Universal sequential outlier hypothesis testing

Yun Li*, Sirin Nitinawarat**, and Venugopal V. Veeravalli

Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA

ABSTRACT
Universal outlier hypothesis testing is studied in a sequential setting. Multiple observation sequences are collected, a small subset of which are outliers. A sequence is considered an outlier if the observations in that sequence are generated by an "outlier" distribution, distinct from a common "typical" distribution governing the majority of the sequences. Apart from being distinct, the outlier and typical distributions can be arbitrarily close. The goal is to design a universal test to best discern all the outlier sequences. A universal test with the flavor of the repeated significance test is proposed and its asymptotic performance, as the error probability goes to zero, is characterized under various universal settings. The proposed test is shown to be universally consistent. For the model with at most one outlier, conditioned on the outlier being present, the test is shown to be asymptotically optimal universally when the typical distribution is known and as the number of sequences goes to infinity when neither the outlier nor the typical distribution is known. With multiple identical outliers, the test is shown to be asymptotically optimal universally when the number of outliers is the largest possible and with the typical distribution being known, and its asymptotic performance when neither the outlier nor the typical distribution being known is also characterized. Extensions of the findings to models with multiple distinct outliers are also discussed. In all cases, it is shown that the asymptotic performance guarantees for the proposed test when neither the outlier nor the typical distribution is known converge to those when the typical distribution is known as the number of sequences goes to infinity.

ARTICLE HISTORY
Received 20 March 2017
Revised 12 April 2017
Accepted 8 July 2017

KEYWORDS
Anomaly detection; consistency; data-driven classification; exponential consistency; fraud detection; generalized likelihood test; multihypothesis sequential probability ratio test; nonparametric sequential testing; outlier detection; repeated significance test

1. Introduction
We consider the following inference problem of outlier hypothesis testing in a sequential setting. Among a fixed number of independent and memoryless observation sequences, it is assumed that a small subset (possibly empty) of these sequences are outliers. 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 a nonparametric setting, in which the outlier and
typical distributions are not fully known and can be arbitrarily close. The goal is to design a
universal test to identify all of the outlier sequences using the fewest observations.

In Li et al. (2014), 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 (Pearson, 1911; Ziv, 1988; Gutman, 1989).
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 (Ziv, 1988; Gutman, 1989).
We also showed that the GL test is asymptotically optimal in the limit of the large number of
sequences. In this article, 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
(1945). 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., Wald, 1945; Wald and Wolfowitz, 1948; Baum and Veeravalli, 1994; Veeravalli
and Baum, 1995; Dragalin et al., 1999, 2000). 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., Zacks, 1971), and the
other based on a maximum (generalized) likelihood function for each hypothesis (see, e.g.,
Lai, 1988). There have also been a limited number of papers on nonparametric 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., Mosteller, 1948). Sequential outlier hypothesis
testing is closely related to the so-called slippage problem studied in the sequential setting (see,
e.g., Ferguson, 1967). 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 (and sometimes no information regarding
the typical distribution either). 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. On the other hand, we make the simplifying
assumption that each instantaneous observation takes a 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 (Woodroofe, 1982; Siegmund,
1985), wherein the test stops when the generalized likelihood 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.
2. Preliminaries

Throughout the article, random variables (rvs) are denoted by capital letters, and their realizations are denoted by the corresponding lowercase letters. All rvs are assumed to take values in finite alphabets, and all logarithms are the natural one.

Our results will be stated in terms of certain distance metrics between a pair of distributions $p, q$ on $\mathcal{Y}$: the Bhattacharyya distance and the relative entropy, denoted by $B(p, q)$ and $D(p||q)$, respectively, and defined as (see, e.g., Cover and Thomas, 2006)

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

and

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

respectively. There is a relationship between these two distances (cf. Li et al., 2014; lemma 2):

$$2B(\mu, \pi) = \min_{\gamma} \left[ D(\gamma||\mu) + D(\gamma||\pi) \right].$$  \hspace{1cm} (2.1)

For a finite set $\mathcal{Y}$, let $\mathcal{Y}^n$ denote the $n$ 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 $y = y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$, denoted by $\gamma = \gamma_y = \gamma^{(n)} \in \mathcal{P}(\mathcal{Y})$, is defined as

$$\gamma(y) \triangleq \frac{1}{n} \left| \left\{ k = 1, \ldots, n : y_k = y \right\} \right|,$$

$y \in \mathcal{Y}$.

The following technical facts will be useful; their derivations can be found in chapter 11 of Cover and Thomas (2006). Consider random variables $Y^n$ that are independent and identically distributed (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 a sequence $y^n$ under $p$ is

$$p(y^n) = \exp \left\{ -n \left( D(\gamma||p) + H(\gamma) \right) \right\},$$ \hspace{1cm} (2.2)

where $H(\gamma)$ is the entropy of $\gamma$, defined as

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

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

$$\gamma(y^n) = \exp \{ -nH(\gamma) \}. \hspace{1cm} (2.3)$$

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, (2.2), and a bound on the size of the set of sequences with the same empirical distribution (see, e.g., theorem 11.1.3 in Cover and Thomas (2006) for details), it can be shown that the probability that the i.i.d. sequence $Y^n$ that is distributed according $p$ has the empirical
distribution $\gamma = q$ (for a feasible $q$) satisfies
\[ P\{\gamma = q\} \leq e^{-nD(q\|p)}. \tag{2.4} \]

We shall also find the following “sum centroid” inequality and its consequence useful. Consider any collection of distributions on $Y : p_i, i \in C$. Then, for any arbitrary distribution $q$,
\[ \sum_{i \in C} D\left(p_i \left\| \frac{\sum_{j \in C} p_j}{|C|}\right\| q\right) \leq \sum_{i \in C} D(p_i \| q) \tag{2.5}. \]

The proof of (2.5) follows from the fact that for any distribution $q$,
\[ \sum_{i \in C} D\left(p_i \left\| \frac{\sum_{j \in C} p_j}{|C|}\right\| q\right) = \sum_{i \in C} D(p_i \| q) - |C|D\left(\frac{\sum_{i \in C} p_i}{|C|} \| q\right). \]

Now for a pair of distributions $p, \tilde{p}$ on $Y$, particularizing (2.5) to the special case where $C$ comprises one $p$ distribution and $L$ copies of the $\tilde{p}$ distribution and with $q$ in (2.5) being $\tilde{p}$, we have that
\[ D\left(p \left\| \frac{p + L\tilde{p}}{L + 1}\right\| q\right) + LD\left(\frac{p + L\tilde{p}}{L + 1} \| q\right) \leq D(p \| \tilde{p}). \tag{2.6} \]

3. **Universal sequential outlier hypothesis testing**

3.1. **Model with at most one outlier**

Consider $M \geq 3$ independent sequences, each of which consist of i.i.d. observations. Denote the $k$th observation of the $i$th sequence by $Y_k^{(i)} \in Y$. It is assumed that there is at most one outlier among the $M$ sequences. In particular, if the $i$th sequence is the outlier, the observations in that sequence are distributed (i.i.d.) according to the outlier distribution $\mu \in P(Y)$, while all of the other sequences are distributed according to the typical distribution $\pi \in P(Y)$. Under the hypothesis with no outlier, namely, the null hypothesis, all sequences are commonly distributed according to the typical distribution. We consider the settings where either only $\mu$ is unknown or both $\mu$ and $\pi$ are unknown. In all cases, it is assumed that both $\mu$ and $\pi$ have full supports over $Y$. The assumption of $\mu, \pi$ having full supports rules out special cases where it is easier to identify the outlier sequence.

When the $i$th sequence is the outlier, the joint distribution of the first $n$ observations is
\[ p_i (y^n) = p_i (y_1, \ldots, y_n) = \prod_{k=1}^{n} \left\{ \mu \left(y_k^{(i)}\right) \prod_{j \neq i} \pi \left(y_k^{(j)}\right) \right\}, \tag{3.1} \]
where
\[ y_k = (y_k^{(1)}, \ldots, y_k^{(M)}), \quad k = 1, \ldots, n. \]

Under the null hypothesis, the joint distribution of the first $n$ observations is given by
\[ p_0 (y^n) = \prod_{k=1}^{n} \prod_{i=1}^{M} \pi \left(y_k^{(i)}\right). \]
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 taken until a final decision is made. At the stopping time $N = n$, the decision is made based on a rule $\delta : Y^Mn \rightarrow \{0, 1, \ldots, M\}$. 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 $\pi$ is known and the completely universal setting where neither $\mu$ nor $\pi$ is known. Consequently, a universal test is not allowed to be a function of $\mu$, and of $\mu$ or $\pi$, in the respective settings.

The accuracy of a sequential test is gauged using the maximal error probability $P_{\text{max}}$, which is a function of both the test and $(\mu, \pi)$ and is defined as

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

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; that is, there exists $\alpha_i > 0$ such that for any $\mu, \pi, \mu \neq \pi$, as $P_{\text{max}} \rightarrow 0$, it holds that

$$\mathbb{E}_i [N] \leq -\frac{\log P_{\text{max}}}{\alpha_i} (1 + o(1)), \quad (3.2)$$

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

We first consider the setting where both the typical and outlier distributions are known. In this nonuniversal setting, the multihypothesis sequential probability ratio test (MSPRT) was shown to be asymptotically optimal in the limit as the error probability goes to zero (Dragalin et al., 1999). For a given threshold $T > 1$ and with $\hat{i}(y^n) \triangleq \arg\max_{i=0,1,\ldots,M} p_i (y^n)$, denoting the instantaneous maximum likelihood (ML) estimate of the hypothesis at time $n$, the stopping time $N^*$ and the final decision rule $\delta^*$ of the MSPRT are defined as follows:

$$N^* = \arg\min_{n \geq 1} \left[ \frac{p_i (y^n)}{\max_{j \neq i} p_j (y^n)} > T \right], \quad (3.3)$$

$$\delta^* = \hat{i} \left( Y^{N^*} \right). \quad (3.4)$$

The following result (cf. Veeravalli and Baum, 1995; Dragalin et al., 1999) characterizes the asymptotic optimality of the MSPRT when the distributions of the observations are known.

**Proposition 3.1.** As the threshold $T$ in (3.3) approaches infinity, the MSPRT in (3.3) and (3.4) satisfies

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

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

$$\mathbb{E}_i [N^*] = \frac{\log T}{D (\mu \| \pi)} (1 + o(1)) = -\frac{\log P_{\text{max}}}{D (\mu \| \pi)} (1 + o(1)),$$
and
\[ E_0[N^*] = \frac{\log T}{D(\pi \| \mu)} (1 + o(1)) = -\frac{\log P_{\text{max}}}{D(\pi \| \mu)} (1 + o(1)). \]

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests \((N, \delta)\) with vanishing maximal error probability, it holds for every \(i = 1, \ldots, M\), that
\[ E_i[N] \geq -\frac{\log P_{\text{max}}}{D(\mu \| \pi)} (1 + o(1)), \tag{3.5} \]
and that
\[ E_0[N] \geq -\frac{\log P_{\text{max}}}{D(\pi \| \mu)} (1 + o(1)). \tag{3.6} \]

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 shown in Li et al. (2014) 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 for the sequential setting.

