Classification of Homogeneous Data with Large Alphabets

Benjamin G. Kelly†, Aaron B. Wagner†, Thitidej Tularak†, and Pramod Viswanath‡

Abstract

Given training sequences generated by two distinct, but unknown, distributions sharing a common alphabet, we study the problem of determining whether a third test sequence was generated according to the first or second distribution using only the training data. To better model sources such as natural language, for which the underlying distributions are difficult to learn, we allow the alphabet size to grow and therefore the probability distributions to change with the blocklength. Our primary focus is the situation in which the underlying probabilities are all of the same order, and in this regime we give conditions on the alphabet growth rate and distributions guaranteeing the existence of universally consistent tests, i.e. tests having a probability of error tending to zero with the blocklength for any underlying distributions. We show that some commonly used statistical tests are universally consistent provided the alphabet is sub-linear but these tests are inconsistent for linear growth rates. We then propose a classifier that is universally consistent with up-to quadratic alphabet growth and that no classifier can handle the case in which the alphabet grows quadratically or faster. If the tester is given the underlying distributions in place of the training data, we prove that consistent testing is possible regardless of the growth of the underlying alphabet. Our results are then used to illuminate the problem of classifying arbitrary (i.e. non-homogeneous) distributions on growing alphabets.

I. Introduction

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

One model for this problem is to suppose that $\mathbf{X} = X_1^n$ is a realization of a discrete memoryless source (DMS) emitting symbols with some fixed, but unknown, distribution p on a finite alphabet $\mathcal{A}$ (and similarly $\mathbf{Y} = Y_1^n$ is generated by a DMS with a different unknown distribution q). The problem is then to decide whether $\mathbf{Z} = Z_1^n$ was generated by distribution p or distribution q, using only $\mathbf{X}$ and $\mathbf{Y}$. The classical information-theoretic approach is to let the blocklength, n, increase 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 $\mathbf{X}$ and $\mathbf{Y}$ are strings containing n words each generated 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*; and 2) English text tends to comprise a large number of words that occur $\Theta(1)$ times. Yet in the traditional asymptotic with a fixed and finite alphabet, the law of large numbers (LLN) applies, implying that all words will eventually appear and the count of any word will increase without bound. Notice that this observation precludes the use of the Zipf-Mandelbrot distribution, often used to model (ranked) word frequencies, because as the blocklength tends to infinity, a string generated according to this distribution would still be dominated by $\Theta(1)$ words appearing $\Theta(n)$ times. The presence of a LLN is roughly equivalent to being able to "learn" the underlying distributions from the data via the convergence of empirical distributions, and can itself be another reason to reject the asymptotic if such an assumption is unrealistic for the application. Note that if we model language with some fixed-order Markov chain, similar issues arise.

† School of Electrical and Computer Engineering, Cornell University. Ithaca, NY 14853. Email: {bgk6,wagner,tt224}@cornell.edu.

‡Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL 61801. Email pramodv@illinois.edu.

A. Contributions

We investigate the classification problem in an alternative asymptotic, where the (discrete) alphabet and underlying distributions generating the data can vary with n. To tackle the problem we formulate it as a sequence of composite[1] binary hypothesis testing problems and ask under what conditions on the distributions p_n, q_n and alphabet $\mathcal{A}_n$ is it possible to have *universally consistent* tests, i.e. a sequence of tests (one for each n) that asymptotically makes no error for any sequence of pairs of distributions on $\mathcal{A}_n$. Note that this problem is non-trivial because here, unlike in the classical asymptotic, the empirical distributions of the test and training data need not converge to the underlying distributions.

Our primary focus is the case in which the underlying distributions belong to the class of α-large-alphabet distributions, i.e. distributions whose underlying symbol probabilities are all order $n^{-\alpha}$ and alphabet size order is order n^{α} (see Def. 1, Sect II for a precise definition). For these sources we provide a simple test and prove that it is universally consistent when $0 \leq \alpha < 2$. We also show that universally consistent classification for these sources is impossible when $\alpha \geq 2$. We also prove that two commonly used tests from classical statistics, the chi-squared test and generalized likelihood ratio test (GLRT), are universally consistent for $0 \leq \alpha < 1$, but both tests fail when $\alpha = 1$.

Our study of α-large-alphabet sources offers insights into the hypothesis testing problem for inhomogeneous sources (i.e. non α-large-alphabet sources whose symbol probabilities are arbitrary) with growing alphabets. Firstly, our results show that universally consistent tests for up-to sub-linear alphabet growth exist. Secondly, our converse result implies that testing for arbitrary sources is not possible when the underlying alphabet grows quadratically or faster. Finally, we illustrate that a key problem in classifying inhomogeneous data concerns how to handle symbols whose probabilities are of different orders. The chi-squared test and GLRT employ a kind of normalization, which attempts to put the differences between the symbol counts in the data on the same scale. Yet, for α-large alphabet sources these differences are naturally on the same scale and we show that this normalization can cause a systematic inconsistency. Our new test relies solely on the unnormalized counts, and we show that for inhomogeneous data our test is inconsistent precisely due to its lack of normalization.

We conclude by proving that when given an infinite amount of training data (i.e. the classifier exactly knows the underlying distributions p_n and q_n) consistent testing is possible for any rate of alphabet growth; we also provide an achievable error exponent for this problem.

B. Related Work

The case of hypothesis testing between fixed distributions on a finite alphabet has been well studied. For this simple-versus-simple case, a fundamental result on the existence of optimum tests is due to Neyman and Pearson, [2]; Chernoff [3], [4] also provides exponential error guarantees. For the simple-versus-composite case, a key result concerning the problem of asymptotically optimum tests (in an error exponent sense) is Hoeffding [5].

The composite-versus-composite case with fixed distributions on finite alphabets has also received some attention. The problem of determining a test with a prescribed exponential error decay under one hypothesis and that is uniformly most powerful under the other is considered by Gutman [6] (see also Ziv [7]). Feder and Merhav [8] propose a "competitive minimax" approach, in which one minimizes the worst case ratio between the probability of error of a universal test and the minimum probability of error attainable when the distributions are known.

For the case of growing alphabets, the existence of tests for the simple-versus-composite problem is studied by Barron [9], Paninski [10] and Ermakov [11]. The works [9], [10] also address the converse problem of determining the the smallest growth rate beyond which (respectively) uniformly exponentially consistent and consistent tests do not exist.

[1]Using the nomenclature from statistics, a hypothesis is *simple* if the distribution is fully known and otherwise we say the hypothesis is *composite*.

