Universal Hypothesis Testing in the Learning-Limited Regime

Benjamin G. Kelly†, Thitidej Tularak†, Aaron B. Wagner†, and Pramod Viswanath‡
†School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14850
Email: bgk6,tt224,wagner@cornell.edu
‡Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, Urbana-Champaign, IL 61801
Email: pramodv@illinois.edu

Abstract—Given training sequences generated by two distinct, but unknown distributions sharing a common alphabet, we seek a classifier that can correctly decide whether a third test sequence is generated by the first or second distribution using only the training data. To model ‘limited learning’ we allow the alphabet size to grow and therefore probability distributions to change with the blocklength. We prove that a natural choice, namely a generalized likelihood ratio test, is universally consistent (has a probability of error tending to zero with the blocklength for all underlying distributions) when the alphabet size is sub-linear in the blocklength, but inconsistent for linear alphabet growth. For up-to quadratic alphabet growth, in a regime where linear in the blocklength, but inconsistent for linear alphabet for all underlying distributions) when the alphabet size is sub-

generalized likelihood ratio test, is universally consistent (has

with the blocklength. We prove that a natural choice, namely a
training data. To model ‘limited learning’ we allow the alphabet

to be related to a different topic two. We are given a third
sequence \(X\) and we are to perform a binary classification (i.e. a hypothesis test), to decide whether \(Z\) is related to topic one or topic two.

One model for this problem might be to suppose that \(X = X_1^n\) is a realization of a discrete memoryless source (DMS) emitting symbols with some fixed, but unknown, distribution \(p\) on a finite alphabet \(A\) (and similiary \(Y = Y_1^n\) is generated by a DMS with a different unknown distribution \(q\)). The problem is then to decide whether \(Z = Z_1^n\) was generated by distribution \(p\) or distribution \(q\), using only \(X\) and \(Y\). The typical information-theoretic approach is then to suppose that the blocklength, \(n\), increases so that we see longer realizations, and be satisfied by a classifier that performs well in the limit as \(n\) goes to infinity.

For certain scenarios this classical asymptotic is inappropriate. For example in natural language, if we take words as our base symbols, then \(X\) and \(Y\) are strings containing \(n\) words each generated i.i.d. according to \(p\) and \(q\). Studies of English text [1] however, suggest that 1) as the blocklength grows, so does the number of words we encounter, without bound (albeit slowly); and 2) English text tends to comprise a large number of words that occur \(\Theta(1)\) times. Yet in the traditional asymptotic, the alphabet size is necessarily fixed (and so all words will eventually appear), and the count of any word will increase without bound. Even if we model language with some fixed-order Markov chain, similar issues arise.

In this paper we investigate an alternative asymptotic, where the alphabet and underlying distributions generating the training data \(X\) and \(Y\) can vary with \(n\). We may then ask under what conditions on the distributions \(p_n\), \(p_{n}^\alpha\) and alphabet \(A_{n}\) is it possible to have universally consistent classification, i.e. a single classifier which asymptotically makes no errors for any pair of distributions on \(A_{n}\). Note that the present problem is not simply a modification of the classical Neyman-Pearson problem [2] to permit growing alphabets. Here both distributions \(p_n\) and \(q_n\) are unknown and the tester must make his or her inferences about \(Z\) only from the training data \(X\) and \(Y\). We show that the rate of growth of the alphabet as a function of the blocklength is of key importance.

We start by considering a rule based on the principle of maximum likelihood (ML). ML is often used in the absence of prior information and has been widely applied in previous information-theoretic studies of hypothesis testing in the guise of the generalized likelihood ratio test (GLRT) [3, 4]. Our analysis shows that if the alphabet size grows like \(\alpha(n)\), then the GLRT is universally consistent, but when the alphabet size grows linearly with \(n\) we show that the GLRT is no longer consistent, i.e. there exists distributions for which the GLRT rule routinely misclassifies.

The aforementioned troublesome distributions have symbol probabilities that are order \(n^{-1}\), and so it is natural to ask whether universal classification with such distributions is possible. We answer this question in the affirmative. In fact, for any \(0 < \alpha < 2\) we prove the consistency of a new simple test that can handle any pair of sources whose underlying symbol probabilities are order \(n^{-\alpha}\). We dub these sources \(\alpha\)-large-alphabet sources and we show that there are no universally consistent classifiers for \(\alpha\)-large-alphabet sources when \(\alpha \geq 2\).

