfortunately, the minimization problem in (6.6) was shown to be intractable (NP hard) in [58]. Spectral clustering algorithms circumvent the problem of NP hardness by solving relaxed versions of (6.6). We now introduce one particular spectral clustering algorithm due to Ng, Jordan, and Weiss [53], which we compare with the FSS test in Section 3.1 in the next section. In particular, for a given number of K clusters, the algorithm consists of the following steps.

1. Compute the first K eigenvectors $u_1, u_2, \ldots, u_K$ of L_{sym} defined in (6.5).
2. Let $U \in \mathbb{R}^{n \times K}$ be the matrix containing $u_1, u_2, \ldots, u_K$ as column vectors.
3. Normalize the rows of U to produce the matrix $T \in \mathbb{R}^{n \times K}$.
4. Let $y_i \in \mathbb{R}^K$ be the vectors corresponding to the i-th row of T, $i = 1, \ldots, K$.
5. Cluster the vectors $y_i, i = 1, \ldots, n$, with the K-means algorithm to produce clusters $C_1, C_2, \ldots, C_K$.
6. Assign observation x_i to cluster C_j if $y_i \in C_j, i = 1, \ldots, n$.

It is shown in [51] that this algorithm solves a relaxed version of the Ncut problem in (6.6).

6.2 Fixed Sample Sizes Test as Clustering Algorithm

There is an interesting connection between universal outlier hypothesis testing and cluster analysis. In universal outlier hypothesis testing, an entire sequence can be considered a data object. Typical sequences are more closely related to one another than to an outlier sequence in the sense that the observations in them are distributed according to the same typical distribution. The same holds for outlier sequences when the outliers are identically distributed. In this case, outliers can be identified by clustering the sequences (objects) into two clusters, where the cluster with more members contains all typical sequences, and the other outliers. When the outliers are distinctly distributed, it is sufficient to identify one "dense region" among the sequences, and any sequence outside such a region is considered an outlier.

In fact, the FSS test in Section 3.4 can be interpreted as performing cluster analysis over the probability simplex. For instance, for the multiple outlier setting where the outliers are identically distributed and the number of outliers known, the decision of the FSS test is given by

$$\delta(y^{Mn}) = \underset{S \in \{1,\ldots,M\}, |S|=K}{\arg\min} \sum_{i \in S} D\left(\gamma_i \middle\| \frac{\sum_{k \in S} \gamma_k}{|S|}\right) + \sum_{j \notin S} D\left(\gamma_j \middle\| \frac{\sum_{k \notin S} \gamma_k}{|S^c|}\right). \tag{6.7}$$

First, by taking the empirical distribution of each sequence, the original data objects (sequences of n observations) that take values in $\mathbb{R}^n$ are transformed into the probability simplex. The center of a particular cluster is defined as the average of all objects belonging to the cluster, which is the same as in the K-means algorithm. In particular, in (6.7), the center

for the outlier cluster is $\frac{\sum_{k\in S}\gamma_k}{|S|}$, and $\frac{\sum_{k\notin S}\gamma_k}{|S^c|}$ for the typical cluster. However, distinct from the K-means algorithm, instead of using the Euclidean distance, we use the KL divergence as the dissimilarity measure in the FSS test. Then for every possible cluster assignment, the corresponding total heterogeneity is computed as the sum of dissimilarities between each object and its cluster center. It is easy to see that the first sum in (6.7) corresponds to the heterogeneity of the outlier cluster, and the second sum the typical cluster. Lastly, the test decides on the cluster assignment that yields the minimal total heterogeneity as in (6.7).

Our technical contributions are as follows. First, by transforming the sequences into the probability simplex, the dimension of the objects to be clustered becomes $|\mathcal{Y}|$, which does not scale with the number of observations n. In addition, our theoretical results suggest the use of the KL divergence as the dissimilarity measure in clustering sequences of i.i.d. observations. Specifically, when the outliers are identically distributed, and the number of outliers is known, the achievable error exponent of the FSS test converges to the optimal one when both μ and π are known as the number of sequences goes to infinity.