An alternate line of investigation into the simple-versus-composite case with growing alphabets studied the Pitman and Bahadur efficiencies of the likelihood and chi-square tests [12], [13]. Moderate and large deviation results for these statistics in the same regime are also available [14]. In [15, Ch.4 §3] Read and Cressie study the power divergence family with growing alphabets, which includes the chi-square and likelihood tests as members; the Bahadur efficiency of this family with growing alphabets is investigated in [16].

The composite-versus-composite case with growing alphabets is addressed in limited form by Wagner et al. [17], who develop a probability estimator for the "rare-events" regime where underlying probabilities are all order $\Theta(n^{-1})$ and therefore alphabet size is order $\Theta(n)$. Other practical approaches may also be taken, see for example Orlitsky-Santhanam-Zhang (OSZ) [18], [19], support vector machines [20], and techniques from pattern recognition and machine learning [21].

C. Outline

The remainder of the paper is organized as follows. In Section II we give definitions and formally state the problem. Section III includes our main results on hypothesis testing for α-large-alphabet sources; we state our test, study its performance, and derive a converse result on the maximum permissible growth rate of the alphabet. In Section IV we study the performance of the GLRT and chi-squared test for α-large-alphabet sources. Section V studies the hypothesis testing when the tester is given an "infinite" amount of training data (i.e. given access to the underlying distributions). Section VI concludes with discussion on the problem of classifying inhomogeneous sources and gives some suggestions for extensions. Proofs of ancillary technical results are deferred to the appendices.

II. Definitions and Problem Statement

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

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

We use $\Lambda_{\mathbf{x}}$ to denote the empirical distribution or *type* of string $\mathbf{x}$, i.e.

$$\Lambda_{\mathbf{x}} = n^{-1}\big[N(a_1|\mathbf{x}) \cdots N(a_{|\mathcal{A}|}|\mathbf{x})\big].$$

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_Q^n (again we usually omit n since it is clear from the context). The set of all type variables $Q \in \mathcal{P}(\mathcal{A})$, i.e. those for which $T_Q^n \neq \emptyset$, is denoted $\mathcal{P}^n(\mathcal{A})$. For other information theoretic notations we use the standard definitions, see e.g. [22]. If p is a distribution on $\mathcal{A}$ then p^n is the n-fold i.i.d. product measure on $\mathcal{A}^{\times n}$, i.e.

$$p^n(\mathbf{x}) = \prod_{i=1}^{n} p(x_i).$$

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. $X^n = X_{n,1}, \ldots, X_{n,n}$. We use $\|\cdot\|_p$ to denote the pth Euclidean norm and $\langle\cdot\rangle$ to denote the standard inner product.

For any distribution p on a finite set $\mathcal{A}$, $\operatorname{supp}(p)$ denotes its support and we define

$$\check{p} = \min_{a \in (\mathcal{A} \cap \operatorname{supp}(p))} p(a) \text{ and } \hat{p} = \max_{a \in \mathcal{A}} p(a).$$

Our primary focus in the paper will be the following class of distributions.

Definition 1. *The sequence $\{p_n, q_n, \mathcal{A}_n\}$ is an α-large-alphabet source pair if for all n*

$$\frac{\check{c}}{n^\alpha} \leq \min(\check{p}_n, \check{q}_n) \leq \max(\hat{p}_n, \hat{q}_n) \leq \frac{\hat{c}}{n^\alpha}, \tag{1}$$

where $\check{c}$ and $\hat{c}$ are positive constants independent of n; and where

$$\mathcal{A}_n = \mathcal{A}'_n \cup \mathcal{X}_n \cup \mathcal{Y}_n$$

with

$$\begin{aligned} \mathcal{A}'_n &= \operatorname{supp}(p_n) \cap \operatorname{supp}(q_n) \\ \mathcal{X}_n &= \operatorname{supp}(p_n) \cap \{a : q_n(a) = 0\} \\ \textit{and } \mathcal{Y}_n &= \operatorname{supp}(q_n) \cap \{a : p_n(a) = 0\}. \end{aligned}$$

Note that for any α-large-alphabet source, $|\mathcal{A}_n| = \Theta(n^\alpha)$. This can easily be seen since

$$1 \geq \sum_{a \in \mathcal{A}'_n} p_n(a) \geq |\mathcal{A}'_n| \frac{\check{c}}{n^\alpha} \text{ and } 1 \leq |\mathcal{A}_n| \frac{\hat{c}}{n^\alpha}$$

which along with $1 \geq |\mathcal{X}_n| \frac{\check{c}}{n^\alpha}$ and $1 \geq |\mathcal{Y}_n| \frac{\check{c}}{n^\alpha}$ implies

$$\frac{3n^\alpha}{\check{c}} \geq |\mathcal{A}_n| \geq \frac{n^\alpha}{\hat{c}}.$$

Such distributions may arise from sampling a probability density. For example, suppose $f(x)$ is (almost everywhere) continuous on $[0, 1]$ satisfying $\int f(x)dx = 1$ and $\check{c} \leq f(x) \leq \hat{c}$. If X is a random variable with density f and we define p_n as the distribution of $\lceil n^\alpha X \rceil$, then the sequence $\{p_n\}$ is α-large-alphabet with alphabet $\{1, \ldots, n^\alpha\}$. As we will see later studying this class sheds light on the general classification problem.

A. Problem Statement

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

$$\liminf_{n\to\infty} \|p_n - q_n\|_1 = \liminf_{n\to\infty} \sum_{a \in \mathcal{A}_n} |p_n(a) - q_n(a)| > 0. \tag{2}$$

For each n we observe independent realizations X^n and Y^n, the nth 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

$$\begin{aligned} \mathcal{H}_0 &: Z^n \sim p_n^n \text{ for all } n, \\ \text{or } \mathcal{H}_1 &: Z^n \sim q_n^n \text{ for all } n \end{aligned}$$

is in effect. One may think of X^n and Y^n as 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 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 of tests.

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

$$\begin{aligned} P_n(T_n(X^n, Y^n, Z^n) = 0) &\to 1 \\ \textit{and } Q_n(T_n(X^n, Y^n, Z^n) = 1) &\to 1 \textit{ as } n \to \infty. \end{aligned}$$

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

$$P_n(T_n(X^n, Y^n, Z^n) = 0) \to 1$$
$$\text{and } Q_n(T_n(X^n, Y^n, Z^n) = 1) \to 1 \text{ as } n \to \infty.$$

Note: Implicit in both definitions of universal consistency is that the classifier knows the underlying alphabet, however the classifiers considered in this work do not require knowledge of the symbols that do not appear in the training data. When proving impossibility results, however, we assume the classifier knows the alphabet.