I. INTRODUCTION

Suppose we are given two training sequences \(X\) and \(Y\) where \(X\) is known to be related to topic one and \(Y\) known to be related to a different topic two. We are given a third sequence \(Z\) and we are to perform a binary classification (i.e. a hypothesis test), to decide whether \(Z\) is related to topic one or topic two.

One model for this problem might be to suppose that \(X = X_1^n\) is a realization of a discrete memoryless source (DMS) emitting symbols with some fixed, but unknown, distribution \(p\) on a finite alphabet \(A\) (and similiary \(Y = Y_1^n\) is generated by a DMS with a different unknown distribution \(q\)). The problem is then to decide whether \(Z = Z_1^n\) was generated by distribution \(p\) or distribution \(q\), using only \(X\) and \(Y\). The typical information-theoretic approach is then to suppose that the blocklength, \(n\), increases so that we see longer realizations, and be satisfied by a classifier that performs well in the limit as \(n\) goes to infinity.

For certain scenarios this classical asymptotic is inappropriate. For example in natural language, if we take words as our base symbols, then \(X\) and \(Y\) are strings containing \(n\) words each generated i.i.d. according to \(p\) and \(q\). Studies of English text [1] however, suggest that 1) as the blocklength grows, so does the number of words we encounter, without bound (albeit slowly); and 2) English text tends to comprise a large number of words that occur \(\Theta(1)\) times. Yet in the traditional asymptotic, the alphabet size is necessarily fixed (and so all words will eventually appear), and the count of any word will increase without bound. Even if we model language with some fixed-order Markov chain, similar issues arise.

In this paper we investigate an alternative asymptotic, where the alphabet and underlying distributions generating the training data \(X\) and \(Y\) can vary with \(n\). We may then ask under what conditions on the distributions \(p_n\), \(p_{n}^\alpha\) and alphabet \(A_{n}\) is it possible to have universally consistent classification, i.e. a single classifier which asymptotically makes no errors for any pair of distributions on \(A_{n}\). Note that the present problem is not simply a modification of the classical Neyman-Pearson problem [2] to permit growing alphabets. Here both distributions \(p_n\) and \(q_n\) are unknown and the tester must make his or her inferences about \(Z\) only from the training data \(X\) and \(Y\). We show that the rate of growth of the alphabet as a function of the blocklength is of key importance.

We start by considering a rule based on the principle of maximum likelihood (ML). ML is often used in the absence of prior information and has been widely applied in previous information-theoretic studies of hypothesis testing in the guise of the generalized likelihood ratio test (GLRT) [3, 4]. Our analysis shows that if the alphabet size grows like \(\alpha(n)\), then the GLRT is universally consistent, but when the alphabet size grows linearly with \(n\) we show that the GLRT is no longer consistent, i.e. there exists distributions for which the GLRT rule routinely misclassifies.

The aforementioned troublesome distributions have symbol probabilities that are order \(n^{-1}\), and so it is natural to ask whether universal classification with such distributions is possible. We answer this question in the affirmative. In fact, for any \(0 < \alpha < 2\) we prove the consistency of a new simple test that can handle any pair of sources whose underlying symbol probabilities are order \(n^{-\alpha}\). We dub these sources \(\alpha\)-large-alphabet sources and we show that there are no universally consistent classifiers for \(\alpha\)-large-alphabet sources when \(\alpha \geq 2\).

II. NOTATION AND \(\alpha\)-LARGE-ALPHABET MODEL

Alphabets are denoted using caligraphic letters, e.g. \(A = \{a_1, \ldots, a_{|A|}\}\). The set \(A^n\) is the \(n\)-fold cartesian product of \(A\). Strings are denoted in bold face, e.g. \(x = x_1 \cdots x_n\) (usually length is clear from the context). \(1\{A\}\) is the indicator function.
for event $A$ and

$$N(a|x) = \sum_{i=1}^{n} 1\{x_i = a\}.$$  

We use $\Lambda_x$ to denote the empirical distribution or type of string $x$, i.e. $\Lambda_x = n^{-1} \sum \{N(a_1|\ldots|a_n)\}$. The set of all discrete distributions on alphabet $\mathcal{A}$ is denoted $\mathcal{P}(\mathcal{A})$. The set of all sequences of length $n$ with type $q$ is denoted $T_n$. The set of all type variables $Q \in \mathcal{P}(\mathcal{A})$, i.e. those for which $T_q \neq \emptyset$, is denoted $\mathcal{P}_n(\mathcal{A})$. For other information theoretic notations we use the standard definitions, see e.g. [5].