### 3.1.1. Proposed universal test

For each \(i = 1, \ldots, M\), denote the empirical distribution of \(y^{(i)}\) by \(\gamma_i\). When only \(\pi\) is known, we compute the generalized likelihood (GL) of \(y^n\) under each hypothesis \(i\) by replacing the unknown \(\mu\) in (3.1) with its ML estimate \(\hat{\mu}_i \triangleq \gamma_i\) as
\[
\hat{p}_i^{\text{typ}}(y^n) = \prod_{k=1}^n \left\{ \hat{\mu}_i \left( y^{(i)}_k \right) \prod_{j \neq i} \pi \left( y^{(j)}_k \right) \right\}. \tag{3.7}
\]

Similarly, when neither \(\pi\) nor \(\mu\) is known, for each \(i = 1, \ldots, M\), we compute the GL of \(y^n\) under each hypothesis \(i\) by replacing the unknown \(\mu\) and \(\pi\) in (3.1) with their ML estimates \(\hat{\mu}_i \triangleq \gamma_i\) and \(\hat{\pi}_i \triangleq \frac{1}{M-1} \sum_{j \neq i} \gamma_j\) as
\[
\hat{p}_i^{\text{univ}}(y^n) = \prod_{k=1}^n \left\{ \hat{\mu}_i \left( y^{(i)}_k \right) \prod_{j \neq i} \hat{\pi}_i \left( y^{(j)}_k \right) \right\}. \tag{3.8}
\]

Our universal test has stopping and final decision rules similar to those of the MSPRT in (3.3) and (3.4), but with the unknown likelihood functions \(p_i(y^n)\), \(i = 1, \ldots, M\), being replaced with the corresponding GL functions. Another key idea in the test is the adoption of a time-dependent threshold.

When only \(\pi\) is known and with \(\hat{i} \triangleq \arg\max_{i=1,\ldots,M} \hat{p}_i^{\text{typ}}(y^n) = \arg\max_{i=1,\ldots,M} D(\gamma_i \| \pi)\), denoting the instantaneous estimate of the non-null hypothesis (using the GL) at time \(n\),
consider the following (stopping) time:

\[
\tilde{N} \triangleq \arg\min_{n \geq 1} \left[ \frac{\hat{p}_i^{\text{typ}} (y^n)}{\max_{j \neq i} \hat{p}_j^{\text{typ}} (y^n)} > T(n + 1)^M Y \right]
\]

\[
= \arg\min_{n \geq 1} \left[ \min_{j \neq i} n (D(\gamma_j || \pi) - D(\gamma_i || \pi)) > \log T + M |Y| \log(n + 1) \right].
\] (3.9)

Our test stops at this time or at \([T \log T]\), depending on which one is smaller; that is,

\[
N^* = \min \left( \tilde{N}, [T \log T] \right)
\] (3.10)

and, correspondingly, the final decision is made according to

\[
\delta^* = \begin{cases} 
\hat{i} (Y^{N^*}) & \text{if } \tilde{N} \leq T \log T; \\
0 & \text{if } \tilde{N} > T \log T.
\end{cases}
\] (3.11)

Similarly, when neither \(\mu\) nor \(\pi\) is known, the test can be described by the following stopping and final decision rules:

\[
N^* = \min \left( \tilde{N}, [T \log T] \right),
\] (3.12)

\[
\delta^* = \begin{cases} 
\hat{i} (Y^{N^*}) & \text{if } \tilde{N} \leq T \log T; \\
0 & \text{if } \tilde{N} > T \log T,
\end{cases}
\] (3.13)

where

\[
\tilde{N} \triangleq \arg\min_{n \geq 1} \left[ \min_{j \neq i} n \left( \sum_{k \neq j} D(\gamma_k || \pi) - \sum_{k \neq \hat{i}} \left( \gamma_k \sum_{\ell \neq \hat{i}} \gamma_{\ell} M - 1 \right) \right) \right. \\
\left. > \log T + M |Y| \log(n + 1) \right],
\] (3.14)

and \(\hat{i} = \hat{i} (y^n) \triangleq \arg\min_{i=1,...,M} \sum_{k \neq i} D(\gamma_k || \pi) \sum_{\ell \neq \hat{i}} \gamma_{\ell} M - 1 \). This proposed universal test has the flavor of the repeated significance test (Woodroofe, 1982; Siegmund, 1985), wherein 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 happens before a predetermined deadline.

### 3.1.2. Performance of proposed test

**Theorem 3.1.** When only \(\pi\) is known, for every \(M\) and any \(\mu \neq \pi\), the proposed test in (3.9), (3.10) and (3.11) is universally consistent and yields for every \(T\) that

\[
P_{\max} \leq O \left( \frac{1}{T} \right),
\] (3.15)
where the constant in the term $O \left( \frac{1}{T} \right)$ in (3.15) depends only on $M, \mu, \pi$, but not on $T$. In addition, for each $i = 1, \ldots, M$, as $T \to \infty$,

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

(3.16)

Remark 3.1. While attaining universal consistency under the null hypothesis, the proposed test in (3.9), (3.10), and (3.11) 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. (3.5)).

Theorem 3.2. When neither $\mu$ nor $\pi$ is known, for every $M$ and any $\mu \neq \pi$, the proposed test in (3.12), (3.13), and (3.14) is universally consistent and yields for every $T$ that

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

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

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

(3.17)

Remark 3.2. Applying (2.6) with $L = M - 2, p = \mu$, and $\tilde{\rho} = \pi$, we get that

$$
D \left( \mu \parallel \frac{1}{M - 1} \mu + \frac{M - 2}{M - 1} \pi \right) + (M - 2)D \left( \pi \parallel \frac{1}{M - 1} \mu + \frac{M - 2}{M - 1} \pi \right) \leq D(\mu \parallel \pi).
$$

(3.18)

This is consistent with (3.19) and (3.5). It also follows from (3.19) that

$$
\lim_{M \to \infty} D \left( \mu \parallel \frac{1}{M - 1} \mu + \frac{M - 2}{M - 1} \pi \right) + (M - 2)D \left( \pi \parallel \frac{1}{M - 1} \mu + \frac{M - 2}{M - 1} \pi \right) = D(\mu \parallel \pi);
$$

that is, as $M \to \infty$, the asymptotic performance of the test in (3.12), (3.13), and (3.14) under each non-null hypothesis (cf. 3.18) when neither $\mu$ nor $\pi$ is known, approaches the (optimal) asymptotic performance of the test in (3.9), (3.10), and (3.11) under each of those hypotheses (cf. 3.16 and 3.5) when only $\pi$ is known.

Remark 3.3. The particular functional forms of the time-dependent threshold in (3.9) and (3.14) and of the deterministic time horizon in (3.10) and (3.12) are chosen solely for the simplicity of exposition. In fact, it follows from our proofs that the results in Theorems 3.1 and 3.2 continue to hold when the stopping time takes a more general form as follows. Consider

$$
\tilde{N} \triangleq \arg\min_{n \geq 1} \left[ \frac{\hat{p}_i(y^n)}{\max_{j \neq i} \hat{p}_j(y^n)} > C \left( T(n + 1)^M \right) \right],
$$

(3.20)

where for each $i = 1, \ldots, M, \hat{p}_i = \hat{p}_i^{\text{typ}}$ for the setting where $\pi$ is known, and $\hat{p}_i = \hat{p}_i^{\text{univ}}$ 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 $[f(T)]$, depending on which one is smaller; that is,

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

and, correspondingly, the final decision is made according to

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

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

4. Model with multiple outliers

We now generalize our results in the previous section to model with multiple outliers. In particular, we assume that there are up to $K$ outliers among the $M$ sequences with $K < \frac{M}{2}$ and with all of the outliers being identically distributed according to $\mu$. It was shown in Li et al. (2014) that in the fixed sample size setting, this 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 $K$). In Section 5, we shall look at the extension with possibly distinctly distributed outliers but with their total number being known.

Let $S \subset \{1, \ldots, M\}$, $|S| < \frac{M}{2}$, denote the set (possibly empty) of outliers. Conditioned on $S$ being the set of outliers, the joint distribution of the first $n$ observations is given by

$$p_S \left( y^n \right) = p_S \left( y_1, \ldots, y_n \right) = \prod_{k=1}^{n} \left\{ \prod_{i \in S} \mu \left( y_k^{(i)} \right) \prod_{j \notin S} \pi \left( y_k^{(j)} \right) \right\}.$$  \hspace{1cm} (4.1)

As previously, we refer to the unique hypothesis with no outlier as the null hypothesis with the joint distribution $p_0 \left( y^n \right) = p_\emptyset \left( y^n \right)$, wherein all of the sequences are distributed (i.i.d.) according to $\pi$.

The test for the outliers is done based on a universal rule $\delta \left( Y^N \right) \in S$, where $S$ 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

$$P_{\text{max}} \triangleq \max_{S \in S} \mathbb{P} \{ \delta \left( Y^N \right) \neq S \}.$$ \hspace{1cm} (4.2)

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 (3.2).

As in the previous section, for the setting with both the typical and outlier distributions being known and with $\hat{S} \left( y^n \right) \triangleq \arg\max_{S \in S} p_S \left( y^n \right)$, the MSPRT with the stopping and final decision rules being

$$N^* = \arg\min_{n \geq 1} \left[ \frac{\hat{P}_{\hat{S}} \left( Y^n \right)}{\max_{S \neq \hat{S}} p_S \left( Y^n \right)} > T \right],$$ \hspace{1cm} (4.3)

$$\delta^* = \hat{S} \left( Y^{N^*} \right)$$ \hspace{1cm} (4.4)

is asymptotically optimal (cf. Veeravalli and Baum, 1995; Dragalin et al., 1999).
Proposition 4.1. As the threshold $T$ in (4.3) approaches infinity, the MSPRT in (4.3) and (4.4) satisfies

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

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

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

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

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

and

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

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests $(N, \delta)$ with vanishing maximal error probability, it holds for each $S \in \mathcal{S}, |S| = K$, that

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

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

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

and that

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

4.1. Proposed universal test

When only $\pi$ is known, we compute the GL of $y^n$ under each non-null hypothesis corresponding to a nonempty subset $S \subset \{1, \ldots, M\}$ by replacing the unknown $\mu$ in (4.1) with its ML estimate $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} y_i}{|S|}$ as

$$
\hat{P}_S^\text{typ} \left(y^n\right) = \prod_{k=1}^{n} \left\{ \prod_{i \in S} \hat{\mu}_S(y_k^{(i)}) \prod_{j \notin S} \pi(y_k^{(j)}) \right\}. \tag{4.5}
$$

Similarly, when neither $\pi$ nor $\mu$ is known, we compute the GL of $y^n$ under each non-null hypothesis corresponding to a nonempty $S \in \mathcal{S}$ by replacing the unknown $\mu$ and $\pi$ in (4.1) with their ML estimates $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} y_i}{M - |S|}$ and $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} y_j}{M - |S|}$, respectively, as

$$
\hat{P}_S^\text{univ} \left(y^n\right) = \prod_{k=1}^{n} \left\{ \prod_{i \in S} \hat{\mu}_S(y_k^{(i)}) \prod_{j \notin S} \hat{\pi}_S(y_k^{(j)}) \right\}. \tag{4.6}
$$

With $\hat{S}(y^n) \triangleq \arg\max_{S \in \mathcal{S}, S \neq \emptyset} \hat{P}_S^\text{typ} \left(y^n\right) = \arg\min_{S \in \mathcal{S}, S \neq \emptyset} \left[ \sum_{i \in S} D\left(y_i \| \frac{\sum_{k \in S} y_k}{|S|}\right) + \sum_{j \notin S} D(y_j \| \pi) \right]$ 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:

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

$$\delta^* = \begin{cases} \hat{S}(Y^{N^*}) & \text{if } \tilde{N} \leq T \log T \\ 0 & \text{if } \tilde{N} > T \log T, \end{cases} \quad (4.8)$$

where

$$\tilde{N} \triangleq \arg\min_{n \geq 1} \begin{cases} \min_{S' \neq \hat{S}} n & \sum_{i \in S'} D \left( \gamma_i \left\| \sum_{k \in S'} \frac{\gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S'} D \left( \gamma_j \| \pi \right) \\ - \sum_{i \in \hat{S}} D \left( \gamma_i \left\| \sum_{k \in \hat{S}} \frac{\gamma_k}{|\hat{S}|} \right\| \right) - \sum_{j \notin \hat{S}} D \left( \gamma_j \| \pi \right) \end{cases} > \log T + (M + 1)|Y| \log(n + 1). \quad (4.9)$$