III. Testing of α-large-alphabet sources

A. Achievability

In this subsection we show that α-large-alphabet sources can be handled with a simple test based on Euclidean geometric considerations. Loosely speaking, the idea is that under hypothesis $\mathcal{H}_0$, Λ_{Z^n} should be "closer" to Λ_{X^n} than it is to Λ_{Y^n}, despite the fact that $\|\Lambda_{X^n} - p_n\|_1$ need not tend to zero when $|\mathcal{A}_n|$ grows linearly or faster [23].

Theorem 1. *If $0 \le \alpha < 2$ then the test*

$$\|\Lambda_{Z^n} - \Lambda_{X^n}\|_2^2 \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \|\Lambda_{Z^n} - \Lambda_{Y^n}\|_2^2 \tag{3}$$

is α-universally consistent.

To prove the result we need the following lemmas. Throughout we define

$$F = F(X^n, Y^n, Z^n) = \|\Lambda_{Z^n} - \Lambda_{X^n}\|_2^2 - \|\Lambda_{Z^n} - \Lambda_{Y^n}\|_2^2.$$

Lemma 1.

$$\mathbb{E}_0[F] = \sum_{a \in \mathcal{A}_n} -(p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a))$$
$$\text{and } \mathbb{E}_1[F] = \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a))$$

Proof: Using $\mathbb{E}_i$ to denote expectation under $\mathcal{H}_i$, we now compute

$$\mathbb{E}_i[F(X^n, Y^n, Z^n)] = \mathbb{E}_i[\|\Lambda_{X^n}\|_2^2 - \|\Lambda_{Y^n}\|_2^2 - 2\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle].$$

We start with the two-norm of the type

$$\mathbb{E}_i\Big[\|\Lambda_{X^n}\|_2^2\Big] = n^{-2} \sum_{a \in \mathcal{A}_n} \mathbb{E}_i\big[N^2(a|X^n)\big].$$

Since $N(a|X^n)$ is a binomial random variable with parameters $(n, p_n(a))$,

$$\begin{aligned}\mathbb{E}_i\Big[\|\Lambda_{X^n}\|_2^2\Big] &= n^{-2} \sum_{a \in \mathcal{A}_n} np_n(a)(1 - p_n(a)) + n^2 p_n^2(a) \\ &= n^{-1} + \sum_{a \in \mathcal{A}_n} p_n^2(a) - n^{-1} p_n^2(a)\end{aligned}$$

Similarly

$$\mathbb{E}_i\big[\|\Lambda_{Y^n}\|_2^2\big] = n^{-1} + \sum_{a\in\mathcal{A}_n} q_n^2(a) - n^{-1}q_n^2(a).$$

For the final term

$$\begin{aligned}
&\mathbb{E}_i\big[\langle \Lambda_{Z^n}, (\Lambda_{X^n} - \Lambda_{Y^n})\rangle\big] \\
&= n^{-2}\sum_{a\in\mathcal{A}_n} \mathbb{E}_i\Big[N(a|Z^n)(N(a|X^n) - N(a|Y^n))\Big] \\
&= n^{-1}\sum_{a\in\mathcal{A}_n} \mathbb{E}_i\big[N(a|Z^n)](p_n(a) - q_n(a)).
\end{aligned}$$

Under hypothesis $\mathcal{H}_0$, the previous line is

$$\sum_{a\in\mathcal{A}_n} p_n(a)^2 - p_n(a)q_n(a)$$

and under hypothesis $\mathcal{H}_1$ is

$$\sum_{a\in\mathcal{A}_n} -q_n(a)^2 + p_n(a)q_n(a).$$

Therefore

$$\begin{aligned}
\mathbb{E}_0[F] &= \sum_{a\in\mathcal{A}_n} p_n^2(a) - n^{-1}p_n^2(a) - q_n^2(a) + n^{-1}q_n^2(a) - n^{-1}p_n^2(a) + 2p_n(a)q_n(a) \\
&= \sum_{a\in\mathcal{A}_n} -(p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a)),
\end{aligned}$$

and similarly

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

■

Lemma 2. *For all* $0 < \alpha < 2$ *and for* $i = 0, 1$

$$\mathrm{Var}_i[n^\alpha F] \to 0$$

Proof: Follows from direct calculation using binomial moments. See Appendix A for details. ■

Lemma 3. *For any* α*-large-alphabet source pair* $\{p_n, q_n, \mathcal{A}_n\}$

$$\check{c}/3\|p_n - q_n\|_1^2 \le n^\alpha \|p_n - q_n\|_2^2$$

Proof: The result follows from the Cauchy-Schwarz inequality and the bound $|\mathcal{A}_n| \le \frac{3n^\alpha}{\check{c}}$. ■

We are now in a position to prove achievability.

Proof of Theorem 1: Case 1 : $0 < \alpha < 2$. Notice that the test $n^\alpha F \lessgtr 0$ makes the same decision as the test in the statement of the theorem. When hypothesis $\mathcal{H}_1$ is in effect (a subscript on operators denotes this) Lemma 1 tells us

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

where both $\sum_{\mathcal{X}_n} p_n^2(a)$ and $\sum_{\mathcal{Y}_n} q_n^2(a)$ are $O(n^{-\alpha})$. Therefore by Lemma 3 we have

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}_1[n^\alpha F] &= \liminf_{n\to\infty} n^\alpha \sum_{a\in\mathcal{A}_n} (p_n(a) - q_n(a))^2 \\ &\geq \liminf_{n\to\infty} \frac{\check{c}}{3} \|p_n - q_n\|_1^2,\end{aligned}$$

which is strictly positive by hypothesis. Invoking Lemma 2

$$\mathrm{Var}_1(n^\alpha F) \to 0$$

and the result follows from Chebyshev's inequality[2]. The hypothesis $\mathcal{H}_0$ is handled analogously.

Case 2: $\alpha = 0$. For this case we take square root of both sides of (3) so that we are working with norms. Now the result may be be proved using the weak law of large numbers (see for example Lemma 10 in Section IV). Suppose hypothesis $\mathcal{H}_0$ is in effect. The lefthand side of (3) is

$$\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2 \leq \|\Lambda_{X^n} - p_n\|_2 + \|\Lambda_{Z^n} - p_n\|_2$$

and both terms on the right of the previous display tend to zero in probability. For the righthand side, note that by the reverse triangle inequality

$$\Big| \|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2 - \|p_n - q_n\|_2 \Big| \leq \|\Lambda_{Y^n} - q_n\|_2 + \|\Lambda_{Z^n} - p_n\|_2$$

and so for n large enough $\|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2$ is as close to $\|p_n - q_n\|_2$ as we desire. Finally note that the hypothesis $\liminf_{n\to\infty} \|p_n - q_n\|_1 > 0$ implies $\liminf_{n\to\infty} \|p_n - q_n\|_2 > 0$ if the alphabet is not growing with n. ■

B. Converse

We next show that the result in Theorem 1 cannot be improved.