6.3 Numerical Results

We compare the performance of the FSS test in Section 3.1 with two other clustering algorithms on a synthetic data set for outlier hypothesis testing. In particular, we compare the FSS test with the spectral clustering algorithm outlined in Section 6.1.2. For the spectral clustering algorithm, for two sequences of observations $Y^{(i)} = (Y_1^{(i)}, \ldots, Y_n^{(i)})$, and $Y^{(j)} = (Y_1^{(j)}, \ldots, Y_n^{(j)})$, $i \neq j, i, j \in \{1, \ldots, M\}$, we adopt the pairwise Hamming distance as the similarity measure, i.e.,

$$w(i,j) \triangleq \sum_{k=1}^{n}\sum_{l=1}^{n} \mathbb{I}(Y_k^{(i)} = Y_l^{(j)}).$$

We also applied a combinatorial clustering algorithm using the L_2 distance as the dissimilarity measure [13]. Specifically, the combinatorial clustering algorithm solves the same optimization problem as in (6.7), but with the KL divergence replaced by the L_2 distance.

The particular choice of typical and outlier distributions are $\pi = (0.25, 0.41, 0.34)$ and $\mu = (0.1, 0.55, 0.35)$. There is exactly one outlier among $M = 5$ sequences. For different sample size n, we evaluate the probabilities of error incurred by the FSS test, the spectral clustering algorithm, and the combinatorial clustering algorithm, respectively. As we can see from the results in Figure 6.1, the spectral clustering algorithm using the pairwise Hamming distance outperforms the FSS test when n is small. For sufficiently large n, for this synthetic

data set, the FSS test outperforms the other two algorithms. The result suggest that it may be beneficial to use spectral clustering when the number of observations is limited, and when n is sufficiently large, the simulation results corroborate our theoretical findings in Theorem 3.

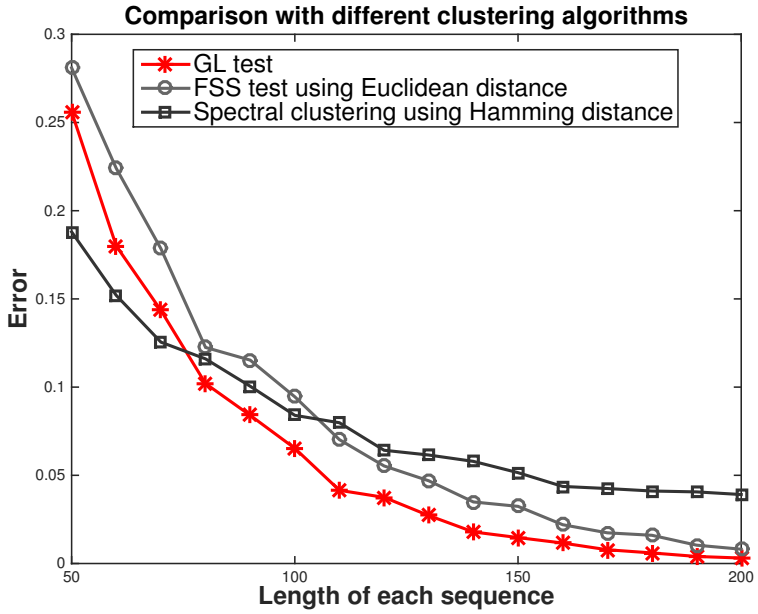

Figure 6.1: Performance of the FSS test, the spectral clustering algorithm, and the combinatorial clustering algorithm

CHAPTER 7

APPLICATION TO ANOMALY DETECTION

In this chapter, we evaluate the performance of the various proposed tests in Chapters 3, 4 and 5 on a spam detection data set. Multiple sequences of emails are collected. One of the sequences contains only spams, while the rest non-spams. The goal is to identify the outlier sequence that consists of only spams.

The data set contains information from 4610 emails (each being labeled as a spam or non-spam) addressed to an employee at Hewlett-Packard (HP)[13]. The information for each email consists of relative frequencies of a set of 48 words and 6 punctuation marks. We shall refer to the relative frequencies of such words and punctuation marks as *features*. There are 1813 spams among the 4601 emails.

The specific application that we envision pertains to identifying spam sources of an individual email account. Consider the situation where an email account may be spammed by a few vicious IP addresses, which constitute a small fraction of all possible IP addresses. Cast into the formulation of outlier hypothesis testing, each sequence consists of emails from a certain IP address. When an account is compromised, a small subset of the sequences are outliers that contain only spams, while the majority of the sequences are typical with non-spams. The goal is to decide whether an email account is compromised, and if so, which are the sources of spams.