For triangular arrays, $X_{n,m}$, $1 \leq m \leq n$, $n \geq 1$, the notation $X^n$ refers to the rows of the array, i.e. $X^n = X_{n,1}, \ldots, X_{n,n}$. For any distribution $p$ on alphabet $\mathcal{A}$ define

$$\hat{p} = \min_{a \in \mathcal{A}} p(a) \quad \text{and} \quad \hat{p} = \max_{a \in \mathcal{A}} p(a).$$

**Definition 1:** Let $\{p_n, q_n\}$ be a sequence of pairs of distributions, with the $n$th having alphabet $\mathcal{A}_n$; if for each $n$

$$\frac{\tilde{c}}{n^a} \leq \min(\hat{p}_n, \hat{q}_n) \leq \frac{\hat{c}}{n^a}, \quad \text{(1)}$$

where $\hat{c}$ and $\tilde{c}$ are constants that are independent of $n$, then we say the sequence $\{p_n, q_n, \mathcal{A}_n\}$ is an $\alpha$-large-alphabet two-source.

**III. STATEMENT OF PROBLEM AND RELATED RESULTS**

For each $n$, let $X_{n,m}$, $1 \leq m \leq n$ be i.i.d. random variables with distribution $p_n$ and similarly let $Y_{n,m}$, $1 \leq m \leq n$ be i.i.d. with distribution $q_n$. We assume that $p_n$ and $q_n$ are unknown distributions with a common finite alphabet $\mathcal{A}_n$. We also assume that $p_n$ and $q_n$ satisfy

$$\liminf_{n \to \infty} \sum_{a \in \mathcal{A}_n} |p_n(a) - q_n(a)| > 0. \quad \text{(2)}$$

For each $n$ we observe independent realizations $X^n$ and $Y^n$, the $i$th rows of the corresponding triangular arrays. Given a third independent row $Z_{n,m}$, $1 \leq m \leq n$ generated i.i.d, we wish to test which of hypotheses

$$\mathcal{H}_0 : Z^n \sim p_n^n \quad \text{for all } n,$$

$$\mathcal{H}_1 : Z^n \sim q_n^n \quad \text{for all } n$$

is in effect. One may think of $X^n$ and $Y^n$ being training data and the problem is to determine whether $Z^n$ came from the unknown distribution $p_n$ or $q_n$. We refer to this general problem as the **triangular array hypothesis testing problem**.

Let $P_n = p_n^n \times q_n^n \times p_n^n$ and $Q_n = p_n^n \times q_n^n \times q_n^n$. We will be concerned with the following asymptotic properties.

**Definition 2 (Universal Consistency):** For a given sequence of alphabets $\{\mathcal{A}_n\}_{n=1}^{\infty}$ we say a sequence of tests $T_n : \mathcal{A}_n^\times \times \mathcal{A}_n^\times \times \mathcal{A}_n^\times \to \{0, 1\}$ is universally consistent if for every sequence of distributions $\{p_n, q_n\}$ on $\{\mathcal{A}_n\}$ satisfying condition (2),

$$P_n(T_n(X^n, Y^n, Z^n) = 0) \to 1$$

and

$$Q_n(T_n(X^n, Y^n, Z^n) = 1) \to 1 \quad \text{as } n \to \infty.$$

**Definition 3 (α-Universal Consistency):** For a given sequence of alphabets $\{\mathcal{A}_n\}_{n=1}^{\infty}$ with $|\mathcal{A}_n| = \Theta(n^a)$, we say a sequence of tests $T_n : \mathcal{A}_n^\times \times \mathcal{A}_n^\times \times \mathcal{A}_n^\times \to \{0, 1\}$ is α-universally consistent if for every sequence $\{p_n, q_n\}$ on $\{\mathcal{A}_n\}$ satisfying (1) and (2),

$$P_n(T_n(X^n, Y^n, Z^n) = 0) \to 1$$

and

$$Q_n(T_n(X^n, Y^n, Z^n) = 1) \to 1 \quad \text{as } n \to \infty.$$

**Note:** Implicit in both definitions of universal consistency is that the classifier knows the underlying alphabet, however the classifiers considered here do not make use of this knowledge.