Theorem 2 (Converse)**.** *If $\alpha \geq 2$, then there are alphabets with growth rate $\Theta(n^\alpha)$ for which there are no α-universally consistent tests.*

To prove the result we need the following additional machinery.

Definition 4 (Testing Affinity)**.** *Suppose P and Q are probability measures on some space $\mathbb{X}$ dominated by λ with densities f and g. Let the density $f \wedge g$ define the (sub-probability) measure $P \wedge Q$, i.e.*

$$(P \wedge Q)(A) = \int_A (f \wedge g) d\lambda.$$

with $f \wedge g$ denoting the pointwise minimum of f and g.

Note that $2(a \wedge b) = a + b - |a - b|$, and so we may also write

$$\|P \wedge Q\|_1 = 1 - \frac{1}{2}\|P - Q\|_1. \tag{4}$$

Following Le Cam [24, Ch.16 §4] we associate with a hypothesis $\mathcal{H}_0$ (resp. $\mathcal{H}_1$) a set of measures, say A (resp. B). Let $0 \leq \phi \leq 1$ be a randomized test function, i.e. a function which gives the probability of accepting hypothesis $\mathcal{H}_0$. For a given ϕ and sets of measures A and B we define the worst case "average" error probability as follows

$$\mathfrak{R}(A, B, \phi) = \sup_{P\in A, Q\in B} \Big[\int (1 - \phi) dP + \int \phi dQ \Big],$$

[2]Sharper concentration results can be obtained using martingale techniques; see Theorem 9 in the Appendix for one such result.

and define the minimax error probability (or risk) as

$$\mathfrak{R}(A,B) = \inf_{\phi} \mathfrak{R}(A,B,\phi)$$

i.e. $\mathfrak{R}(A,B)$ is the best universally achievable risk. We recall the following result.

Lemma 4. *[Kraft [24, Ch.16 §4, Lem. 1]]*

$$\mathfrak{R}(A,B) = \sup_{P\in\mathrm{conv}(A),Q\in\mathrm{conv}(B)} \|P \wedge Q\|$$

where $\mathrm{conv}(A)$ *denote the convex hull of the set A.*

Equality (4) and Lemma 4 allow us to express minimax risk in terms of L_1 distances between convex hulls. We will also need the following result.

Lemma 5. *For any pair of probability measures P and Q, both dominated by a probability measure λ,*

$$\|P-Q\|_1^2 \le \int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda.$$

Proof: Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}\|P-Q\|_1 &= \int \Big|\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big| d\lambda \\ &\le \sqrt{\int d\lambda \int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda} \\ &= \sqrt{\int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda}.\end{aligned}$$

■

We now use these facts to establish a converse result. We first give a lower bound on the risk for a suitably chosen hypothesis testing problem on the sequence of alphabets $\mathcal{A}_n = \{1,\dots,\lceil n^\alpha\rceil_2\}$, where $\lceil\cdot\rceil_2$ denotes rounding up to the next even integer. Define sets

$$\begin{aligned}\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}} &= \{(p_n,q_n)\in\mathcal{P}(\mathcal{A}_n^{\times 2}) : \|p_n-q_n\|_1 \ge \epsilon, \check{c}n^{-\alpha} \le \min(\check{p}_n,\check{q}_n) \le \max(\hat{p}_n,\hat{q}_n) \le \hat{c}n^{-\alpha} \\ &\qquad \forall a\in\mathcal{A}_n : \max(p_n(a),q_n(a)) > 0\}, \\ A_{n,\alpha,\epsilon,\check{c},\hat{c}} &= \{p_n^n \times q_n^n \times p_n^n : (p_n,q_n)\in\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}\}, \\ \text{and } B_{n,\alpha,\epsilon,\check{c},\hat{c}} &= \{p_n^n \times q_n^n \times q_n^n : (p_n,q_n)\in\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}\}.\end{aligned}$$

Observe that for any choice of $\epsilon > 0$ and constants $\check{c},\hat{c}$ any sequence of pairs distributions $\{p_n,q_n\}$ with the nth chosen from $\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ is by definition α-large alphabet and moreover

$$\liminf_{n\to\infty} \|p_n-q_n\|_1 \ge \epsilon.$$

The following upper bound on the L_1 distance between the convex hulls of the sets for this testing problem combined with (4) give the aforementioned lower bound on the risk. The proof of the bound is similar in spirit to that of [10, Th. 4], which in turn borrows ideas from [25], using a so-called "mixture measure" to construct bad convex combinations. In our proof we apply the mixture measure idea to address the composite-versus-composite problem studied here.

Lemma 6. *Let* $0 < \epsilon < 1$. *For* $0 < \check{c} \le \frac{1-\epsilon}{3} < 1+\epsilon \le \hat{c}$ *there exists* $P_n \in \mathrm{conv}(A_{n,\alpha,\epsilon,\check{c},\hat{c}})$ *and* $Q_n \in \mathrm{conv}(B_{n,\alpha,\epsilon,\check{c},\hat{c}})$ *so that*

$$\|Q_n - P_n\|_1 \le \sqrt{2}\exp\Big(\frac{n^2\epsilon^4}{2n^\alpha}\Big).$$

Proof: Define $m = \lceil n^\alpha \rceil_2$. Let u_n be the uniform distribution on $\{1, \ldots, m\}$. Let $\Pi = \{-1, 1\}^{\times(m/2)}$ i.e. the set of all $\{-1, 1\}$ vectors of length $m/2$. For any $\pi \in \Pi$ let

$$\nu(i, \pi) = \begin{cases} \pi_{i/2} & i \text{ even} \\ -\pi_{(i+1)/2} & i \text{ odd,} \end{cases}$$

and define the distribution $q_{n,\pi}$ as

$$q_{n,\pi}(i) = (1 + \epsilon\nu(i, \pi))m^{-1} \text{ for } i \in \{1, \ldots, m\}.$$