Similarly, when neither $\mu$ nor $\pi$ is known, the test can be written as in (4.7) and (4.8) but with $\hat{S}(y^n) \triangleq \arg\max_{S \subseteq S, S \neq \emptyset} \hat{p}^\text{inv}(y^n) = \arg\min_{S \subseteq S, S \neq \emptyset} \left[ \sum_{i \in S} D \left( \gamma_i \left\| \sum_{k \in S} \frac{\gamma_k}{|S|} \right\| \right) + \sum_{j \notin S} D \left( \gamma_j \left\| \sum_{k \notin S} \frac{\gamma_k}{|S|} \right\| \right) \right], and

$$\tilde{N} \triangleq \arg\min_{n \geq 1} \begin{cases} \min_{S' \neq \hat{S}} n & \sum_{i \in S'} D \left( \gamma_i \left\| \sum_{k \in S'} \frac{\gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S'} D \left( \gamma_j \left\| \sum_{k \notin S'} \frac{\gamma_k}{|S'|} \right\| \right) \\ - \sum_{i \in \hat{S}} D \left( \gamma_i \left\| \sum_{k \in \hat{S}} \frac{\gamma_k}{|\hat{S}|} \right\| \right) - \sum_{j \notin \hat{S}} D \left( \gamma_j \left\| \sum_{k \notin \hat{S}} \frac{\gamma_k}{|\hat{S}|} \right\| \right) \end{cases} > \log T + (M + 1)|Y| \log(n + 1). \quad (4.10)$$

### 4.2. Performance of proposed test

**Theorem 4.1.** When only $\pi$ is known, the test in (4.7), (4.8), and (4.9) is universally consistent and yields for every $T$ that

$$P_{\text{max}} \leq O \left( \frac{1}{T} \right). \quad (4.11)$$

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

$$\mathbb{E}_S[N^*] = \frac{\log T}{\alpha_S} (1 + o(1)) \quad (4.12)$$

$$\leq \begin{cases} - \frac{\log P_{\text{max}}}{D(\mu \| \pi)} (1 + o(1)), & |S| = K; \\ - \frac{\log P_{\text{max}}}{\min (D(\mu \| \pi), \eta_S(\mu \| \pi))} (1 + o(1)), & 1 \leq |S| < K, \end{cases} \quad (4.13)$$
where
\[
\alpha_S \triangleq \min_{S' \neq S, S' \neq \emptyset} \left[ |S \cap S'| D \left( \mu \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|} \right) + |S \setminus S'| D(\mu \| \pi) \right.
\]
\[
+ \left. |S' \setminus S| D \left( \pi \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|} \right) \right] > 0,
\]
(4.14)
and
\[
\eta_S(\mu \| \pi) \triangleq |S| D \left( \mu \left\| \frac{|S| \mu + \pi}{|S| + 1} \right) + D \left( \pi \left\| \frac{|S| \mu + \pi}{|S| + 1} \right) \right.,
\]
(4.15)

**Theorem 4.2.** When neither \( \mu \) nor \( \pi \) is known, the universal test in (4.7), (4.8), and (4.10) is universally consistent and yields for every \( T \) that
\[
P_{\max} \leq O \left( \frac{1}{T} \right).
\]
(4.16)

In addition, for each non-null hypothesis \( S \in \mathcal{S}, S \neq \emptyset \), as \( T \to \infty \),
\[
\mathbb{E}_S \left[ N^{*} \right] = \frac{\log T}{\alpha_S} (1 + o(1))
\]
(4.17)
\[
\leq \begin{cases} 
- \log P_{\max} + (1 + o(1)), & |S| = K; \\
- \log \frac{P_{\max}}{\eta(\mu \| \pi)} (1 + o(1)), & 1 \leq |S| < K,
\end{cases}
\]
(4.18)

where
\[
\alpha_S \triangleq \min_{S' \neq S, S' \neq \emptyset} \left[ |S \cap S'| D \left( \mu \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|} \right) + |S \setminus S'| D \left( \mu \left\| \frac{|S' \setminus S| \mu + |S \cap S'| \pi}{M - |S'|} \right) \right.
\]
\[
+ \left. |S' \setminus S| D \left( \pi \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|} \right) \right] > 0,
\]
(4.19)
and
\[
\eta_S(\mu \| \pi) \triangleq D \left( \mu \left\| \frac{\mu + (M - K - |S|) \pi}{M - K - |S| + 1} \right) \right.
\]
\[
+ (M - K - |S|) D \left( \pi \left\| \frac{\mu + (M - K - |S|) \pi}{M - K - |S| + 1} \right) \right.,
\]
(4.20)

and \( \eta_S(\mu \| \pi) \) is as in (4.15).

**Remark 4.1.** It follows from (4.20) and (2.6) that as \( M \to \infty \) (while \( K \) is kept fixed),
\[
\pi_S(\mu, \pi) \to D(\mu \| \pi);
\]
(4.21)
that is, the asymptotic performance for the test in (4.7), (4.8), and (4.10) when neither \( \mu \) nor \( \pi \) is known (cf. 4.18) converges to that for the test in (4.7), (4.8), and (4.9) when \( \pi \) is known (cf. 4.13) as \( M \to \infty \).
Remark 4.2. Similar to Remark 3.3, the results in Theorems 4.1 and 4.2 continue to hold when the deterministic time horizon \( T \log T \) in (4.7) 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 (4.9) and (4.10) on the right-side of the inequalities.

5. Extension to model with distinct outliers

It was shown in Li et al. (2014) that in the fixed sample size setting 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. We now describe this extension with distinctly distributed outliers but with their number being known in the sequential setting. 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.

In particular, for \( S \subset \{1, \ldots, M\} \), \(|S| = K\), denoting the set of \( K \) outliers, the joint distribution of all observations under the hypothesis with the outlier subset being \( S \) is

\[
p_S(y^n) = p_S(y_1, \ldots, y_n) = \prod_{k=1}^{n} \left\{ \prod_{i \in S} \mu_i \left( y_k^{(i)} \right) \prod_{j \not\in S} \pi \left( y_k^{(j)} \right) \right\},
\]

where each of the \( i \)th outliers, \( i \in S \), is distributed as \( \mu_i \), which can be arbitrarily distinct from one another as long as each \( \mu_i \neq \pi \). The test for the outliers is done based on a rule \( \delta(Y^n) \in S_K \), for an appropriate stopping time \( N \) and where \( S_K \) 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 (4.2) but with the maximum being over \( S_K \) instead.

As previously, for the setting with both the typical and outlier distributions being known and with \( \hat{S} \left( y^n \right) \hat{\Delta} \arg\max_{S \in S_K} p_S \left( y^n \right) \), the MSPRT has the stopping and final decision rules as in (4.3) and (4.4) but with the joint distribution \( p_S \left( y^n \right) \) being as in (5.1) instead of (4.1) and with the maximum in the denominator in (4.3) being over \( S_K \setminus \{\hat{S}\} \) instead. This MSPRT is asymptotically optimal (cf. Veeravalli and Baum, 1995; Dragalin et al., 1999).

**Proposition 5.1.** As the threshold \( T \) in (4.3) approaches infinity, the MSPRT in (4.3) and (4.4), with \( p_S \left( y^n \right) \) as in (5.1), \( \hat{S} \) being computed over \( S_K \), and the maximum in the denominator in (4.3) being over \( S_K \setminus \{\hat{S}\} \) satisfies

\[
P_{\text{max}} \leq O \left( \frac{1}{T} \right).
\]

In addition, for each \( S \in S_K \), as \( T \to \infty \),

\[
\mathbb{E}_S [N^*] = \frac{\log T(1 + o(1))}{\left( \min_{i \in S} D \left( \mu_i \| \pi \right) \right) + \left( \min_{j \not\in S} D \left( \pi \| \mu_j \right) \right)} = \frac{-\log P_{\text{max}}(1 + o(1))}{\left( \min_{i \in S} D \left( \mu_i \| \pi \right) \right) + \left( \min_{j \not\in S} D \left( \pi \| \mu_j \right) \right)}.
\]
Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests \((N, \delta)\) with vanishing maximal error probability, for each \(S \in \mathcal{S}_K\),

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

### 5.1. Proposed universal test

When only \(\pi\) is known, we can compute the corresponding GL of \(y^n\) under each hypothesis \(S \in \mathcal{S}_K\) by replacing the unknown \(\mu_i, i \in S\), in (5.1) with its ML estimate \(\hat{\mu}_i = y_i\). In particular, with \(\hat{S}(y^n) = \arg\min_{S \in \mathcal{S}_K} \sum_{j \notin S} D(y_j \| \pi)\) 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:

\[
N^* = \arg\min_{n \geq 1} \min_{S' \neq \hat{S}} \left( \sum_{j \notin S'} D(y_j \| \pi) - \sum_{j \notin \hat{S}} D(y_j \| \pi) \right) > \log T + (M + 1)|\mathcal{Y}| \log(n + 1);
\]

\[
\delta^* = \hat{S}(Y^{N^*}).
\]

Similarly, when neither \(\mu\) nor \(\pi\) is known, the test can be written as

\[
N^* = \arg\min_{n \geq 1} \min_{S' \neq \hat{S}} \left( \sum_{j \notin S'} D\left(y_j \| \frac{\sum_{k \notin S} y_k}{M - |S'|}\right) - \sum_{j \notin \hat{S}} D\left(y_j \| \frac{\sum_{k \notin \hat{S}} y_k}{M - |\hat{S}|}\right) \right) > \log T + (M + 1)|\mathcal{Y}| \log(n + 1);
\]

\[
\delta^* = \hat{S}(Y^{N^*}),
\]

but with \(\hat{S}(y^n) = \arg\min_{S \in \mathcal{S}} \sum_{j \notin S} D\left(y_j \| \frac{\sum_{k \notin S} y_k}{M - |S|}\right)\). 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 (4.7).

### 5.2. 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 5.1.** With the number of outliers \(K\) being known and when only \(\pi\) is known, the test in (5.2) and (5.3) is universally exponentially consistent and yields for every \(T\) that

\[
P_{\text{max}} \leq O\left(\frac{1}{T}\right).
\]
In addition, for each non-null hypothesis $S \in S_K$ as $T \to \infty$,

$$
E_S \left[ N^* \right] = \frac{\log T}{\min_{i \in S} D (\mu_i \| \pi)} (1 + o(1)) \leq -\frac{\log P_{\text{max}}}{\min_{i \in S} D (\mu_i \| \pi)} (1 + o(1)). 
$$

(5.6)

**Theorem 5.2.** With the number of outliers $K$ being known but neither $\mu$ nor $\pi$ being known, the test in (5.4) and (5.5) is universally exponentially consistent and yields for every $T$ that

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

In addition, for each non-null hypothesis $S \in S_K$ as $T \to \infty$,

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

(5.7)

**Remark 5.1.** As $M \to \infty$, the denominator of (5.7) converges to $\min_{i \in S} D (\mu_i \| \pi)$, which is the asymptotic performance of the universal test in (5.2) and (5.3) when $\pi$ is known (cf. (5.6)).