**A. Generalized Likelihood Ratio Test**

A natural test for the triangular array hypothesis testing problem would be the following form of a generalized likelihood ratio test, based on the idea of maximum likelihood (ML). For each $n$, decide according to

$$\max_{\hat{p} \in \mathcal{H}_0} \frac{p_n^n(X^n)q_n^n(Y^n)p_n^n(Z^n)}{p_n^n(X^n)q_n^n(Y^n)q_n^n(Z^n)} \quad \text{and} \quad \max_{\hat{q} \in \mathcal{H}_1} \frac{p_n^n(X^n)q_n^n(Y^n)p_n^n(Z^n)}{p_n^n(X^n)q_n^n(Y^n)q_n^n(Z^n)}.$$  

Using the well known identity [5, Ch 1, Lemma 2.6]

$$p(x) = \exp(-n[D(\Lambda_x||p) + H(\Lambda_x)]) \quad \text{(3)}$$

and Lemma 1 (which follows) we see that the GLRT is equivalently decided according to

$$D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n)$$

and

$$D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n). \quad \text{(4)}$$

where $\hat{p}_n = (\Lambda_{X^n} + \Lambda_{Z^n})/2$ and $\hat{q}_n = (\Lambda_{Y^n} + \Lambda_{Z^n})/2$.

**Lemma 1:** For any three probability distributions $x, y$ and $z$ on a common alphabet $\mathcal{A}$

$$\min_{p, q \in \mathcal{P}(\mathcal{A})} D(x||p) + D(y||q) + D(z||p)$$

$$= D(x||\hat{p}) + D(z||\hat{p}),$$

where

$$\hat{p} = (x + z)/2.$$  

**Proof:** Choosing $q = y$ yields $D(y||q) = 0$. For the optimal $p$, the result follows from the parallelogram identity [5, Ex 1.3.19],

$$D(x||p) + D(z||p) = D(x||(x + z)/2) + D(z||(x + z)/2)$$

$$+ 2D((x + z)/2||p).$$

The GLRT developed above is closely related to the test considered by Gutman [4] (see also Ziv [3]). Gutman was concerned with the fixed distribution setting (i.e. $p_n = p, q_n = q$ for all $n$) and showed that if one requires the error probability
under hypothesis $\mathcal{H}_0$ to decay exponentially in $n$ with rate $\lambda > 0$, then the test
\[
T_n(X^n, Y^n, Z^n) = \begin{cases} 0 & \text{if } D(\Lambda_{X^n} || \hat{p}_n) + D(\Lambda_{Z^n} || \hat{p}_n) - \lambda + \rho_n < 0 \\ 1 & \text{otherwise}, \end{cases}
\]
where $\rho_n = O(n^{-1} \log n)$, is most powerful i.e. has smallest error probability under $\mathcal{H}_1$.

For hypothesis testing in the classical regime, the following result holds.

**Lemma 2**: Suppose that for all $n$, $p_n = p$ and $q_n = q$, and that $p$ and $q$ are distributions on a finite alphabet $\mathcal{A}$ satisfying $\sum_{a \in \mathcal{A}} |p(a) - q(a)| > 0$. Then the GLRT \[\mathcal{H}\] is universally consistent.

**Proof**: This result is a direct consequence of Theorem 1 which follows. We note that it may also be proven more directly using the strong law of large numbers and continuity arguments.

For determining consistency of the GLRT in the general triangular array problem it turns out the growth rate of the alphabet $\mathcal{A}_n$ is of critical interest. In the next two sections we examine the cases of sub-linear and linear growth respectively.

### IV. GLRT and Sub-linear Alphabet Growth

The following lemma allows us to prove a ‘weak law’ for empirical distributions when the alphabet grows sub-linearly with $n$.