We note that

$$\|q_{n,\pi} - u_n\|_1 = \epsilon \text{ for all } \pi. \tag{5}$$

Also since for all positive real x

$$x \leq \lceil x \rceil_2 \leq x + 2,$$

one has

$$\frac{1}{3} \leq \frac{n^\alpha}{m} \leq 1. \tag{6}$$

Define measures

$$P_{n,\pi} = u_n^n \times q_{n,\pi}^n \times u_n^n \text{ and } Q_{n,\pi} = q_{n,\pi}^n \times u_n^n \times u_n^n$$

and observe that (5), and (6) combined with

$$\frac{1 - \epsilon}{m} \leq \min(\check{u}_n, \check{q}_{n,\pi}) \leq \max(\hat{u}_n, \hat{q}_{n,\pi}) \leq \frac{1 + \epsilon}{m}$$

imply that $P_{n,\pi} \in A_{n,\alpha,\epsilon,\check{c},\hat{c}}$ and $Q_{n,\pi} \in B_{n,\alpha,\epsilon,\check{c},\hat{c}}$ for the $\epsilon, \check{c}$ and $\hat{c}$ of the theorem. Let μ denote the uniform distribution on the set Π and define mixtures

$$P_n = \sum_{\pi \in \Pi} P_{n,\pi}\mu(\pi) \text{ and } Q_n = \sum_{\pi \in \Pi} Q_{n,\pi}\mu(\pi).$$

Note that $P_n \in \text{conv}(A_{n,\alpha,\epsilon,\check{c},\hat{c}})$ and $Q_n \in \text{conv}(B_{n,\alpha,\epsilon,\check{c},\hat{c}})$ and further

$$P_n(\mathbf{x}, \mathbf{y}, \mathbf{z}) = m^{-2n} \sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(\mathbf{y})$$

and

$$Q_n(\mathbf{x}, \mathbf{y}, \mathbf{z}) = m^{-2n} \sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(\mathbf{x}).$$

We will now show that the stated L_1 bound holds for this choice of P_n and Q_n.

Taking $\lambda = u_n^n \times u_n^n \times u_n^n$ and invoking Lemma 5 we have

$$\begin{aligned}
\|P_n - Q_n\|_1^2 &\leq \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \Big(\frac{P_n(\mathbf{x}, \mathbf{y}, \mathbf{z}) - Q_n(\mathbf{x}, \mathbf{y}, \mathbf{z})}{\lambda(\mathbf{x}, \mathbf{y}, \mathbf{z})} \Big)^2 \lambda(\mathbf{x}, \mathbf{y}, \mathbf{z}) \\
&= E_\lambda \Big[\Big(\frac{P_n(X^n, Y^n, Z^n) - Q_n(X^n, Y^n, Z^n)}{\lambda(X^n, Y^n, Z^n)} \Big)^2 \Big] \\
&= \mathbb{E}_\lambda \Big[\Big(\frac{m^{-2n} \sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(Y^n) - m^{-2n} \sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(X^n)}{m^{-3n}} \Big)^2 \Big] \\
&= m^{2n} \mathbb{E}_\lambda \Big[\Big(\sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(Y^n) - \sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(X^n) \Big)^2 \Big] \\
&\leq m^{2n} \mathbb{E}_\lambda \Big[\Big(\sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(Y^n) \Big)^2 \Big] + m^{2n} \mathbb{E}_\lambda \Big[\Big(\sum_{\pi \in \Pi} \mu(\pi) q_{n,\pi}^n(X^n) \Big)^2 \Big].
\end{aligned}$$

Noting that under λ, Y^n and X^n have the same distribution and then expanding the square, we see that

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\leq 2m^{2n} \sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\mathbb{E}_\lambda\left[q_{n,\pi}^n(Y^n)q_{n,\gamma}^n(Y^n)\right] \\ &= 2m^{2n} \sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\mathbb{E}_\lambda\left[\prod_{i=1}^n q_{n,\pi}(Y_i)q_{n,\gamma}(Y_i)\right] \\ &= 2m^{2n} \sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\Big(\mathbb{E}_{u_n}[q_{n,\pi}(Y_i)q_{n,\gamma}(Y_i)]\Big)^n, \end{aligned} \tag{7}$$

where on the previous line we used the fact under λ the Y_i are i.i.d. uniform random variables. Focusing on the expectation alone

$$\begin{aligned}\mathbb{E}_{u_n}[q_{n,\pi}(Y_i)q_{n,\gamma}(Y_i)] &= \sum_i u_n(i)(1+\epsilon\nu(i,\pi))m^{-1}(1+\epsilon\nu(i,\gamma))m^{-1} \\ &= m^{-3}\sum_i 1 + \epsilon[\nu(i,\pi)+\nu(i,\gamma)] + \epsilon^2\nu(i,\pi)\nu(i,\gamma) \\ &= m^{-3}\sum_i 1 + \epsilon^2\nu(i,\pi)\nu(i,\gamma) \\ &= m^{-2} + m^{-3}\epsilon^2\sum_{i \text{ even}} \pi_{i/2}\gamma_{i/2} + \sum_{i \text{ odd}} \pi_{(i+1)/2}\gamma_{(i+1)/2} \\ &= m^{-2} + 2m^{-3}\epsilon^2\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\end{aligned}$$

where $\phi(\pi_i,\gamma_i) = 1$ when $\pi_i = \gamma_i$ and $\phi(\pi_i,\gamma_i) = -1$ otherwise. Applying this calculation to (7) yields

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\leq 2m^{2n} \sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\Big(m^{-2} + 2m^{-3}\epsilon^2\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\Big)^n \\ &= 2\sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\Big(1 + 2m^{-1}\epsilon^2\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\Big)^n \\ &\leq 2\sum_{\pi\in\Pi}\sum_{\gamma\in\Pi} \mu(\pi)\mu(\gamma)\exp\Big(\frac{2n\epsilon^2}{m}\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\Big)\end{aligned}$$

where we used the inequality $\log(1+x) \leq x$. Recalling that μ is uniform over $\{-1,1\}^{\times(m/2)}$ we may write

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\leq 2\mathbb{E}_{\pi,\gamma}\Big[\exp\Big(\frac{2n\epsilon^2}{m}\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\Big)\Big] \\ &= 2\Big(\frac{1}{2}\exp\Big(-\frac{2n\epsilon^2}{m}\Big) + \frac{1}{2}\exp\Big(\frac{2n\epsilon^2}{m}\Big)\Big)^{m/2}.\end{aligned}$$

Applying the inequality

$$\frac{1}{2}(\exp(u)+\exp(-u)) \leq \exp\Big(\frac{u^2}{2}\Big),$$

which follows from Hoeffding's Lemma (or by simply comparing the series expansions), gives

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\leq 2\exp\Big(\frac{2n^2\epsilon^4}{m^2}\Big)^{m/2} \\ &= 2\exp\Big(\frac{n^2\epsilon^4}{m}\Big),\end{aligned}$$

i.e.

$$\|P_n - Q_n\|_1 \leq \sqrt{2}\exp\Big(\frac{n^2\epsilon^4}{2m}\Big) \leq \sqrt{2}\exp\Big(\frac{n^2\epsilon^4}{2n^\alpha}\Big).$$

■

We are now in a position to prove Theorem 2. Roughly the argument is as follows. Recall that the setup of Lemma 6 provides the tester with ϵ, the minimum L_1 distance between distributions and constants $\check{c}, \hat{c}$. But even for this "easier" problem, there is some choice of $\check{c}, \hat{c}, \epsilon$ and distributions $P_n \in \mathrm{conv}(A_{n,\epsilon,\check{c},\hat{c}})$ and $Q_n \in \mathrm{conv}(B_{n,\epsilon,\check{c},\hat{c}})$ so that when $\alpha \geq 2$

$$\limsup_{n\to\infty} \|P_n - Q_n\|_1 < 2$$

implying that no α-universally consistent test is exists.