### 6. Numerical results

We now provide some numerical results for an example with $|\mathcal{Y}| = 4$. We compare the performance of the sequential test and the fixed sample size (FSS) GL test studied in Li et al. (2014). 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 1 are for the case where the underlying hypothesis has one outlier, and those in Figure 2 are for the case with two outliers. Depending on the nature of the test, the horizontal axis corresponds to the average stopping time for the sequential testing or the length of each sequence for the FSS test. In both figures, the vertical axis corresponds to the logarithm of the conditional error probabilities incurred by each test, conditioned on the underlying hypothesis. It follows from the result in Li et al. (2014) that there cannot exist a universally exponentially consistent FSS test with respect to $P_{\text{max}}$ 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 GL test and the sequential test are universally exponentially consistent. When comparing the FSS test to our sequential test, it is natural to compare the sample size of the FSS test to the expected stopping time under the 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 (4.18) is with respect to the maximum error probability, $P_{\text{max}}$, the same asymptote is also achievable when $P_{\text{max}}$ is replaced by a conditional error probability, simply because the latter is dominated by the former. 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 (3.20) is set to be $C = 15.05$. For the FSS test, the lengths 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. The result in (4.18) 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 1 and 2 corroborate such theoretical findings. It is also interesting to note that in both figures, there is a sharp 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. This drop in the conditional error probability can be explained by the fact that the sequential test is more adaptive than the FSS test.

7. Discussion

In practice, it is 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 (A.11), (A.14), and (A.15) in the proof of Theorem 3.1, we see that although arbitrarily small probabilities of $\mathbb{P}_i \{\delta^* \neq i, \delta^* \neq 0\}, i = 1, \ldots, M$, $\mathbb{P}_0 \{\delta^* \neq 0\}$ can be achieved with a suitably large $T$ universally for every $\mu$, the same cannot be achieved universally for the probability $\mathbb{P}_i \{\delta^* = 0\}, i = 1, \ldots, M$, unless we are given a lower bound on the distance between $\mu$ and $\pi$ (without having to know them precisely). 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 $\mu, \pi$ sufficiently close to each other that will incur a 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 $T$, which will be a function of $\mu, \pi$, and, hence, cannot be predetermined. Note that we are only working on one set of (test) data. Additional training data, when available separately from the test data, could be used to facilitate setting an appropriate threshold value.

8. Conclusions

We studied universal outlier hypothesis testing in the sequential setting. The goal was to design a universal test to best discern a few outlier sequences from multiple observation sequences. The observations in an outlier sequence were assumed to be distributed according to an outlier distribution, distinct from the common typical distribution governing the majority of the sequences. For the universal setting under consideration, except for being distinct, the outlier and typical distributions could be arbitrarily close. 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 when the outlier is present, we showed that the test is asymptotically optimal universally when the typical distribution is known. 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 showed 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. Lastly, we evaluated the performance of the proposed test numerically on a synthetic data set and compared its performance with that of the universal fixed sample size test that we studied in our previous work.

We end with a discussion of possible future work. First, the complexity of the implementation of the proposed tests scales exponentially with the number of outliers. When the number
of outliers is large, it is desirable to seek a more practical test that might sequentially pick out one outlier at a time instead of getting all the outliers at once as in the current test. Secondly, it is suggested by the numerical results in Figure 1 and 2 (and also from the results in Li et al., 2014) that although the proposed tests are shown to be asymptotically optimal in the limit of the large number of sequences and sequence length, there may exist other tests that outperform the proposed tests for small sequence length or small number of sequences. A challenging future research question is to discover such alternative tests and to systematically evaluate their performance in the regime of finite sequence length or finite number of sequences. Lastly, for certain data science applications such as cancer screening and telematics wherein each observation comprises a large number of components (also called dimensions or features), it would be interesting to understand the interplay between feature selection and outlier hypothesis testing. Numerous feature selection methods have been proposed for supervised and semi-supervised learning problems such as subset selection, shrinkage and cross-validation (Hastie et al., 2009). Nevertheless, the aforementioned techniques require the training features and labels and, hence, are not applicable to our universal setting for outlier hypothesis testing. It would be interesting to investigate how to perform such a feature selection in the context of universal outlier hypothesis testing.

Appendix A: Proof of Theorem 3.1

For each $i = 1, \ldots, M$, let

$$
\tilde{N}_i \triangleq \arg\min_{n \geq 1} \left\{ \min_{j \neq i} n \left( D(\gamma_i \| \pi) - D(\gamma_j \| \pi) \right) > \log T + M|\mathcal{Y}| \log(n + 1) \right\}. \quad (A.1)
$$

Our proofs rely on the following lemmas.

Lemma A.1. Under every non-null hypothesis $i = 1, \ldots, M$,

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

Proof. We get by the definition of $\tilde{N}_i$ in (A.1) that

$$
P_i\{\tilde{N}_i \geq n\} \\
\leq \sum_{j \neq i} P_i \left\{ (n - 1) \left[ D(\gamma_i^{(n-1)} \| \pi) - D(\gamma_j^{(n-1)} \| \pi) \right] \leq \log T + M|\mathcal{Y}| \log n \right\} \\
\leq \sum_{j \neq i} P_i \left\{ D(\gamma_i \| \mu) + D(\gamma_j \| \pi) \geq -\frac{1}{n-1} \left( \log T + M|\mathcal{Y}| \log n \right) + (D(\gamma_i \| \mu) + D(\gamma_j \| \pi)) \right\} \\
\leq \sum_{j \neq i} P_i \left\{ D(\gamma_i \| \mu) + D(\gamma_j \| \pi) \geq -\frac{1}{n-1} \left( \log T + M|\mathcal{Y}| \log n \right) + 2B(\mu, \pi) \right\}, \quad (A.2)
$$

where the last inequality follows from (2.1). Continuing from (A.2) by using (2.4) upon noting the independence of the $i$th and $j$th sequences and that the number of feasible empirical distributions $\gamma_i, \gamma_j$, each from sequences of length $(n - 1)$ are both upper bounded by $n|\mathcal{Y}|$
(cf. Section 2), we get that
\[
\mathbb{P}_i\{\tilde{N}_i \geq n\} \leq M \left( T n^M |\mathcal{Y}| \right) n^2 |\mathcal{Y}| e^{-(n-1)2B(\mu, \pi)} \\
\leq M T n^{(M+2)} |\mathcal{Y}| e^{-(n-1)2B(\mu, \pi)}.
\]
(A.3)

**Lemma A.2.** Under each non-null hypothesis \( i = 1, \ldots, M \),
\[
\lim_{T \to \infty} \mathbb{E}_i \left[ \frac{\tilde{N}_i}{\log T} - \frac{1}{D(\mu \| \pi)} \right] = 0.
\]
(A.4)

**Proof.** First observe that under hypothesis \( i = 1, \ldots, M \), we obtain by the strong law of large numbers that for every \( y \in \mathcal{Y} \), \( \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{\gamma^{(i)}_k = y\} \) 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 A.1 that \( \tilde{N}_i \) is finite a.s. under \( \mathbb{P}_i \); that is,
\[
\mathbb{P}_i\{\tilde{N}_i \geq n\} \to 0 \quad \text{as} \quad n \to \infty.
\]
(A.5)

It then follows from this a.s. finiteness and the definition of \( \tilde{N}_i \) in (A.1) that with probability 1 under \( \mathbb{P}_i \),
\[
\min_{j \neq i} \left( D\left( \gamma_i^{(N_i)} \| \pi \right) - D\left( \gamma_j^{(N_i)} \| \pi \right) \right) > \frac{\log T + M |\mathcal{Y}| \log(\tilde{N}_i + 1)}{\tilde{N}_i};
\]
(A.6)
\[
\min_{j \neq i} \left( D\left( \gamma_i^{(N_i-1)} \| \pi \right) - D\left( \gamma_j^{(N_i-1)} \| \pi \right) \right) \leq \frac{\log T + M |\mathcal{Y}| \log \tilde{N}_i}{\tilde{N}_i - 1}.
\]
(A.7)

Next, by observing that for any distribution \( q \), \( D(q \| \pi) \leq \log \left( \frac{1}{\min_{y} \mu(y)} \right) < \infty \) (\( \pi \) has a full support), we get from (A.6) that
\[
\mathbb{P}_i\{\tilde{N}_i \leq n\} \leq \mathbb{P}_i \left\{ \tilde{N}_i D\left( \gamma_i^{(N_i)} \| \pi \right) > \log T; \tilde{N}_i \leq n \right\} \\
\leq \mathbb{P}_i \left\{ n \log \left( \frac{1}{\min_{y} \mu(y)} \right) > \log T \right\} \\
= 0, \quad \text{for every} \quad n < \frac{\log T}{\log \left( \frac{1}{\min_{y} \mu(y)} \right)},
\]
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 \( \mu \) and \( \gamma_j^{(n)} \), \( j \neq i \), to \( \pi \) that under \( \mathbb{P}_i \),
\[
\min_{j \neq i} \left( D\left( \gamma_i^{(N_i)} \| \pi \right) - D\left( \gamma_j^{(N_i)} \| \pi \right) \right),
\]
\[
\min_{j \neq i} \left( D\left( \gamma_i^{(N_i-1)} \| \pi \right) - D\left( \gamma_j^{(N_i-1)} \| \pi \right) \right) \to D(\mu \| \pi),
\]
a.s., as $T \to \infty$. It now follows from this, (A.6) and (A.7) that under $\mathbb{P}_i$, $\tilde{N}_{i/\log T}$ 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 $\tilde{N}_{i/\log T}$ is uniformly integrable as $T \to \infty$. To this end, for any $v > 0$ sufficiently large, we upper bound the following quantity using Lemma A.1 as follows:

\[
\mathbb{E}_i \left[ \frac{\tilde{N}_i}{\log T} \mathbb{I}_{\left\{ \tilde{N}_{i/\log T} \geq v \right\}} \right] \leq \mathbb{E}_i \left[ \frac{\left( \tilde{N}_i - \lfloor v \log T \rfloor + v \log T \right)}{\log T} \mathbb{I}_{\left\{ \tilde{N}_i \geq \lfloor v \log T \rfloor \right\}} \right] \\
\leq \frac{1}{\log T} \mathbb{E}_i \left[ \left( \tilde{N}_i - \lfloor v \log T \rfloor \right) \mathbb{I}_{\left\{ \tilde{N}_i - \lfloor v \log T \rfloor \geq 0 \right\}} \right] \\
+ \frac{v \log T}{\log T} \mathbb{P}_i \left\{ \tilde{N}_i \geq \lfloor v \log T \rfloor \right\} \\
= \frac{1}{\log T} \sum_{\ell=1}^{\infty} \mathbb{P}_i \left\{ \tilde{N}_i \geq \lfloor v \log T \rfloor + \ell \right\} + v \mathbb{P}_i \left\{ \tilde{N}_i \geq \lfloor v \log T \rfloor \right\} \\
\leq \frac{MT}{\log T} \sum_{\ell=1}^{\infty} e^{-\left( \lfloor v \log T + \ell - 2 \right) 2B(\mu, \pi)} \left( \lfloor v \log T \rfloor + \ell \right)^{(M+2)|\mathcal{Y}|} \\
+ vMT e^{-\left( \lfloor v \log T - 2 \right) 2B(\mu, \pi)} \left( \lfloor v \log T \rfloor \right)^{(M+2)|\mathcal{Y}|}. \tag{A.8}
\]

Continuing from (A.8), it then follows that for all $T$ sufficiently large so that $\lfloor v \log T \rfloor \geq 1$,

\[
\mathbb{E}_i \left[ \frac{\tilde{N}_i}{\log T} \mathbb{I}_{\left\{ \tilde{N}_{i/\log T} \geq v \right\}} \right] \leq \frac{MT}{\log T} \left( 2\lfloor v \log T \rfloor \right)^{(M+2)|\mathcal{Y}|} e^{-2vB(\mu, \pi) \log T} \\
\times \left( e^{4B(\mu, \pi)} \sum_{\ell=1}^{\infty} e^{-2B(\mu, \pi) \ell} \ell^{(M+2)|\mathcal{Y}|} \right) \\
+ vMT \left( \lfloor v \log T \rfloor \right)^{(M+2)|\mathcal{Y}|} e^{-2vB(\mu, \pi) \log T} \times e^{4B(\mu, \pi)}, \tag{A.9}
\]

which vanishes as $T \to \infty$, for any $v > \frac{1}{2B(\mu, \pi)}$, thereby establishing the uniform integrability and, hence, (A.4). \hfill \Box

We start by proving (3.15). It follows from the description of the test in (3.9), (3.10), and (3.11) that for any $i, j = 1, \ldots, M$, $i \neq j$,

\[
\mathbb{P}_i \left\{ \delta^* = j \right\} \leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ N^* = \tilde{N} = n, \delta^* = j \right\} \\
\leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ n \left( D(\gamma_j || \pi) - D(\gamma_i || \pi) \right) > \log T + M|\mathcal{Y}| \log(n + 1) \right\}
\]
\[
\leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ nD(\gamma_i || \pi) > \log T + M|\mathcal{Y}| \log(n + 1) \right\}
\]

\[
\leq \frac{1}{T} \sum_{n=1}^{\infty} (n + 1)^{-(M-1)|\mathcal{Y}|} \quad (A.10)
\]

\[
\leq \frac{C'(|\mathcal{Y}|, M)}{T}, \quad (A.11)
\]

where (A.10) follows from (2.4) and the polynomial upper bound on the number of empirical distributions.