The experiment is designed such that there is *exactly one* outlier sequence among $M = 6$ number of sequences. The outlier sequence contains only spams, and typical sequences non-spams. It is known that the values of certain features, such as the relative frequencies of "RE", "FREE", the name of the recipient, and the name of the company where the recipient is employed ("HP" and HP laboratory ("HPL")), tend to vary greatly between spams and non-spams [13]. In this experiment, we choose the relative frequencies of "HP", "HPL" and "RE" as the observations. Specifically, the k-th observation of sequence i, $i = 1, \ldots, M$, is $y_k^{(i)} = \left(y_{k,1}^{(i)}, y_{k,2}^{(i)}, y_{k,3}^{(i)}\right)$, where $y_{k,1}^{(i)}$ is the relative frequency of "HP" in the corresponding email, $y_{k,2}^{(i)}$ of "HPL", and $y_{k,3}^{(i)}$ of "RE." It is assumed that the coordinates of an observation are mutually independent, and identically distributed across the observations.

In the original data set, the features take continuous values in the finite interval of $[0, 100]$.

The tests described in Chapters 3 and 4 are only applicable when the observations take values in finite alphabets. In order to apply our proposed tests, the observations are first quantized, where the quantization intervals of a certain feature are appropriately chosen based on the distribution of the feature values over all emails, regardless of their labels. Specifically, for a certain feature, the region in [0, 100] which finds the majority of the values of said feature is quantized more finely than other regions. There are five levels in the quantizations for "HP" and "HPL", and six levels for "RE". The value of each quantization interval is chosen to be the midpoint of that interval.

We first compare the fixed sample size GL test (3.41), and the MMD-based tests in [45]. One advantage of the MMD-based test is that it is applicable when the underlying distributions are continuous. In this experiment, we implement the MMD-based test using the original data (continuous), the quantized data, and the indices of the quantization intervals, respectively. The numerical results are obtained by averaging over a number of trials. It is shown in Figure 7.1 that the GL test outperforms all three MMD-based tests for large enough n, which agrees with the optimality result of the GL test in Theorem 3. In particular, the GL test outperforms the MMD-based tests when the length of the sequences n is larger than 20. This is due to the fact that an intermediate step of the GL test is to estimate the KL divergence between the underlying distributions (cf. (3.10)), which becomes more accurate as n increases.

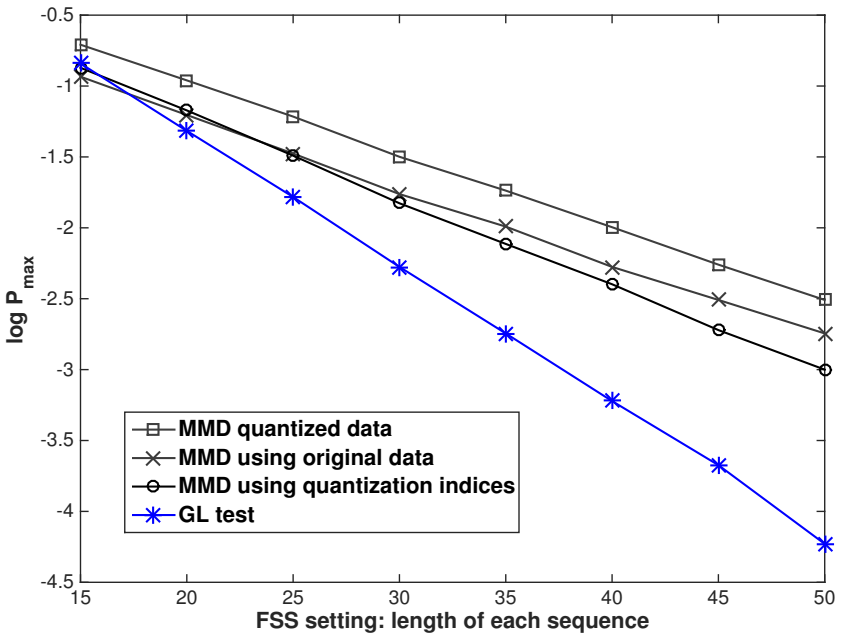

Figure 7.1: Comparison between various fixed sample size tests

The sequential test in (4.9) – (4.11) has a stopping time that depends on a deterministic time horizon, which is appropriately chosen to accommodate the null hypothesis. In this experiment, there is exactly one outlier sequence among M sequences. So the stopping time can be simplified to be

$$N \triangleq \operatorname*{argmin}_{n\geq 1} \Bigg[\min_{j\neq\hat{i}} \; n\Big[\sum_{k\neq j} D\left(\gamma_k \Big\| \tfrac{\sum_{\ell\neq j}\gamma_\ell}{M-1}\right) - \sum_{k\neq\hat{i}} D\left(\gamma_k \Big\| \tfrac{\sum_{\ell\neq\hat{i}}\gamma_\ell}{M-1}\right) \Big] \\ > \log T + M|\mathcal{Y}| \log(n+1) \Bigg]. \tag{7.1}$$

At the stopping time, the test decides on the most probable hypothesis, i.e.,

$$\delta = \hat{i}(Y^{MN}), \tag{7.2}$$

where $\hat{i}(Y^{Mn}) \triangleq \arg\min_{i=1,\ldots,M} \hat{p}_i^{\text{univ}}(Y^{Mn})$.