Theorem (2)**.** *If $\alpha \geq 2$, then there are alphabets with growth rate $\Theta(n^\alpha)$ for which there are no α-universally consistent tests.*

Proof: Let $\alpha \geq 2$, $\mathcal{A}_n = \{1, \ldots, \lceil n^\alpha \rceil_2\}$ and suppose by way of contradiction that there exists $\{T_n\}$, a universally consistent test for the α-large-alphabet hypothesis testing problem having alphabet $\mathcal{A}_n$. Now fix $0 < \epsilon < 1$, $0 < \check{c} \leq \frac{1-\epsilon}{3} < 1 + \epsilon \leq \hat{c}$ and choose $(p_n, q_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ so that

$$p_n^n \times q_n^n \times p_n^n(T_n = 1) \geq \frac{1}{2} \sup_{\tilde{p}_n,\tilde{q}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{p}_n^n \times \tilde{q}_n^n \times \tilde{p}_n^n(T_n = 1).$$

Since $\{T_n\}$ is α-universally consistent we have that

$$p_n^n \times q_n^n \times p_n^n(T_n = 1) \to 0$$

which in turn implies that

$$\sup_{\tilde{p}_n,\tilde{q}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{p}_n^n \times \tilde{q}_n^n \times \tilde{p}_n^n(T_n = 1) \to 0. \tag{8}$$

We now choose $(r_n, s_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ so that

$$r_n^n \times s_n^n \times s_n^n(T_n = 0) \geq \frac{1}{2} \sup_{\tilde{r}_n,\tilde{s}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{r}_n^n \times \tilde{s}_n^n \times \tilde{s}_n^n(T_n = 0),$$

and therefore again by universality we must have

$$\sup_{\tilde{r}_n,\tilde{s}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{r}_n^n \times \tilde{s}_n^n \times \tilde{s}_n^n(T_n = 0) \to 0. \tag{9}$$

Thus the existence of a α-universal test implies that

$$\sup_{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}} P_n(T_n = 1) \to 0$$

and

$$\sup_{Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}} Q_n(T_n = 0) \to 0$$

and therefore

$$\sup_{\substack{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}} \\ Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n = 1) + Q_n(T_n = 0) \to 0.$$

But

$$\sup_{\substack{P_n\in\mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}\\ Q_n\in\mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n=1)+Q_n(T_n=0) \geq \inf_{\tilde{T}_n}\sup_{\substack{P_n\in\mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}\\ Q_n\in\mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(\tilde{T}_n=1)+Q_n(\tilde{T}_n=0) \tag{10}$$

$$= \mathfrak{R}(\mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}},\mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}})$$

$$\geq 1-\frac{\sqrt{2}}{2}\exp\left(\frac{n^2\epsilon^4}{2n^\alpha}\right) \tag{11}$$

where in (10) the infimum is over all (randomized) tests and where (11) follows from Lemma 6 and (4). Note when $\alpha>2$ the exponential term goes to 1 as $n\to\infty$ and $1-\sqrt{2}/2$ is strictly greater than zero. When $\alpha=2$ taking $\epsilon=(1/2\log 2)^{1/4}>0$ gives $1-2^{-1/4}>0$. Thus for any $\alpha\geq 2$, choosing this ϵ and taking limits we obtain the inequality

$$\begin{aligned}0 &= \lim_{n\to\infty}\sup_{\substack{P_n\in\mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}\\ Q_n\in\mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n=1)+Q_n(T_n=0)\\ &\geq \lim_{n\to\infty} 1-\frac{\sqrt{2}}{2}\exp\left(\frac{n^2\epsilon^4}{2n^\alpha}\right)\\ &> 0\end{aligned}$$

a contradiction, and thus no such α-universal test $\{T_n\}$ exists. ■

Although we used a particular choice $\{\mathcal{A}_n\}$ to prove the converse, a slight modification of Theorem 2 goes through for any $\{\mathcal{A}_n\}$ with $|\mathcal{A}_n|=\Theta(n^\alpha)$. Thus we can in fact state the following more general theorem.

Theorem 3. *Let $\{\mathcal{A}_n\}$ by any sequence of alphabets with $|\mathcal{A}_n|=\Theta(n^\alpha)$. Then there are no α-universal consistent tests for any $\alpha\geq 2$.*

IV. Generalized Likelihood Ratio and Chi-Squared Tests

In this section we study the performance of two commonly used statistical tests: the generalized likelihood ratio and chi-squared tests. We show that both tests are α-universally consistent with sub-linear alphabet growth and that both tests are inconsistent with linear alphabet growth. Note that for both tests we actually prove *universal consistency* as opposed to merely α-universal consistency for up-to sub-linear alphabet growth, we return to this point in the conclusion.

A. GLRT and its Consistency

The GLRT is derived from the maximum likelihood method, which compares the likelihood functions evaluated with the most likely distribution in the hypothesis sets $\mathcal{H}_0$ and $\mathcal{H}_1$. This gives

$$\max_{p_n,q_n\in\mathcal{P}(\mathcal{A}_n)} p_n^n(X^n)q_n^n(Y^n)p_n^n(Z^n) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \max_{p_n,q_n\in\mathcal{P}(\mathcal{A}_n)} p_n^n(X^n)q_n^n(Y^n)q_n^n(Z^n),$$

where the maximizations are over *arbitrary distributions* on the alphabet $\mathcal{A}_n$. (Recall that the constants $\check{c},\hat{c}$ defining the α-large-alphabet sequence are unknown by the tester and the L_1 constraint is asymptotic in nature any so any p_n and q_n are feasible.)

The following Lemma allows us to rewrite the GLRT in terms of Kullback-Leibler divergences.