In addition, for each \(i = 1, \ldots, M\),

\[
\mathbb{P}_i \{ \delta^* = 0 \} = \mathbb{P}_i \left\{ \tilde{N} > T \log T \right\}
\]

\[
\leq \frac{\mathbb{E}_i[\tilde{N}]}{T \log T} \quad (A.12)
\]

\[
\leq \frac{\mathbb{E}_i[\tilde{N}_i]}{T \log T} \quad (A.13)
\]

\[
\leq \frac{C'(\mu, \pi, |\mathcal{Y}|, M)}{T}, \quad (A.14)
\]

where (A.12), (A.13), and (A.14) 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 A.2, respectively.

Next, it follows from the definition of \(\tilde{N}, N^*\) in (3.9), (3.10) that

\[
\mathbb{P}_0 \{ \delta^* \neq 0 \} = \mathbb{P}_0 \left\{ N^* = \tilde{N} \right\}
\]

\[
= \mathbb{P}_0 \left\{ \tilde{N} \leq T \log T \right\}
\]

\[
\leq \mathbb{P}_0 \left\{ \tilde{N} \text{ is finite} \right\}
\]

\[
= \sum_{n=1}^{\infty} \mathbb{P}_0 \left\{ \tilde{N} = n \right\}
\]

\[
\leq \sum_{n=1}^{\infty} \sum_{i=1}^{M} \mathbb{P}_0 \left\{ nD(\gamma_i || \pi) > \log T + M|\mathcal{Y}| \log(n + 1) \right\}
\]

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

\[
\leq \frac{C'(|\mathcal{Y}|, M)}{T}. \quad (A.15)
\]

The combination of (A.11), (A.14), and (A.15) constitutes (3.15).

The first equality in (3.16) now follows from that for each \(i = 1, \ldots, M\), the limit in probability of \(\frac{N^*_i}{\log T}\) (under \(\mathbb{P}_i\)) is the same as that of \(\frac{\tilde{N}_i}{\log T}\), which is \(\frac{1}{D(\mu || \pi)}\) (cf. A.4) by virtue
of the fact that (cf. A.11, A.14) for every $\epsilon > 0$, 
\[ \mathbb{P}_i \left\{ \frac{N^*}{\log T} - \frac{1}{D(\mu \parallel \pi)} > \epsilon, \delta = i \right\} + \mathbb{P}_i \{ \delta \neq i \} \]
\[ = \mathbb{P}_i \left\{ \frac{N^*}{\log T} - \frac{1}{D(\mu \parallel \pi)} > \epsilon, \delta = i \right\} + \mathbb{P}_i \{ \delta \neq i \} \]
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 A.2.

**Appendix B: Proof of Theorem 3.2**

For each $i = 1, \ldots, M$, let

\[ \tilde{N}_i \triangleq \arg\min_{n \geq 1} \left\{ \min_{j \neq i} n \left[ \sum_{k \neq j} D \left( \gamma_k \left\| \sum_{\ell \neq j} \gamma_{\ell} / M - 1 \right\| \right) - \sum_{k \neq i} D \left( \gamma_k \left\| \sum_{\ell \neq i} \gamma_{\ell} / M - 1 \right\| \right) \right] > \log T + M|\mathcal{Y}| \log(n + 1) \right\} . \quad (B.1) \]

Our proofs rely on the following lemmas.

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

\[ \mathbb{P}_i \{ \tilde{N}_i \geq n \} \leq (M^2 - 1) Tn^{2M|\mathcal{Y}|} e^{-(n-1)b}, \quad (B.2) \]

where $b$ is a function of $\mu, \pi$ that is always positive.

**Proof.** It follows from the definition of $\tilde{N}_i$ in (B.1) that

\[ \mathbb{P}_i \{ \tilde{N}_i \geq n \} \]
\[ \leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq j} D \left( \gamma_k \left\| \sum_{\ell \neq j} \gamma_{\ell} / M - 1 \right\| \right) - \sum_{k \neq i} D \left( \gamma_k \left\| \sum_{\ell \neq i} \gamma_{\ell} / M - 1 \right\| \right) \right\} \]
\[ \leq \log T + M|\mathcal{Y}| \log n \]
\[ = \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq j} D \left( \gamma_k \left\| \sum_{\ell \neq j} \gamma_{\ell} / M - 1 \right\| \right) \geq - \frac{1}{n - 1} (\log T + M|\mathcal{Y}| \log n) + \sum_{k \neq j} D \left( \gamma_k \left\| \sum_{\ell \neq j} \gamma_{\ell} / M - 1 \right\| \right) \right\} \]
\[ \leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D (\gamma_k \parallel \pi) \geq - \frac{1}{n - 1} (\log T + M|\mathcal{Y}| \log n) + \sum_{k \neq j} D \left( \gamma_k \left\| \sum_{\ell \neq j} \gamma_{\ell} / M - 1 \right\| \right) \right\} , \quad (B.3) \]

where (B.3) follows from the sum centroid inequality (2.5) with $\mathcal{C} = \{ \gamma_k | k = 1, \ldots, M, k \neq i \}$, and $q = \pi$. 
Continuing from (B.3), we have

\[
\mathbb{P}_i \{ \tilde{N}_i \geq n \} 
\leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D(\gamma_k \| \pi) \geq -\frac{1}{n-1} (\log T + M|\mathcal{Y}| \log n) + \sum_{k \neq j} D(\gamma_k \| \frac{\sum_{\ell \neq j} \gamma_{\ell}}{M-1}) \right\} 
\]

\[
D(\gamma_i \| \mu) \leq \epsilon, \text{ and } D(\gamma_j \| \pi) \leq \epsilon, \text{ for all } j \neq i 
\]

\[
+ \sum_{j \neq i} \mathbb{P}_i \left\{ D(\gamma_i \| \mu) > \epsilon, \text{ or } D(\gamma_j \| \pi) > \epsilon, \text{ for some } j \neq i \right\} 
\]

\[
\leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D(\gamma_k \| \pi) \geq -\frac{1}{n-1} (\log T + M|\mathcal{Y}| \log n) + \sum_{k \neq j} D(\gamma_k \| \frac{\sum_{\ell \neq j} \gamma_{\ell}}{M-1}) \right\} 
\]

\[
D(\gamma_i \| \mu) \leq \epsilon, \text{ and } D(\gamma_j \| \pi) \leq \epsilon, \text{ for all } j \neq i 
\]

\[
+ (M - 1) Mn|\mathcal{Y}| e^{-(n-1)\epsilon}, \quad \text{(B.4)} 
\]

where (B.4) is by (2.4). Note that \( \sum_{k \neq j} D\left( \gamma_k \| \frac{\sum_{\ell \neq j} \gamma_{\ell}}{M-1} \right) \) in (B.4) is zero only if for all \( k \neq j \), \( \gamma_k = \gamma \) for some \( \gamma \). This cannot happen if the \( \epsilon \) in (B.4) is chosen to be sufficiently small, because it also holds for the event in (B.4) that \( \mu \neq \pi \), and \( D(\gamma_i \| \mu) \leq \epsilon \), and for any \( j \neq i \) that \( D(\gamma_j \| \pi) \leq \epsilon \). We then conclude that when the \( \epsilon \) is chosen to be sufficiently small (as a function of \( \mu, \pi \)), it follows that \( \sum_{k \neq j} D\left( \gamma_k \| \frac{\sum_{\ell \neq j} \gamma_{\ell}}{M-1} \right) \geq a(\mu, \pi) > 0 \). Continuing from (B.4) with the \( \epsilon \) chosen sufficiently small and upon noting that the number of feasible empirical distributions \( \gamma_{(n)} \) for each \( k \neq i \) is upper bounded by \( n|\mathcal{Y}| \), we obtain

\[
\mathbb{P}_i \{ \tilde{N}_i \geq n \} \leq \sum_{j \neq i} \mathbb{P}_i \left\{ \sum_{k \neq i} D(\gamma_k \| \pi) \geq -\frac{1}{n-1} (\log T + M|\mathcal{Y}| \log n) + a(\mu, \pi) \right\} 
\]

\[
D(\gamma_i \| \mu) \leq \epsilon, \text{ and } D(\gamma_j \| \pi) \leq \epsilon, \text{ for all } j \neq i 
\]

\[
+ (M - 1) Mn|\mathcal{Y}| e^{-(n-1)\epsilon} 
\]

\[
\leq (M - 1) \left( Tn^M|\mathcal{Y}| e^{-(n-1)a(\mu, \pi)} \right) n^{(M-1)|\mathcal{Y}|} + (M - 1) Mn|\mathcal{Y}| e^{-(n-1)\epsilon} 
\]

\[
\leq (M^2 - 1) Tn^{2M|\mathcal{Y}|} e^{-(n-1) \min(a(\mu, \pi), \epsilon)} 
\]

\[
= (M^2 - 1) Tn^{2M|\mathcal{Y}|} e^{-(n-1)b}. \quad \text{(B.5)} 
\]

Lemma B.2. Under each non-null hypothesis \( i = 1, \ldots, M \),

\[
\lim_{T \to \infty} \mathbb{E}_i \left[ \left| \frac{\tilde{N}_i}{\log T} - \frac{1}{D(\mu \| \frac{1}{M-1} \mu + \frac{M-2}{M-1} \pi) + (M - 2)D(\pi \| \frac{1}{M-1} \mu + \frac{M-2}{M-1} \pi)} \right| \right] = 0. \quad \text{(B.6)} 
\]

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; Cover and Thomas, 2006) that under $\mathbb{P}_i$, 
\[
\min_{j \neq i} \left[ \sum_{k \neq j} D\left( Y_k^{(n)} \| \frac{\sum_{\ell \neq j} Y_{\ell}^{(n)}}{M - 1} \right) - \sum_{k \neq i} D\left( Y_k^{(n)} \| \frac{\sum_{\ell \neq i} Y_{\ell}^{(n)}}{M - 1} \right) \right]
\]
as.s. $D\left( \mu \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| + (M-2)D\left( \pi \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| \right)$, 
\begin{equation}
\text{(B.7)}
\end{equation}
as $n \to \infty$.

By Lemma B.1, 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 (B.1) that with probability 1 under $\mathbb{P}_i$, 
\[
\min_{j \neq i} \left[ \sum_{k \neq j} D\left( \tilde{Y}_k \| \frac{\sum_{\ell \neq j} \tilde{Y}_\ell}{M - 1} \right) - \sum_{k \neq i} D\left( \tilde{Y}_k \| \frac{\sum_{\ell \neq i} \tilde{Y}_\ell}{M - 1} \right) \right] > \frac{\log T + M|\mathcal{Y}| \log(\tilde{N}_i + 1)}{\tilde{N}_i}
\]
as $n \to \infty$.