We then apply the sequential test in (7.1) and (7.2) to the quantized data with a series of increasing thresholds T. For each T, the sequential test is repeated a number of trials using bootstrap samples (we randomly permute the emails when we run out of data, and reuse the permuted data). The comparison between the sequential test and the fixed sample size GL test is shown in Figure 7.2. We see that the sequential test starts to outperform the fixed sample size GL test (and all different versions of the MMD tests) when the average stopping time exceeds 30.

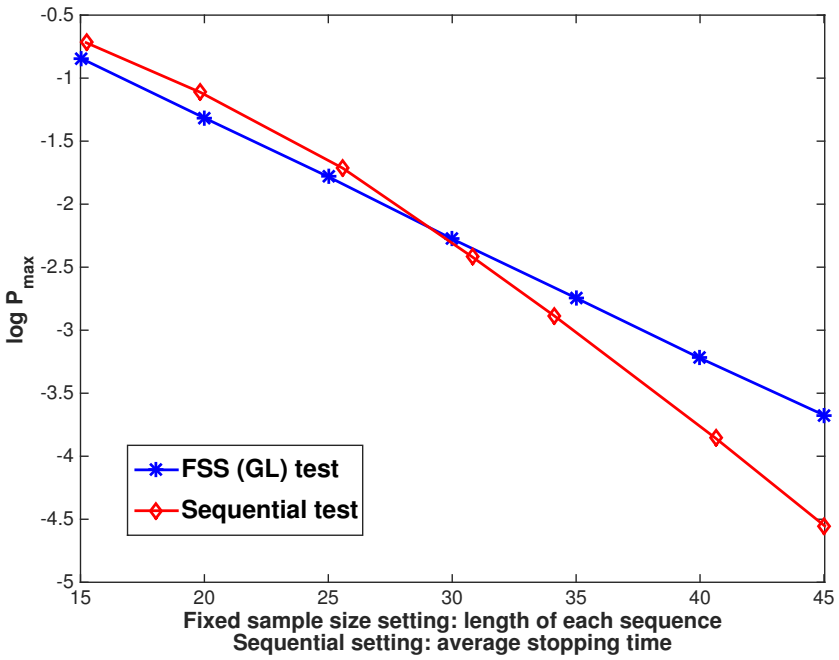

Figure 7.2: Comparison between the sequential test and fixed sample size GL test

CHAPTER 8

CONCLUSION AND FUTURE WORK

We formulated and studied the problem of outlier hypothesis testing in various universal settings. Our main contribution was in proposing tests that yield exponentially decaying probability of error with the number of observations for both the fixed sample size (FSS) and the sequential settings.

In the FSS setting, for the case with exactly one outlier, we showed that the generalized likelihood (GL) test is universally exponentially consistent. We also provided a characterization of the error exponent achievable by the GL test for each $M \geq 3$. Surprisingly the GL test is not only universally exponentially consistent, but also asymptotically optimal as the number of sequences goes to infinity. Specifically, as M goes to infinity, the error exponent achievable by the GL test converges to the absolutely optimal error exponent when both the outlier and typical distributions are known. When there is an additional null hypothesis, we showed that a suitable modification of the GL test achieves exponential consistency under each hypothesis with the outlier, and consistency under the null hypothesis universally. Under every non-null hypothesis, this modified test achieves the same error exponent as that achievable when the null hypothesis is excluded. We then extended our models to cover the case with more than one outlier. When the outliers can be distinctly distributed, even with the typical distribution being known, we proved that there cannot exist a universally exponentially consistent test if the number of outliers is not completely known. For models with a known number of outliers, the distributions of the outliers could be distinct as long as each of them differs from the typical distribution. We showed that The GL test is universally exponentially consistent for such a setting. Furthermore, we characterized the limiting error exponent achieved by the GL test, and established its universally asymptotically efficiency. For models with an *unknown* number of *identically* distributed outliers, we proved that the GL test is universally exponentially consistent when the null hypothesis is excluded. When the null hypothesis is included, we showed that a slight modification of the GL test achieves a positive error exponent under every non-null hypothesis, and also consistency under the null hypothesis universally. We also extended our theoretical findings to the setting with continuous alphabets. We proposed a test similar to the GL test for finite alphabets to

accommodate continuous alphabets, and proved that such a test is universally consistent.