Lemma 7. *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 [22, Ex 1.3.19],

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

■

Using this Lemma combined with the well-known identity [22, Ch 1, Lemma 2.6]

$$p^n(\mathbf{x}) = \exp(-n[D(\Lambda_{\mathbf{x}}||p) + H(\Lambda_{\mathbf{x}})]) \tag{12}$$

we see that the GLRT test is equivalent to

$$D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n), \tag{13}$$

where $\hat{p}_n = (\Lambda_{X^n} + \Lambda_{Z^n})/2$ and $\hat{q}_n = (\Lambda_{Y^n} + \Lambda_{Z^n})/2$. Later it we will find the following useful. Define the functional

$$G(p, q, \mathcal{M}) = \sum_{a \in \mathcal{M}} p(a) \log\Big(\frac{2p(a)}{p(a)+q(a)}\Big) + q(a) \log\Big(\frac{2q(a)}{p(a)+q(a)}\Big)$$

and notice we may equivalently write the GLRT (13) as

$$G(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} G(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n).$$

We will also make use of the following result.

Lemma 8. *Suppose p and q are distributions on an alphabet $\mathcal{A}$, then*

$$G(p, q, \mathcal{A}) = \sum_{a \in \mathcal{A}} \sum_{i:\text{even}} \frac{1}{i(i-1)} \frac{(q(a) - p(a))^i}{(p(a)+q(a))^{i-1}}.$$

Further,

$$p(a) \log \frac{2p(a)}{p(a)+q(a)} + q(a) \log \frac{2q(a)}{p(a)+q(a)} \geq 0.$$

It turns out the growth-rate of the alphabet is of critical interest for proving consistency of the statistical tests. The following result allows us to prove a "weak law" for empirical distributions (to be used later) and Theorem 4, the consistency of the GLRT for sub-linear alphabet growth.

Lemma 9. *If $|\mathcal{A}_n| = o(n)$ then*[3]

$$n^{-1} \log |\mathcal{P}^n| \to 0 \text{ as } n \to \infty.$$

Proof: See [9, Lem. 1] ■

Lemma 10 (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 $\epsilon > 0$*

$$p_n^n(D(\Lambda_{X^n}||p_n) > \epsilon) \leq e^{-n(\epsilon - \delta_n)},$$

where $\delta_n(|\mathcal{A}_n|) \to 0$ as $n \to \infty$.

The final components of our proof of consistency of the GLRT (and chi-squared tests) are the following concentration results, which we include here for completeness.

[3]The sequence a_n has the property $a_n = o(b_n)$ iff $\lim \frac{a_n}{b_n} = 0$.

Definition 5. *A function $g : \mathcal{A}^n \to \mathbb{R}$ has the* bounded differences *property if for some non-negative constants $c_1, \ldots, c_n$,*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{A}} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i, \text{ for } 1 \leq i \leq n. \tag{14}$$

Lemma 11 (Efron-Stein Inequality [26], [27])**.** *Let $\mathcal{A}$ be any set and let $g^n : \mathcal{A}^n \to \mathbb{R}$ be a function of n variables. Define $Z = g(X_1,\ldots,X_n)$, where $X_1,\ldots,X_n$ are arbitrary independent random variables taking values in $\mathcal{A}$. Let $X_1',\ldots,X_n'$ be independent copies of $X_1,\ldots,X_n$ and define*

$$Z_i' = g(X_1,\ldots,X_i',\ldots,X_n)$$

then

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')^2].$$

Corollary 1. *Suppose g satisfies the hypothesis of Lemma 11 and has bounded differences with constant c. Then*

$$\mathrm{Var}(Z) \leq \frac{nc^2}{2}.$$

To establish consistency of the GLRT we also need

Lemma 12. *The quantity*

$$D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2)$$

viewed as a real-valued function of the vector $(\mathbf{x},\mathbf{z}) = (x_1,\ldots,x_n,z_1,\ldots,z_n)$ satisfies the bounded differences property with the single constant

$$\frac{2}{n}(1 + \log 2 + \log(1+n)).$$

Proof: See Appendix B. ■

Theorem 4. *If $|\mathcal{A}_n| = o(n)$ then the GLRT* (13) *is universally consistent.*

Proof: Suppose hypothesis $\mathcal{H}_0$ is in effect. Define the set

$$\mathcal{D}_n^{\epsilon} = \{(\mathbf{x},\mathbf{z}) : G(\Lambda_{\mathbf{x}},\Lambda_{\mathbf{z}}) > \epsilon\}.$$

By definition

$$\begin{aligned} P_n((X^n,Z^n) \in \mathcal{D}_n^{\epsilon}) &= \sum_{(\mathbf{x},\mathbf{z}) \in \mathcal{D}_n^{\epsilon}} p_n^n(\mathbf{x}) p_n^n(\mathbf{z}) \\ &= \sum_{\substack{Q_X \in \mathcal{P}^n(\mathcal{A}_n) \\ Q_Z \in \mathcal{P}^n(\mathcal{A}_n):\\ G(Q_X,Q_Z) > \epsilon}} \sum_{\substack{\mathbf{x} \in T(Q_X) \\ \mathbf{z} \in T(Q_Z)}} p_n^n(\mathbf{x}) p_n^n(\mathbf{z}). \end{aligned}$$

Using identity (12) and the bound [22, Ch 1, Lemma 2.5]

$$|T(Q_X)| \leq \exp(nH(Q_X)),$$

it follows that

$$\begin{aligned} &\sum_{\mathbf{x} \in T(Q_X)} \sum_{\mathbf{z} \in T(Q_Z)} p_n^n(\mathbf{x}) p_n^n(\mathbf{z}) \\ &\leq \exp(-n[D(Q_X||p_n) + D(Q_Z||p_n)]). \end{aligned}$$

Further, as in the proof of Lemma 7 we have for all distributions Q_X, Q_Z, p_n

$$D(Q_X||p_n) + D(Q_Z||p_n) \geq G(Q_X, Q_Z)$$

and therefore

$$P_n((X^n, Z^n) \in \mathcal{D}_n^\epsilon) \leq |\{\mathcal{P}(\mathcal{A}_n)\}|^2 e^{-n\epsilon}.$$

By way of Lemma 9 and the hypothesis, this implies that for all $\epsilon > 0$

$$P_n(D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) > \epsilon) \to 0 \text{ as } n \to \infty.$$

It remains to show that for some $\delta > 0$

$$\lim_{n\to\infty} P_n(D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n) > \delta) = 1. \tag{15}$$

Chebyshev's inequality tells us for any $\delta > 0$

$$\begin{aligned} &P_n(|D(\Lambda_{Y^n}||\hat{q}_n) - \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)]| > \delta) \\ &\leq \frac{\mathrm{Var}(D(\Lambda_{Y^n}||\hat{q}_n))}{\delta^2}. \end{aligned}$$