**Lemma 3**: If $|\mathcal{A}_n| = o(n)$\[\mathcal{H}\] then
\[
n^{-1} \log |\mathcal{P}_n| \to 0 \text{ as } n \to \infty.
\]

**Proof Sketch**: Proceed by tightly bounding the exact number of types using \[\mathcal{H}\] Ch.2 §9.5.15, then takes logs, divide by $n$ and examine the limit as $n$ goes to infinity.

**Lemma 4 (Empirical Weak Law)**: Let $X_{n,m}, 1 \leq m \leq n$ be i.i.d. with distribution $p_n$ on alphabet $\mathcal{A}_n$. If $|\mathcal{A}_n| = o(n)$ then for any $\epsilon > 0$
\[
p_n^n(\Lambda_{X^n} || p_n) > \epsilon \leq e^{-n(\epsilon - \delta_n)},
\]
where $\delta_n(|\mathcal{A}_n|) \to 0$ as $n \to \infty$.

**Proof**: Omitted due to space constraints.

Our motivation for studying growing alphabets was to make it more difficult to ‘learn’ the distributions. From the above weak law we see that in the $|\mathcal{A}_n| = o(n)$ case, we can still learn distributions in some sense and it should be no surprise that the GLRT classifier is consistent in this regime.

**Theorem 1**: If $|\mathcal{A}_n| = o(n)$ then the GLRT \[\mathcal{H}\] is universally consistent.

**Proof**: Suppose hypothesis $\mathcal{H}_0$ is in effect. For all distributions $p$ and $q$ define
\[
F(p, q) = D(p || (p + q)/2) + D(q || (p + q)/2)
\]
and define the set
\[
\mathcal{D}_n = \{(x, z) : F(\Lambda_x, \Lambda_z) > \epsilon\}.
\]

By definition
\[
P_n((X^n, Z^n) \in \mathcal{D}_n) = \sum_{(x, z) \in \mathcal{D}_n} p_n^n(x)p_n^n(z) - \lambda + \rho_n < 0
\]
and
\[
\sum_{Q_x \in \mathcal{P}_n(A_x) \times T(Q_x), Q_z \in \mathcal{P}_n(A_z) : x \in T(Q_x), Q_z \in F(Q_x, Q_z) > 0}
\]

Using identity \[\mathcal{H}\] and the bound \[\mathcal{H}\] Ch 1, Lemma 2.5
\[
|T(Q_x)| \leq \exp(nH(Q_x)),
\]
it follows that
\[
\sum_{x \in T(Q_x)} \sum_{z \in T(Q_z)} p_n^n(x)p_n^n(z)
\]
\[
\leq \exp(-n[D(\Lambda_{X^n} || p_n) + D(\Lambda_{Z^n} || p_n)]).
\]
Further, as in the proof of Lemma 1 we have for all distributions $Q_X, Q_Z, p_n$
\[
D(\Lambda_{X^n} || p_n) + D(\Lambda_{Z^n} || p_n) \geq F(Q_X, Q_Z)
\]
and therefore
\[
P_n((X^n, Z^n) \in \mathcal{D}_n) \leq |\{\mathcal{P}(\mathcal{A}_n)\}|^2 e^{-nc}.
\]
By way of Lemma 3 and the hypothesis, this implies that for all $\epsilon > 0$
\[
P_n(\Lambda_{X^n} || \hat{p}_n) + D(\Lambda_{Z^n} || \hat{p}_n) > \epsilon \to 0 \text{ as } n \to \infty.
\]
It remains to show that for some $\delta > 0$
\[
\lim_{n \to \infty} P_n(D(\Lambda_{X^n} || \hat{q}_n) + D(\Lambda_{Z^n} || \hat{q}_n) > \delta) = 1.
\]

Chebyshev’s inequality tells us for any $\delta > 0$
\[
P_n([D(\Lambda_{X^n} || \hat{q}_n) - E[D(\Lambda_{X^n} || \hat{q}_n)]) > \delta])
\]
\[
\leq \frac{\text{Var}(D(\Lambda_{X^n} || \hat{q}_n))}{\delta^2}.
\]