The a.s. convergence of $\frac{\tilde{N}_i}{\log T}$ to $\frac{1}{D\left( \mu \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| + (M-2)D\left( \pi \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| \right)}$ under hypothesis $i$ will follow from (B.7), (B.8), and (B.9) by an argument based on sample paths similar to that in the proof of Lemma A.2 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 $j \neq i$, $\sum_{k \neq j} D\left( \tilde{Y}_k \| \frac{\sum_{\ell \neq j} \tilde{Y}_\ell}{M - 1} \right) \leq M \log (M - 1)$, and, hence, we get from (B.8) that for any $n \geq 1$ and any $j \neq i$, 
\[
\mathbb{P}_i \left\{ \tilde{N}_i \leq n \right\} \leq \mathbb{P}_i \left\{ \tilde{N}_i^{(n)} \left[ \sum_{k \neq j} D\left( \tilde{Y}_k \| \frac{\sum_{\ell \neq j} \tilde{Y}_\ell}{M - 1} \right) \right] > \log T; \tilde{N}_i \leq n \right\}
\]
\[
\leq \mathbb{P}_i \left\{ nM \log (M - 1) > \log T \right\}
\]
\[
= 0, \text{ for every } n < \frac{\log T}{M \log (M - 1)},
\]
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 $\frac{1}{D\left( \mu \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| + (M-2)D\left( \pi \left\| \frac{1}{M-1} \mu + \frac{M-2}{M} \pi \right\| \right)}$ under hypothesis $i$.

The main claim (B.6) follows by using the exponential tail bound for $\tilde{N}_i$ in Lemma B.1 to establish the uniform integrability of the sequence $\frac{\tilde{N}_i}{\log T}$ as in the argument leading to (A.9). \hfill \square
We start by proving (3.17). First, note that for any \( i, j = 1, \ldots, M, \ i \neq j \),

\[
\mathbb{P}_i \{ \delta^* = j \} \leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ N^* = \tilde{N} = n, \ \delta^* = j \right\}
\]

\[
\leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ n \left[ \sum_{k \neq i} D \left( \gamma_k \| \sum_{\ell \neq i, \ell \neq j} \gamma_{\ell} \right) - \sum_{k \neq j} D \left( \gamma_k \| \sum_{\ell \neq j} \gamma_{\ell} \right) \right] > \log T + M|\mathcal{Y}| \log(n+1) \right\}
\]

\[
\leq \sum_{n=1}^{\infty} \mathbb{P}_i \left\{ n \sum_{k \neq i} D \left( \gamma_k \| \pi \right) > \log T + M|\mathcal{Y}| \log(n+1) \right\}
\]

\[
\leq \frac{1}{T} \sum_{n=1}^{\infty} (n + 1)^{-|\mathcal{Y}|}
\]

\[
= \frac{C(|\mathcal{Y}|)}{T},
\]

(B.12)

where (B.10) follows from (2.5) and (B.11) follows from (2.4) and the polynomial upper bound on the number of empirical distributions.

In addition, for each \( i = 1, \ldots, M \), it follows from (B.6), the Markov inequality, and the fact that \( \tilde{N} \leq \bar{N}_i \) with probability 1 that

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

(B.13)

Next, it follows from the definitions of \( \bar{N}, N^* \) in (3.14), (3.12), and (2.5) that

\[
\mathbb{P}_0 \{ \delta^* \neq 0 \} = \mathbb{P}_0 \left\{ N^* = \bar{N} \right\} = \mathbb{P}_0 \left\{ \bar{N} \leq T \log T \right\}
\]

\[
\leq \mathbb{P}_0 \left\{ \bar{N} \text{ is finite} \right\} = \sum_{n=1}^{\infty} \mathbb{P}_0 \left\{ \bar{N} = n \right\} = \sum_{n=1}^{\infty}
\]

\[
\leq \sum_{n=1}^{\infty} \sum_{i=1}^{M} \mathbb{P}_0 \left\{ n \left[ \sum_{k \neq i} D \left( \gamma_k \| \sum_{\ell \neq i, \ell \neq j} \gamma_{\ell} \right) \right] > \log T + M|\mathcal{Y}| \log(n+1) \right\}
\]

\[
\leq \sum_{n=1}^{\infty} \sum_{i=1}^{M} \mathbb{P}_0 \left\{ n \sum_{k \neq i} D \left( \gamma_k \| \pi \right) > \log T + M|\mathcal{Y}| \log(n+1) \right\}
\]

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

(B.14)

The combination of (B.12), (B.13), and (B.14) constitutes (3.17).

The proof of (3.18) follows similar steps as in the proof of (3.16): first, under \( \mathbb{P}_i \), \( i = 1, \ldots, M \), the limit in probability of \( \frac{N^*}{\log T} \) is identical to that of \( \frac{\tilde{N}}{\log T} \) (cf. B.12, B.13);
and, second, the uniform integrability of $\frac{N^*}{\log T}$ follows from the uniform integrability of $\tilde{N}_i\log T$, which was already established, by virtue of the fact that for each $i = 1, \ldots, M$, $N^* \leq \tilde{N}_i$ with probability 1.

**Appendix C: Proof of Theorem 4.1**

For each $S \in \mathcal{S}, S \neq \emptyset$, let

$$
\tilde{N}_S \triangleq \arg\min_{n \geq 1} \left( \min_{S' \neq S, S' \neq \emptyset} n \left[ \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S'} \gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S} D \left( \gamma_j \| \pi \| \right) 
- \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S} \gamma_k}{|S|} \right\| \right) - \sum_{j \notin S} D \left( \gamma_j \| \pi \| \right) \right] > \log T + (M + 1)|\mathcal{Y}| \log (n + 1) \right) 
\right).
$$

(C.1)

Our proofs rely on the following lemmas.

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

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

(C.2)

where $b > 0$ is a function only of $\mu$ and $\pi$.

**Proof.** It follows by the definition of $\tilde{N}_S$ in (C.1) and (2.5) that

$$
\mathbb{P}_S[\tilde{N}_S \geq n] \leq \sum_{S' \neq S} \mathbb{P}_S \left\{ (n - 1) \left[ \sum_{i \in S'} D \left( \gamma_i \left\| \frac{\sum_{k \in S} \gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S'} D \left( \gamma_j \| \pi \| \right) 
- \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S} \gamma_k}{|S|} \right\| \right) - \sum_{j \notin S} D \left( \gamma_j \| \pi \| \right) \right] \leq \log T + (M + 1)|\mathcal{Y}| \log n \right) \right\}
$$

$$
\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S} \gamma_k}{|S|} \right\| \right) + \sum_{j \notin S} D \left( \gamma_j \| \pi \| \right) 
\geq - \frac{1}{n - 1} (\log T + (M + 1)|\mathcal{Y}| \log n) + \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S' \neq S} \gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S'} D \left( \gamma_j \| \pi \| \right) \right\}
$$

$$
\leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \sum_{i \in S} D \left( \gamma_i \| \mu \| \right) + \sum_{j \notin S} D \left( \gamma_j \| \pi \| \right) 
\geq - \frac{1}{n - 1} (\log T + (M + 1)|\mathcal{Y}| \log n) + \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S' \neq S} \gamma_k}{|S'|} \right\| \right) + \sum_{j \notin S'} D \left( \gamma_j \| \pi \| \right) \right\}
$$

(C.3)
Note that the term \( \sum_{i \in S} D (y_i || \mu) \) is zero only if \( y_i = \gamma \) for all \( i \in S' \) for some \( \gamma \) and \( y_j = \pi \) for all \( j \notin S' \). As in the event whose probability is concerned in (C.4), 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 \( \epsilon \) in (C.4) is chosen sufficiently small, since \( S' \neq S \). We conclude that when \( \epsilon \) is chosen to be sufficiently small (as a function of \( (\mu, \pi) \)), it holds therein that \( \sum_{i \in S} D (y_i || \mu) \) is zero. Continuing from (C.4) with the \( \epsilon \) chosen sufficiently small, we get that

\[
\mathbb{P}_S \{ \tilde{N}_S \leq n \} \leq \sum_{S' \neq S} \mathbb{P}_S \left\{ \sum_{i \in S} D (y_i || \mu) + \sum_{j \notin S} D (y_j || \pi) \geq -\frac{1}{n-1} \left( \log T + (M + 1)|\mathcal{Y}| \log n + a(\mu, \pi) \right); \right. \\
\left. \quad D(\gamma_i || \mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j || \pi) \leq \epsilon \text{ for every } j \notin S \right\} + M^{K+1} n|\mathcal{Y}| e^{-(n-1)\epsilon}.
\]

\[
\leq M^K Tn^{(2M+1)|\mathcal{Y}|} e^{-a(n-1)\epsilon} + M^{K+1} n|\mathcal{Y}| e^{-(n-1)\epsilon}.
\]

\[
\leq (M + 1) M^K Tn^{(2M+1)|\mathcal{Y}|} e^{-(n-1)\min(a, \epsilon)}.
\]

\[\text{(C.5)}\]

**Lemma C.2.** Under each non-null hypothesis \( S \in \mathcal{S}, S \neq \emptyset \),

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

where \( \alpha_S \) is defined in (4.14).

**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 \(P_S\),
\[
\sum_{i \in S'} D \left( y_i^{(n)} \left\| \frac{\sum_{k \in S'} y_k^{(n)}}{|S'|} \right\| \pi \right) + \sum_{j \notin S'} D \left( y_j^{(n)} \| \pi \right)
\]
\[ \xrightarrow{a.s.} |S \cap S'|D \left( \mu \left\| \frac{|S \cap S'|\mu + |S\setminus S'|\pi}{|S'|} \right\| + |S \setminus S'|D(\mu \| \pi) \right)
\]
\[ + |S \setminus S'|D \left( \pi \left\| \frac{|S \cap S'|\mu + |S\setminus S'|\pi}{|S'|} \right\| \right), \tag{C.7} \]
as \(n \to \infty\). Taking minimum over \(S' \neq S\) on both sides of (C.7), we see that under \(P_S\),
\[
\min_{S' \neq S} \sum_{i \in S'} D \left( y_i \left\| \frac{\sum_{k \in S'} y_k}{|S'|} \right\| \pi \right) + \sum_{j \notin S'} D \left( y_j \| \pi \right) \to \alpha_S \quad \text{a.s.,}
\]
as \(n \to \infty\), where \(\alpha_S\) is defined in (4.14).

By Lemma C.1, we see that \(\tilde{N}_S\) is finite a.s. under \(P_S\), \(S \in S, S \neq \emptyset\). It then follows from this a.s. finiteness and the definition of \(\tilde{N}_S\) in (C.1) that with probability 1 under \(P_S\),
\[
\min_{S' \neq S} \left[ \sum_{i \in S'} D \left( y_i \left\| \frac{\sum_{k \in S'} y_k}{|S'|} \right\| \pi \right) + \sum_{j \notin S'} D \left( y_j \| \pi \right) - \sum_{i \in S} D \left( y_i \left\| \frac{\sum_{k \in S} y_k}{|S|} \right\| \pi \right) - \sum_{j \notin S} D \left( y_j \left\| \pi \right\| \pi \right) \right] > \frac{\log T + (M + 1)|\mathcal{Y}| \log(\tilde{N}_S + 1)}{\tilde{N}_S}; \tag{C.8} \]
\[
\min_{S' \neq S} \left[ \sum_{i \in S'} D \left( y_i \left\| \frac{\sum_{k \in S'} y_k}{|S'|} \right\| \pi \right) + \sum_{j \notin S'} D \left( y_j \left\| \pi \right\| \pi \right) - \sum_{i \in S} D \left( y_i \left\| \frac{\sum_{k \in S} y_k}{|S|} \right\| \pi \right) - \sum_{j \notin S} D \left( y_j \left\| \pi \right\| \pi \right) \right] \leq \frac{\log T + (M + 1)|\mathcal{Y}| \log \tilde{N}_S}{\tilde{N}_S - 1}. \tag{C.9} \]
The claim in (C.6) now follows from (C.7), (C.8) and (C.9) if we can establish that under \(P_S\), \(\tilde{N}_S \to \infty\), 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 (C.8) that for any \(n \geq 1\), and any \(S' \neq S, S' \neq \emptyset\),
\[
P\{\tilde{N}_S \leq n\} \leq P_S \left\{ \tilde{N}_S \left[ \sum_{i \in S'} D \left( y_i \left\| \frac{\sum_{k \in S'} y_k}{|S'|} \right\| \pi \right) + \sum_{j \notin S'} D \left( y_j \left\| \pi \right\| \pi \right) \right] > \log T; \tilde{N}_S \leq n \right\} \]
\[
\leq \mathbb{P}_S\left\{ n \left( M \max \left( \log M, \log \left( \frac{1}{\min \pi(y)} \right) \right) \right) > \log T \right\} \\
= 0, \text{ for every } n < \frac{\log T}{M \max \left( \log M, \log \left( \frac{1}{\min \pi(y)} \right) \right)}, \tag{C.10}
\]

thereby yielding that \( \tilde{N}_S \to \infty \) a.s. as \( T \to \infty \). Using the exponential tail bound in Lemma C.1 to establish the uniform integrability of the sequence \( \frac{\tilde{N}_S}{\log T} \) similarly as in the previous proofs (details skipped), we obtain (C.6).