In the sequential setting, we proposed a sequential test with the flavor of the repeated significance test and showed that it is universally consistent. With at most one outlier and with the typical distribution being known, we showed that the achievable error exponent of the proposed sequential test is the same as the absolutely optimal one when the outlier is present. The test is also asymptotically optimal in the limit of the large number of sequences when neither the outlier nor typical distribution is known. When there might be multiple outliers, we established that the test is asymptotically optimal universally when the number of outliers is the largest possible and when the typical distribution is known. We also characterized the asymptotic performance of the test when the typical distribution is not known either. We then extended our findings to the model with multiple distinct outliers. In all cases, we proved that as the number of sequences goes to infinity, the asymptotic performance of the proposed sequential test when neither the outlier nor the typical distribution is known converges to that when the typical distribution is known.

We end with a discussion of possible future work. In the case with multiple outliers, although the proposed test was shown to be asymptotically optimal, the complexity of its implementation scales exponentially with the number of outliers. When the number of outliers can be large, it is desirable to seek a more practical sub-optimal test that sequentially picks out one outlier at a time and terminates when it is determined that there are no outliers left. It is suggested by the numerical results in Section 6.3 and Chapter 7 that although the proposed tests are shown to be asymptotically optimal in the limit as the number of sequences goes to infinity in various universal settings, when the number of observations is limited, there may exist other tests that outperform the proposed tests (cf. Figures 6.1, 7.1 and 7.2). A direction for future research is to study tests that are more appropriate for small sample size. Toward this end, an intermediate step is to derive the exact asymptotics for the proposed tests, since such asymptotics can often be much more precise for small to moderate sample size as compared to the standard exponential approximation that we studied in this dissertation [59]. Another interesting extension is to study feature selection methods for universal outlier hypothesis testing. In applications such as cancer screening and telematics analysis, the observations can have a large number of dimensions. An individual dimension is usually referred to as a *feature*. It is possible that only a few of the large number of features are relevant in detecting outliers. Our proposed test may not perform well if all features are regarded as equally important in detecting outliers. If that is the case, before applying the proposed tests, a critical step is to identify relevant features and "filter" out irrelevant ones. Numerous feature selection methods have been proposed for supervised and semi-supervised learning problems including methods for subset selection, shrinkage methods,

and cross-validation [13]. However, the aforementioned techniques require training samples with known class labels (for classification) or prediction values (for regression). It remains to investigate how one can perform feature selection in the completely universal setting that is considered here.

REFERENCES

[1] R. J. Bolten and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, pp. 235–249, 2002.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15.1–15.58, 2009.

[3] J. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, pp. 16–25, 2007.

[4] V. H. Poor, *An Introduction to Signal Detect and Estimation.* Springer, 1994.

[5] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inf. Theory*, vol. 38, pp. 1597–1602, 1992.

[6] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369–401, Apr. 1965.

[7] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, pp. 250–254, 1911.

[8] O. Shiyevitz, "On Rényi measures and hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 31-Aug. 5 2011, pp. 894–898.

[9] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 1-6 2012, pp. 2027–2031.

[10] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, pp. 278–286, Mar. 1988.

[11] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, pp. 401–408, Mar. 1989.

[12] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed. John Wiley and Sons, Inc., 2011.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements in Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. Available at http://statweb.stanford.edu/tibs/ElemStatLearn/.

[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[15] P. Than, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Pearson, 2005. Available at http://www-users.cs.umn.edu/ kumar/dmbook/index.php.

[16] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the 5th Berkeley Symp. on Math. Statist. and Prob.*, Jun. 21-Jul. 18 1967, Dec. 27-Jan. 7 1968, pp. 281–297.

[17] G. Ball and D. Hall, "A clustering technique for summarizing multivariate data," *Behavior Science*, vol. 12, pp. 153–155, 1967.

[18] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, Inc., 1990.

[19] P. Macnaughton-Smith, W. T. Williams, M. B. Dale, and G. Mockett, "Dissimilarity analysis: A new technique of hierarchical sub-division," *Nature*, vol. 202, pp. 1034–1035, 1964.