The Efron-Stein inequality \[\mathcal{H}, \mathcal{H}\] and fact that the quantity $D(\Lambda_{X^n} || (\Lambda_x + \Lambda_z)/2)$ viewed as a real-valued function of the vector $(x, z) = (x_1, \ldots, x_n, z_1, \ldots, z_n)$ is coordinate-wise Lipschitz with constant $O(\log(n)/n)$ imply that this variance goes to zero. Thus it follows with probability tending to one, $D(\Lambda_{X^n} || \hat{q}_n) + D(\Lambda_{Z^n} || \hat{q}_n)$ ‘concentrates’ around $E[D(\Lambda_{X^n} || \hat{q}_n)] + E[D(\Lambda_{Z^n} || \hat{q}_n)]$. Recalling $D(p || q)$ is convex in the pair $(p, q)$, by Jensen’s inequality
\[
E[D(\Lambda_{X^n} || \hat{q}_n)] + E[D(\Lambda_{Z^n} || \hat{q}_n)]
\]
\[
\geq D(E[\Lambda_{X^n}] || E[\Lambda_{X^n}]) + D(E[\Lambda_{Z^n}] || E[\Lambda_{Z^n}])
\]
\[
= D(q_n || (p_n + q_n)/2) + D(p_n || (p_n + q_n)/2),
\]
and from \[\mathcal{H}\] and Pinsker’s inequality \[\mathcal{H}\] Ch. 1.3.17
\[
\lim_{n \to \infty} \inf \frac{1}{4 \log 2} \left( \sum_{a \in \mathcal{A}_n} |p_n(a) - q_n(a)| \right)^2
\]
\[
> 0.
\]
Thus for $n$ sufficiently large $D(\Lambda Y^n||\hat{q}_n) + D(\Lambda Z^n||\hat{q}_n)$ concentrates around a strictly positive quantity, which is enough to establish \( \mathcal{H}_0 \). Under hypothesis \( \mathcal{H}_1 \) the proof is similar. ■

V. GLRT AND LINEAR ALPHABET GROWTH

In this section we show that when the alphabet growth is linear the GLRT is not universally consistent.

**Theorem 2:** There exists a sequence of alphabets having linear growth for which \( \mathcal{H}(\alpha) \) is not universally consistent.

**Proof:** We let \( \mathcal{A}_n = \{1, \ldots, 9n\} \) and will show there exists a pair of sources for which the GLRT fails. Define distributions

\[
p_n(a) = \begin{cases} \frac{1}{2n} & \text{if } a \in \{1, \ldots, n\} \\ \frac{1}{10n} & \text{if } a \in \{n+1, \ldots, 9n\} \end{cases}
\]

and \( q_n(a) = \begin{cases} \frac{5}{4n} & \text{if } a \in \{1, \ldots, n/2\} \\ \frac{1}{4n} & \text{if } a \in \{n/2+1, \ldots, n\} \\ \frac{1}{12n} & \text{if } a \in \{n+1, \ldots, 9n\}. \end{cases} \)

For this source pair, it is possible to analytically compute limits of the form \( \mathbb{E}[D(\Lambda Y^n||\hat{p}_n)] \) in terms of mixtures of moments of functions of Poisson random variables. Carrying out this analysis and numerically evaluating the resulting limits, we see that under hypothesis \( \mathcal{H}_0 \),

\[
\lim_{n \to \infty} \mathbb{E}[D(\Lambda Y^n||\hat{p}_n) + D(\Lambda Z^n||\hat{p}_n)] = 1.085078578
\]

\[
\lim_{n \to \infty} \mathbb{E}[D(\Lambda Y^n||\hat{q}_n) + D(\Lambda Z^n||\hat{q}_n)] = 1.026320785
\]

whereas under hypothesis \( \mathcal{H}_1 \),

\[
\lim_{n \to \infty} \mathbb{E}[D(\Lambda Y^n||\hat{p}_n) + D(\Lambda Z^n||\hat{p}_n)] = 1.026320785
\]

\[
\lim_{n \to \infty} \mathbb{E}[D(\Lambda Y^n||\hat{q}_n) + D(\Lambda Z^n||\hat{q}_n)] = 0.772879166.
\]

From the Efron-Stein inequality and Lipschitz property stated in the proof of Theorem 1, the random variables concentrate around their respective means, which by the previous calculation are converging to the values above. It follows that under hypothesis \( \mathcal{H}_0 \), the test incorrectly declares \( \mathcal{H}_1 \).