We now proceed to prove (4.11). First, note that for any \( S, S' \in S, S \neq S', S, S' \neq \emptyset \), we get from (2.5) that
\[
\mathbb{P}_S\{\delta^* = S'\} \\
\leq \sum_{n=1}^{\infty} \mathbb{P}_S\{N^* = n, \delta^* = S'\} \\
\leq \sum_{n=1}^{\infty} \mathbb{P}_S \left\{ n \left[ \sum_{i \in S} D\left( \gamma_i \left\| \sum_{k \in S} y_k \right/ |S| \right) + \sum_{j \notin S} D\left( \gamma_j \right) \pi \right] \right. \\
- \left. \sum_{i \in S'} \sum_{j \notin S'} D\left( \gamma_i \left\| \sum_{k \in S'} y_k \right/ |S'| \right) + \sum_{j \notin S'} D\left( \gamma_j \right) \pi \right\} > \log T + (M+1)|\mathcal{Y}| \log(n+1) \\
\leq \sum_{n=1}^{\infty} \mathbb{P}_S \left\{ n \left[ \sum_{i \in S} D\left( \gamma_i \left\| \sum_{k \in \mathcal{Y}} y_k \right/ |S| \right) + \sum_{j \notin S} D\left( \gamma_j \right) \pi \right] \\
> \log T + (M+1)|\mathcal{Y}| \log(n+1) \right\} \\
\leq \sum_{n=1}^{\infty} \frac{1}{n} \left( \log T + (M+1)|\mathcal{Y}| \log(n+1) \right) \\
\leq \frac{1}{T} C'(|\mathcal{Y}|). \tag{C.11}
\]

In addition, for each \( S \in S, S \neq \emptyset \), we obtain from (C.6), the Markov inequality, and that \( \tilde{N} \leq \tilde{N}_S \) with probability 1, that
\[
\mathbb{P}_S\{\delta^* = 0\} = \mathbb{P}_S\{\tilde{N} > T \log T\} \leq \frac{C'(\mu, \pi, |\mathcal{Y}|, M, K)}{T}. \tag{C.13}
\]

Next, it follows from the definitions of \( \tilde{N} \) and \( N^* \) in (4.9), (4.7), and (2.5) that
\[
\mathbb{P}_0\{\delta^* \neq 0\} = \mathbb{P}_0\{N^* = \tilde{N}\} = \mathbb{P}_0\{\tilde{N} \leq T \log T\} \\
\leq \mathbb{P}_0\{\tilde{N} \text{ is finite}\} = \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} = n\} = \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} \geq n\} = \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} \geq n\}.
\]
where (C.17) follows from (2.5) with combination of (C.16) and (C.17) constitutes the claim in (4.13).

It is now left to prove (4.13). First observe that when \(|S| = K\), it holds for any \(S' \in \mathcal{S}, S' \neq S, S' \neq \emptyset\), that \(|S\setminus S'| \geq 1\). Consequently, we get when \(|S| = K\) that

\[
\alpha_S = \min_{S' \neq \emptyset \atop S' \neq S} \left[ |S \cap S'| D \left( \mu \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|} \right) \right] + |S \setminus S'| D(\mu \| \pi)
\]

\[
\geq |S| D \left( \mu \left\| \frac{|S| \mu + |S' \setminus S| \pi}{|S'|} \right) \right) + D(\pi \left\| \frac{|S| \mu + |S' \setminus S| \pi}{|S'|} \right) \right)
\]

\[
geq \min_{\nu \in \mathcal{P}(\mathcal{Y})} |S| D(\mu \| \nu) + D(\pi \| \nu)
\]

\[
= |S| D \left( \mu \left\| \frac{|S| \mu + \pi}{|S| + 1} \right) \right) + D \left( \pi \left\| \frac{|S| \mu + \pi}{|S| + 1} \right) \right) = \eta_S(\mu \| \pi),
\]

where (C.17) follows from (2.5) with \(\mathcal{C}\) therein comprising \(|S|\) copies of \(\mu\) and one \(\pi\). The combination of (C.16) and (C.17) constitutes the claim in (4.13).

**Appendix D: Proof of Theorem 4.2**

For each \(S \in \mathcal{S}, S \neq \emptyset\), let

\[
\bar{N}_S \triangleq \arg\min_{n \geq 1} \left( \min_{S' \neq S \atop S' \neq \emptyset} \left[ \sum_{i \in S'} D \left( \gamma_i \left\| \frac{\sum_{k \in S'} y_k}{|S'|} \right) \right] + \sum_{j \not\in S} D \left( \gamma_j \left\| \frac{\sum_{k \not\in S} y_k}{M - |S|} \right) \right] \right) - \sum_{i \in S} D \left( \gamma_i \left\| \frac{\sum_{k \in S} y_k}{|S|} \right) \right] - \sum_{j \not\in S} D \left( \gamma_j \left\| \frac{\sum_{k \not\in S} y_k}{M - |S|} \right) \right] > \log T + (M + 1)|\mathcal{Y}| \log(n + 1).
\]

(C.14)
Our proofs rely on the following lemmas.

**Lemma D.1.** Under every non-null hypothesis $S \in \mathcal{S}, S \neq \emptyset$, and each $n \geq 1$,
\[
\mathbb{P}_S[\tilde{N}_S \geq n] \leq (M + 1)M^K\text{Tr}_{(2M+1)\mathcal{Y}} e^{-(n-1)d},
\]}
(D.2)
where $b > 0$ is a function only of $\mu$ and $\pi$.

**Proof.** We get from (2.5) and (D.1) that
\[
\mathbb{P}_S[\tilde{N}_S \geq n] \leq \sum_{S \neq \emptyset} \mathbb{P}_S \left\{(n - 1) \left[ \sum_{i \in S} D(\gamma_i \| \mu) + \sum_{j \not\in S} D(\gamma_j \| \pi) \right. \right.
\]
\[
\left. \left. - \sum_{i \in S} D\left(\gamma_i \left\| \sum_{k \in S'} \frac{Y_k}{|S'|} \right) \right) \right) \right) - \sum_{j \not\in S} D\left(\gamma_j \left\| \sum_{k \not\in S} \frac{Y_k}{M - |S|} \right) \right)
\]
\[
\leq \log T + (M + 1)|\mathcal{Y}| \log n
\]
(D.3)
\[
\mathbb{P}_S \left\{ \left[ \sum_{i \in S} D(\gamma_i \| \mu) + \sum_{j \not\in S} D(\gamma_j \| \pi) \right. \right.
\]
\[
\left. \left. \geq -\frac{1}{n - 1} \left( \log T + (M + 1)|\mathcal{Y}| \log n \right) \right) \right) + \sum_{i \not\in S} D\left(\gamma_i \left\| \sum_{k \in S'} \frac{Y_k}{|S'|} \right) \right) \right) + \sum_{j \not\in S} D\left(\gamma_j \left\| \sum_{k \not\in S} \frac{Y_k}{M - |S|} \right) \right)
\]
and
\[
D(\gamma_i \| \mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j \| \pi) \leq \epsilon \text{ for every } j \not\in S
\]
\[
+ \sum_{S \neq \emptyset} \mathbb{P}_S \left\{ D(\gamma_i \| \mu) > \epsilon \text{ for some } i \in S, \text{ or } D(\gamma_j \| \pi) > \epsilon \text{ for some } j \not\in S \right) \right)
\]
(D.4)
Similar to the previous proofs, since $S' \neq S, S' \neq \emptyset$, \sum_{i \in S} D(\gamma_i \left\| \sum_{k \in S'} \frac{Y_k}{|S'|} \right) + \sum_{j \not\in S} D(\gamma_j \left\| \sum_{k \not\in S} \frac{Y_k}{M - |S|} \right) is zero only if $\gamma_i = \gamma$ for all $i \in S$, for some $\gamma$ and $\gamma_j = \gamma'$ for
all \( j \notin S' \), for some \( \gamma' \). For sufficiently small \( \epsilon \), in the event whose probability is concerned in (D.4), 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 \( \epsilon \) is chosen to be sufficiently small (as a function of \( (\mu, \pi) \)), it holds that
\[
\sum_{i \in S} D(\gamma_i \| \mu) + \sum_{j \notin S} D(\gamma_j \| \pi) \geq \alpha(\mu, \pi) > 0.
\]
Continuing from (D.4) with \( \epsilon \) chosen sufficiently small, we get that
\[
\sum_{i \in S} D(\gamma_i \| \mu) + \sum_{j \notin S} D(\gamma_j \| \pi) \geq - \frac{1}{n-1} \left( \log T + (M+1)|\mathcal{Y}| \log n \right) + a(\mu, \pi) \text{ and } D(\gamma_i \| \mu) \leq \epsilon \text{ for every } i \in S, \text{ and } D(\gamma_j \| \pi) \leq \epsilon \text{ for every } j \notin S
\]
\[
+ M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon}
\]
\[
\leq M^K T n^{(2M+1)|\mathcal{Y}|} e^{-a(n-1)\epsilon} + M^{K+1} n^{|\mathcal{Y}|} e^{-(n-1)\epsilon}
\]
\[
\leq (M+1) M^K T n^{(2M+1)|\mathcal{Y}|} e^{-(n-1)\min(a, \epsilon)}.
\] (D.5)

\[\square\]

**Lemma D.2.** Under each non-null hypothesis \( S \in \mathcal{S}, S \neq \emptyset \),
\[
\lim_{T \rightarrow \infty} \mathbb{E}_S \left[ \left\lfloor \frac{\tilde{N}_S}{\log T} - \frac{1}{\tilde{\alpha}_S} \right\rfloor \right] = 0,
\] (D.6)

where \( \tilde{\alpha}_S \) is defined in (4.19).