[20] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *KDD Workshop on Test Mining, 2000*, 2000.

[21] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of the 2nd Intl. Conf. on Knowledge Discovery and Data Mining*, Aug. 2-4 1996, pp. 226–231.

[22] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications," *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, 1998.

[23] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, pp. 27–64, 2007.

[24] V. Barnett, "The study of outliers: Purpose and model," *Appl. Stat.*, vol. 27, no. 3, pp. 242–250, 1978.

[25] D. Hawkins, *Identification of Outliers.* Chapman and Hall, 1980.

[26] V. V. Veeravalli and C. W. Baum, "Asymptotic efficiency of a sequential multihypothesis test," *IEEE Trans. Inf. Theory*, vol. 41, pp. 1994–1997, Nov. 1995.

[27] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests—Part I: Asymptotic optimality," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2448–2461, Nov. 1999.

[28] M. Woodroofe, *Nonlinear Renewal Theory in Sequential Analysis*, ser. CBMS-NSF regional conference series in applied mathematics. SIAM, 1982.

[29] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, ser. Springer series in statistics. Springer-Verlag, 1985.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Wiley and Sons, Inc., 2006.

[31] W. Hoeffding and J. Wolfowitz, "Distinguishability of sets of distributions," *Ann. Math. Statist.*, vol. 29, no. 3, pp. 700–718, June 1958.

[32] B. Levy, *Principles of Signal Detection and Parameter Estimation*. New York: Springer, 2008.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[34] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[35] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inf. Theory*, vol. 44, pp. 1726–1745, 1998.

[36] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Math. Contr. Signals Syst.*, vol. 1, pp. 167–182, 1988.

[37] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, pp. 117–186, 1945.

[38] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Statist.*, vol. 19, pp. 326–339, 1948.

[39] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *IEEE Trans. Inf. Theory*, vol. 40, pp. 1994–2007, Nov. 1994.

[40] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests—Part II: Accurate asymptotic expansions for the expected sample size," *IEEE Trans. Inf. Theory*, vol. 46, pp. 1136–1383, July 2000.

[41] S. Zacks, *Theory of Statistical Inference (Probability and Mathematical Statistics)*. John Wiley and Sons, Inc., 1971.

[42] T. L. Lai, "Nearly optimal sequential tests of composite hypotheses," *Ann. Statist.*, vol. 16, pp. 856–886, 1988.

[43] F. Mosteller, "A k-sample slippage test for an extreme population," *Ann. Math. Statist.*, vol. 19, pp. 58–65, 1948.

[44] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

[45] S. Zou, Y. Liang, V. H. Poor, and X. Shi, "Nonparametric detection of anomalous data via kernel mean embedding," *IEEE Trans. Inf. Theory*, submitted, 2014.

[46] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernal two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.

[47] Q. Wang, S. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, pp. 3064–3074, Sep. 2005.

[48] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *J. Classification*, vol. 5, pp. 5–48, 1986.

[49] M. Schwaiger and O. Opitz, *Exploratory Data Analysis in Empirical Research*. Springer, 2002.

[50] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, *The Maximum Clique Problem*. Boston, MA, USA: Kluwer Academic Publishers, 1999.

[51] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[52] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888–905, 2000.

[53] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural Inf. Process. Syst*, vol. 14, pp. 849–856, 2002.

[54] K. Deng, P. Mehta, and S. P. Meyn, "Optimal Kullback-Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. Autom. Contr.*, vol. 56, pp. 2793–2807, 2011.

[55] B. Hendrickson and R. Leland, "Multidimensional spectral load balancing," Sandia National Laboratories, Albuquerque, NM, Tech. Rep. SAND93-0074, 1993.

[56] B. Mohar, "The Laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications*. Kalamazoo, MI: John Wiley and Sons, Inc., 1991, vol. 2, pp. 871–898.

[57] B. Mohar, "Some applications of Laplace eigenvalues of graphs," in *Graph Symmetry: Algebraic Methods and Applications*, G. Hahn and G. Sabidussi, Eds. Kluwer, 1997, vol. NATO ASI Ser. C 497, pp. 871–898.

[58] D. Wagner and F. Wagner, "Between min cut and graph bisection," in *Proc. of the 18th Intl. Symp. on Mathematical Foundations of Computer Science*, 1993, pp. 744–750.

[59] Y. Huang and P. Moulin, "Strong large deviations for composite hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 29-Jul. 4 2014, pp. 556–560.