We return to this inconsistency at the end of the next section. As we will show, the key to understanding why the GLRT fails is to first understand hypothesis testing with \( \alpha \)-large-alphabet sources and so we turn to this next.

VI. BEYOND LINEAR ALPHABET GROWTH

(\( \alpha \)-LARGE-ALPHABET SOURCES)

In the previous section we established that the GLRT is not universally consistent for sources with linear alphabet growth. The distributions we used to exhibit this have probabilities that are all order \( n^{-1} \); in this section we focus on the \( \alpha \)-large-alphabet sources, whose probabilities are \( \Theta(n^{-\alpha}) \) and whose alphabet size is \( \Theta(n^\alpha) \). We show that for all \( 0 < \alpha < 2 \) there exist \( \alpha \)-universally-consistent tests. As a converse we also show that when \( |\mathcal{A}_n| \) grows quadratically or faster, i.e. \( \alpha \geq 2 \), there are no universally consistent tests.

It turns out \( \alpha \)-large-alphabet sources can be handled with a simple test based on geometric considerations. Loosely speaking, the idea is that under hypothesis \( \mathcal{H}_0 \), \( \Lambda Z^\cdot \) is “closer” to \( \Lambda X^\cdot \) than it is to \( \Lambda Y^\cdot \), despite the fact that \( \|\Lambda X^n - p_n\|_1 \) need not tend to zero when \( \mathcal{A}_n \) grows linearly or faster.

**Theorem 3:** If \( 0 < \alpha < 2 \) then the test

\[
\|\Lambda Z^n - \Lambda X^n\|_{\mathcal{H}_0}^{\mathcal{H}_1} \leq \|\Lambda Z^n - \Lambda Y^n\|_{\mathcal{H}_0}^{\mathcal{H}_1}
\]

is \( \alpha \)-universally consistent.

**Proof Sketch:** For brevity let

\[
F = \|\Lambda Z^n - \Lambda X^n\|_{\mathcal{H}_0}^{\mathcal{H}_1} - \|\Lambda Z^n - \Lambda Y^n\|_{\mathcal{H}_0}^{\mathcal{H}_1}.
\]

Suppose \( \mathcal{H}_1 \) is in effect (a subscript on operators denotes this). By using linearity of expectation and the fact that \( N(a|X^n), N(a|Y^n) \), etc. are binomial random variables, we have

\[
\mathbb{E}_1[F] = \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a))
\]

where both \( \sum_{X^n} p_n^2(a) \) and \( \sum_{Y^n} q_n^2(a) \) are \( O(n^{-\alpha}) \). Therefore by the Cauchy Schwarz inequality

\[
\lim_{n \to \infty} \mathbb{E}_1[n^\alpha F] = \lim_{n \to \infty} n^\alpha \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2
\]

\[
\geq \lim_{n \to \infty} \frac{\hat{c}}{3} \|p_n - q_n\|_1^2,
\]

which is strictly positive by hypothesis. Invoking Lemma 5, \( \text{Var}_1(n^\alpha F) \to 0 \) and the result follows from Chebyshev’s inequality. The hypothesis \( \mathcal{H}_0 \) is handled analogously. ■

**Lemma 5:** For all \( 0 < \alpha < 2 \) and for \( i = 0, 1 \)

\[
\text{Var}_i(n^\alpha F) \to 0
\]

**Proof Sketch:** The result is proven by direct calculation of the variance of the random variable \( F \) using the \( \alpha \)-large-alphabet assumptions. ■

**Theorem 4:** For any \( \alpha \geq 2 \) there exists a sequence of alphabets for which there are no \( \alpha \)-universally consistent tests.

**Proof Sketch:** The proof uses a result of Le Cam [11, Ch.16 §4, Lem. 1], which expresses the minimum error probability when testing between two sets of measures in terms of \( L_1 \) distance between convex hulls of the two sets. By choosing a sequence of alphabets with growth rate \( \Theta(n^\alpha) \) and carefully choosing two sets of measures on these alphabets, which in some sense correspond to the testing problem under consideration, one can show that the best achievable error probability, \( P_e \), is bounded away from zero when \( \alpha \geq 2 \). However, for our carefully chosen sets of measures, the existence of an \( \alpha \)-universal test implies that \( P_e \to 0 \). Clearly one has a contradiction whenever \( \alpha \geq 2 \), and therefore there can be no \( \alpha \)-universal tests for \( \alpha \geq 2 \). (See [11, Th. 4] for a similar application of this kind of argument to a simple-versus composite-hypothesis testing problem.) ■