**Proof.** Since under hypothesis \( S \in \mathcal{S}, S \neq \emptyset \), \( \gamma_i \rightarrow \mu \) a.s., for every \( i \in S \), and \( \gamma_j \rightarrow \pi \) a.s., for every \( j \notin S \), we get that under \( \mathbb{P}_S \),
\[
\sum_{i \in S} D(\gamma_i^{(n)} \| \frac{\sum_{k \in S} Y_k^{(n)}}{|S'|}) + \sum_{j \notin S} D(\gamma_j^{(n)} \| \frac{\sum_{k \notin S} Y_k^{(n)}}{M - |S'|})
\]
a.s. \( \rightarrow |S \cap S'| D \left( \mu \| \frac{|S \cap S'| \mu + |S\setminus S| \pi}{|S'|} \right)
\]
\[
+ |S\setminus S'| D \left( \mu \| \frac{|S\setminus S'| \mu + |S^c \cap S^c| \pi}}{M - |S'|} \right) + |S'| \setminus S |D \left( \pi \| \frac{|S' \setminus S| \mu + |S^c \setminus S^c| \pi}}{M - |S'|} \right),
\] (D.7)
as \( n \rightarrow \infty \). Taking minimum over \( S' \in \mathcal{S} \) on both sides of (D.7), we get that under \( \mathbb{P}_S \),
\[
\min_{S' \neq \emptyset} \sum_{i \in S} D(\gamma_i \| \frac{\sum_{k \in S} Y_k}{|S'|}) + \sum_{j \notin S} D(\gamma_j \| \frac{\sum_{k \notin S} Y_k}{M - |S'|}) \rightarrow \tilde{\alpha}_S \text{ a.s.}
\]
as \( n \rightarrow \infty \).
By Lemma D.1, we get that $\tilde{N}_S$ is finite a.s. under $\mathbb{P}_S, S \in S, S \neq \emptyset$. It then follows from this a.s. finiteness and the definition of $\tilde{N}_S$ in (D.1) that with probability 1 under $\mathbb{P}_S$,

$$\min_{S' \neq S} \left[ \sum_{i \in S'} D \left( y_i (\tilde{N}_S) \| \sum_{k \in S'} y_k (\tilde{N}_S) \| S' \right) \right] + \sum_{j \notin S'} D \left( y_j (\tilde{N}_S) \| \sum_{k \notin S'} y_k (\tilde{N}_S) \| M - |S'| \right)$$

$$- \sum_{i \in S} D \left( y_i (\tilde{N}_S) \| \sum_{k \in S} y_k (\tilde{N}_S) \| S \right) - \sum_{j \notin S} D \left( y_j (\tilde{N}_S) \| \sum_{k \notin S} y_k (\tilde{N}_S) \| M - |S| \right) \right]$$

$$> \frac{\log T + (M + 1)\log(\tilde{N}_S + 1)}{\tilde{N}_S}; \quad (D.8)$$

$$\min_{S' \neq S} \left[ \sum_{i \in S'} D \left( y_i (\tilde{N}_S^{-1}) \| \sum_{k \in S'} y_k (\tilde{N}_S^{-1}) \| S' \right) \right] + \sum_{j \notin S'} D \left( y_j (\tilde{N}_S^{-1}) \| \sum_{k \notin S'} y_k (\tilde{N}_S^{-1}) \| M - |S'| \right)$$

$$- \sum_{i \in S} D \left( y_i (\tilde{N}_S^{-1}) \| \sum_{k \in S} y_k (\tilde{N}_S^{-1}) \| S \right) - \sum_{j \notin S} D \left( y_j (\tilde{N}_S^{-1}) \| \sum_{k \notin S} y_k (\tilde{N}_S^{-1}) \| M - |S| \right) \right]$$

$$\leq \frac{\log T + (M + 1)\log \tilde{N}_S}{\tilde{N}_S - 1}. \quad (D.9)$$

The a.s. convergence of $\tilde{N}_S$ to $\frac{1}{\alpha_S}$ follows from (D.7), (D.8), and (D.9) if we can establish that under each hypothesis $S \in S$, $\tilde{N}_S \rightarrow \infty$, a.s. This can be established similarly as in the previous proofs upon noting for any $S' \neq S, S' \neq \emptyset$,

$$\sum_{i \in S'} D \left( y_i (\tilde{N}_S) \| \sum_{k \in S'} y_k (\tilde{N}_S) \| S' \right) + \sum_{j \notin S'} D \left( y_j (\tilde{N}_S) \| \sum_{k \notin S'} y_k (\tilde{N}_S) \| M - |S'| \right) \leq M \log M. \quad \square$$

The proof of (D.6) follows as previously from using Lemma D.1 to prove the uniform integrability of the sequence $\tilde{N}_S \log T$.

We now proceed to prove (4.16). First, note that for any $S, S' \in S, S \neq S', S, S' \neq \emptyset$, we get from (2.5) that

$$\mathbb{P}_S \{ \delta^* = S' \} \leq \sum_{n=1}^{\infty} \mathbb{P}_S \{ N^* = \tilde{N} = n, \delta^* = S' \}$$
Also, for each $S \in \mathcal{S}, S \neq \emptyset$, we get from (D.6), the Markov inequality, and that $\tilde{N} \leq \tilde{N}_S$ with probability 1 that

$$\mathbb{P}_S\{\delta^* = 0\} = \mathbb{P}_S\{\tilde{N} > T \log T\} \leq \frac{C'(\mu, \pi, |\mathcal{Y}|, M, K)}{T}. \quad (D.11)$$

Next, it follows from the definitions of $N^*, \tilde{N}$ in (4.7), (4.10), and (2.5) that

$$\mathbb{P}_0\{\delta^* \neq 0\} = \mathbb{P}_0\{N^* = \tilde{N}\} = \mathbb{P}_0\{\tilde{N} \leq T \log T\} \leq \mathbb{P}_0\{\tilde{N} \text{ is finite}\} \leq \sum_{n=1}^{\infty} \mathbb{P}_0\{\tilde{N} = n\}$$

$$\leq \sum_{n=1}^{\infty} \sum_{S} \mathbb{P}_0\left\{n \left[ \sum_{i \in S} D(\gamma_i | \mu) + \sum_{j \notin S} D(\gamma_j | \pi) \right] \geq \log T + (M + 1)|\mathcal{Y}| \log(n + 1) \right\}$$

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

$$\leq \frac{1}{T} M^K \sum_{n=1}^{\infty} (n + 1)^{-|\mathcal{Y}|} = \frac{1}{T} C'(|\mathcal{Y}|, M, K). \quad (D.12)$$

The combination of (D.10), (D.11) and (D.12) constitutes (4.16).

The claim in (4.17) follows as previously from (D.6), (D.10), (D.11) 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 (4.18). First observe that when $|S| = K$, it holds for any $S' \in \mathcal{S}, S' \neq S, S' \neq \emptyset$, that $|S \setminus S'| \geq 1$, and $|S^c \cap S'^c| \geq M - K - |S|$. It then follows that

$$\bar{\alpha}_S = \min_{S' \neq S \atop S' \neq \emptyset} \left[ |S \cap S'| D(\mu | [S \cap S'] | \mu + |S' \setminus S'| \pi) |S \setminus S'| D(\mu | [S \setminus S'] | \mu + |S' \setminus S'| \pi) \right. + \left. |S' \setminus S| D(\pi | [S' \setminus S'] | \mu + |S' \setminus S'| \pi) |S' \cap S'^c| D(\pi | [S' \cap S'^c] | \mu + |S' \cap S'^c| \pi) \right]$$

$$\quad + \left[ |S' \setminus S| D(\pi | [S' \setminus S'] | \mu + |S' \setminus S'| \pi) |S' \cap S'^c| D(\pi | [S' \cap S'^c] | \mu + |S' \cap S'^c| \pi) \right] \quad (D.13)$$
\[
\begin{align*}
\geq & \ D \left( \mu \left\| \frac{|S \cap S'| \mu + |S^c \cap S'| \pi}{M - |S'|} \right\| \right) + (M - K - |S|) D \left( \pi \left\| \frac{|S \cap S'| \mu + |S^c \cap S'| \pi}{M - |S'|} \right\| \right) \\
\geq & \ \min_{p \in \mathcal{P}(\Omega)} D(\mu\|p) + (M - K - |S|) D(\pi\|p) \\
= & \ D \left( \mu \left\| \frac{\mu + (M - K - |S|) \pi}{M - K - |S| + 1} \right\| \right) + (M - K - |S|) D \left( \pi \left\| \frac{\mu + (M - K - |S|) \pi}{M - K - |S| + 1} \right\| \right) \\
= & \ \overline{\eta}_S(\mu\|\pi),
\end{align*}
\]

where (D.14) follows from (2.5) with \( C \) therein comprising \( M - K - |S| \) copies of \( \pi \) and one \( \mu \).

We continue for the case with \( 1 \leq |S| < K \). For any \( S' \in S, S' \neq \emptyset \), such that \( S \setminus S' \neq \emptyset \), the term inside the minimum on the right-side of (D.13) is still lower bounded by \( \overline{\eta}_S(\mu\|\pi) \). Now, for any other \( S' \neq S \), such that \( S' \supset S \), we obtain that

\[
\begin{align*}
\left| S \cap S' \right| D \left( \mu \left\| \frac{|S \cap S'| \mu + |S^c \cap S'| \pi}{|S'|} \right\| \right) + |S \setminus S'| D(\mu\|p) & + \left| S \cap S' \right| D \left( \pi \left\| \frac{|S \cap S'| \mu + |S^c \cap S'| \pi}{|S'|} \right\| \right) \\
\geq & \ \min_{p \in \mathcal{P}(\Omega)} |S| D(\mu\|p) + D(\pi\|p) \\
= & \ |S| D \left( \mu \left\| \frac{|S \mu + \pi}{|S|} \right\| \right) + D \left( \pi \left\| \frac{|S \mu + \pi}{|S|} \right\| \right) = \eta_S(\mu\|\pi).
\end{align*}
\]

We conclude from (D.14), (D.15) that (4.18) holds.

**Funding**

This work was supported by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-10-1-0458 and by the National Science Foundation under grants CCF 11-11342 and CCF 16-18658, through the University of Illinois at Urbana–Champaign.

**References**

Baum, C. W. and Veeravalli, V. V. (1994). A Sequential Procedure for Multihypothesis Testing, *IEEE Transactions on Information Theory* 40: 1994–2007.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, New York: Wiley.

Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V. V. (1999). Multihypothesis Sequential Probability Ratio Tests—Part I: Asymptotic Optimality, *IEEE Transactions on Information Theory* 45: 2448–2461.

Dragalin, V. P., Tartakovsky, A. G., and Veeravalli, V. V. (2000). Multihypothesis Sequential Probability Ratio Tests—Part II: Accurate Asymptotic Expansions for the Expected Sample Size, *IEEE Transactions on Information Theory* 46: 1136–1383.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

Gutman, M. (1989). Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics, *IEEE Transactions on Information Theory* 35: 401–408.

Hastie, T., Tibshirani R., and Friedman, J. (2009). *Elements in Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Lai, T. L. (1988). Nearly Optimal Sequential Tests of Composite Hypotheses, *Annals of Statistics* 16: 856–886.
Li, Y., Nitinawarat, S., and Veeravalli, V. V. (2014). Universal Outlier Hypothesis Testing, *IEEE Transactions on Information Theory* 60: 4066–4082.

Li, Y., Nitinawarat, S., and Veeravalli, V. V. (2015). Universal Outlier Hypothesis Testing: Application to Anomaly Detection, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP’15)*, April 19–24, 2015, pp. 5595–5599, Brisbane, Australia.

Mosteller, F. (1948). A $k$-Sample Slippage Test for an Extreme Population, *Annals of Mathematical Statistics* 19: 58–65.

Pearson, K. (1911). On the Probability That Two Independent Distributions of Frequency Are Really Samples from the Same Population, *Biometrika* 8: 250–254.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer.

Veeravalli, V. V. and Baum, C. W. (1995). Asymptotic Efficiency of a Sequential Multihypothesis Test, *IEEE Transactions on Information Theory* 41: 1994–1997.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses, *Annals of Mathematical Statistics* 16: 117–186.

Wald, A. and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test, *Annals of Mathematical Statistics* 19: 326–339.

Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, Philadelphia: SIAM.

Zacks, S. (1971). *Theory of Statistical Inference*, New York: Wiley.

Ziv, J. (1988). On Classification with Empirically Observed Statistics and Universal Data Compression, *IEEE Transactions on Information Theory* 34: 278–286.