\(^2\) Sharper concentration results are available using martingale techniques, for example we can improve the concentration from the rate implied by Lemma 5 to \( \exp(-n^{1/3}) \).
A. GLRT versus \(L_2\)-norm test and Weighting

Some intuition behind the failing of the GLRT can be gleaned by introducing the following quantity

\[
\chi^2(p, q) = \sum_a \frac{(p(a) - q(a))^2}{p(a) + q(a)}.
\]

Note that \(\chi^2(p, q)\) is a kind of weighted (squared) \(L_2\) distance and for the distributions used in Theorem 2 one can show, using the series expansion of \(\log(1 + x)\), that under both \(\mathcal{H}_0\) and \(\mathcal{H}_1\),

\[
D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) \approx \ln(2)\chi^2(\Lambda_{X^n}, \Lambda_{Z^n})
\]

and

\[
D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n) \approx \ln(2)\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}).
\]

From the proof of Theorem 3 we know that the random variable

\[
\sum_a n^a(\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - \sum_a n^a(\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2
\]

concentrates around values which guarantee consistent detection: namely asymptotically \(\mathbb{E}_0[n^aF] = \mathbb{E}_1[n^aF] > 0\). Unlike our \(L_2\) test, which weights all terms equally (by \(n^a\)), the \(\chi^2\) test weights the terms in the first sum of (7) by \((\Lambda_{X^n}(a) + \Lambda_{Z^n}(a))^{-1}\) and those in the second sum \((\Lambda_{Y^n}(a) + \Lambda_{Z^n}(a))^{-1}\); clearly there is no guarantee that the inequality

\[
\mathbb{E}_0[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) - \chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})] < 0
\]

should hold for such weights.

When dealing with general sources, i.e. sources whose probabilities are not all of the same order, some kind of weighting is likely to be necessary. For example, one can imagine a pair of sources where a particular symbol has large probability (say 1/2) under both hypotheses and many other symbols which are “rare”; the central limit theorem implies the fluctuations in counts for this dominant symbol will be order \(\sqrt{n}\) potentially dominating the deviations in counts for the other more rare symbols. By weighting based on counts, as is done in the \(\chi^2\) test, these fluctuations are all placed on the same order. For examples like this our two-norm based test would likely fail, since it essentially relies on just the unweighted counts, but, on the other hand, the \(\chi^2\) weighting can be too severe when the rare symbols are \(\alpha\)-large alphabet with \(\alpha \geq 1\).

VII. SIMULATION

In Figure 1 we show the empirical performance (over 10000 trials) of the GLRT classifier (4) versus the two-norm classifier (6) for increasing \(n\) and a uniform prior on the two hypotheses \(\mathcal{H}_0\) and \(\mathcal{H}_1\). Test A refers to the distributions \(p_n, q_n\) appearing in the proof of Theorem 2 Test B is the same sequence \(p_n\) versus \(r_n = 1/(9n)\), the uniform distribution. We see that in Test A the average error probability of the GLRT classifier tends to 1/2, as predicted by Theorem 2. In Test B, even though the GLRT is consistent, the convergence of the performance of our two-norm classifier is much faster.

VIII. CONCLUSIONS AND FUTURE WORK

We have studied universal hypothesis testing when the underlying alphabets and distributions are permitted to change with the blocklength \(n\). We established consistency of the GLRT in the regime where the alphabet size, \(|A_n|\), grows like \(o(n)\), and inconsistency when \(|A_n|\) grows linearly with \(n\).

We introduced \(\alpha\)-large-alphabet sources and proposed a new test which is \(\alpha\)-universally consistent for all \(0 < \alpha \leq 2\) and showed that there are no such tests when \(\alpha \geq 2\).

In the real world setting we must deal with different quantities of training and test data and we plan to address this in future work. On the theoretical side it is desirable to know whether there are tests for the general-source triangular array hypothesis testing problem that are universally consistent for linear alphabet growth rates, or whether a converse along the lines of Theorem 4 can be proven.

REFERENCES