The bounded differences property (Lemma 12) and the Efron-Stein inequality (Lemma 11) imply that this variance goes to zero. Thus it follows with probability tending to one, $D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)$ 'concentrates' around $\mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)] + \mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)]$. Recalling $D(p||q)$ is convex in the pair (p, q), by Jensen's inequality

$$\begin{aligned} &\mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)] + \mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \\ &\geq D(\mathbb{E}[\Lambda_{Y^n}]||\mathbb{E}[\hat{q}_n]) + D(\mathbb{E}[\Lambda_{Z^n}]||\mathbb{E}[\hat{q}_n]) \\ &= D(q_n||(p_n + q_n)/2) + D(p_n||(p_n + q_n)/2), \end{aligned}$$

and from (2) and Pinsker's inequality [22, Ex 1.3.17]

$$\begin{aligned} &\liminf_{n\to\infty} D(p_n||(p_n + q_n)/2) + D(q_n||(p_n + q_n)/2) \\ &\geq \liminf_{n\to\infty} \frac{1}{4\log 2} \left(\sum_{a\in\mathcal{A}_n} |p_n(a) - q_n(a)| \right)^2 \\ &> 0. \end{aligned}$$

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 (15). Under hypothesis $\mathcal{H}_1$ the proof is similar. ■

We now show that when the alphabet growth is linear, i.e. $\alpha = 1$, the GLRT is not α-universally consistent. We do this by means of a particular counterexample which we will refer to throughout the remainder of the paper.

We first need the following technical result.

Lemma 13. *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x, y)$ the shadow (see [17]), i.e. the distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x, y)$ converges weakly to $\mu^2(x, y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\begin{aligned} \mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int \Big[&\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) \\ &- \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \frac{\exp(-x)x^k}{k!} \log(j+k) \Big] \, d\mu^2(x, y) \end{aligned}$$

and

$$\mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \to \int \Big[\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\log(2j) - \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^k}{k!}\log(j+k)\Big]\, d\mu^2(x,y).$$

Proof: See Appendix B. ∎

Theorem 5. *There exists a sequence of alphabets having linear growth for which the GLRT* (13) *is not* α*-universally consistent.*

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

$$p_n(a) = \begin{cases} \frac{1}{2n} & \text{if } a \in \{1,\ldots,n\} \\ \frac{1}{16n} & \text{if } a \in \{n+1,\ldots,9n\} \end{cases}$$

$$\text{and } q_n(a) = \begin{cases} \frac{5}{4n} & \text{if } a \in \{1,\ldots,n/2\} \\ \frac{1}{4n} & \text{if } a \in \{n/2+1,\ldots,n\} \\ \frac{1}{32n} & \text{if } a \in \{n+1,\ldots,9n\}. \end{cases}$$

Using Lemma 13, and numerically evaluating the resulting integrals, we see that under hypothesis $\mathcal{H}_0$,

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n)] = 1.085$$
$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)] = 1.026$$

whereas under hypothesis $\mathcal{H}_1$,

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n)] = 1.026$$
$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)] = 0.773.$$

From the Efron-Stein inequality and bounded differences property (Lemma 12), 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 $\mathcal{H}_1$. This is illustrated in section IV-D. ∎

Another well-known statistical procedure is chi-squared testing and we turn to that next.

B. Chi-Squared Test and its Consistency

For any distributions p and q on alphabet $\mathcal{A}$, and any $\mathcal{M} \subseteq \mathcal{A}$ introduce the functional[4]

$$\chi^2(p,q,\mathcal{M}) = \sum_{a\in\mathcal{M}} \frac{(p(a)-q(a))^2}{p(a)+q(a)}.$$

We will usually write $\chi^2(p,q)$ when the set $\mathcal{M}$ is taken for the full alphabet $\mathcal{A}$. Following [29, Ch 17, Ex. 3], one can apply the following chi-squared procedure to the present problem

$$\sum_{a\in\mathcal{A}_n} \frac{(\Lambda_{X^n}(a)-\hat{p}_n(a))^2}{\hat{p}_n(a)} + \frac{(\Lambda_{Z^n}(a)-\hat{p}_n(a))^2}{\hat{p}_n(a)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \sum_{a\in\mathcal{A}_n} \frac{(\Lambda_{Y^n}(a)-\hat{q}_n(a))^2}{\hat{q}_n(a)} + \frac{(\Lambda_{Z^n}(a)-\hat{q}_n(a))^2}{\hat{q}_n(a)}.$$

[4]For $\mathcal{M} = \mathcal{A}$ this functional is sometimes called the *triangular discrimination*, see [28].

After some manipulation, this yields

$$\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}), \tag{16}$$

which we will refer to as the chi-squared test (see also [24, Ch.4 §2]).

As with the GLRT, the chi-squared test is consistent with sublinear alphabet growth, in particular for $0 \leq \alpha < 1$. The proof is similar to that of the GLRT, and so only outline the argument.

Theorem 6. *Suppose $|\mathcal{A}_n| = o(n)$, then the chi-squared test* (16) *is universally consistent.*

Proof: Suppose hypothesis $\mathcal{H}_0$ is in effect, i.e. $X^n, Y^n, Z^n \sim P_n$. We will show the left side tends to zero in probability, while the other goes to something positive. For brevity we omit writing the alphabet argument in χ^2. Let $\epsilon > 0$. By Lemma taking the first term of the expansion from Lemma 8 we have that

$$D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) \geq \frac{1}{2}\chi^2(\Lambda_{X^n}, \Lambda_{Z^n})$$

therefore the event $\{D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) < \epsilon/2\}$ implies $\chi^2(p, q) < \epsilon$. Thus

$$P_n(\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) > \epsilon) \leq P_n(D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) > \epsilon/2)$$

which goes to zero according to the proof of Theorem 4.

An easy argument (see Lemma 22 in Appendix B) shows that $\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})$ viewed as a function from $\mathbb{R}^{2n} \to \mathbb{R}$ has the bounded differences property with constant $8n^{-1}$. Also, Jensen's inequality and the joint convexity of the function $(p-q)^2/(p+q)$ in p, q imply that

$$\begin{aligned}
\mathbb{E}_{P_n}\Big[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})\Big] &= \sum_a \mathbb{E}_{P_n}\Big[\frac{(\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2}{\Lambda_{Y^n}(a) + \Lambda_{Z^n}(a)}\Big]] \\
&\geq \frac{(\mathbb{E}_{P_n}[\Lambda_{Y^n}(a)] - \mathbb{E}_{P_n}[\Lambda_{Z^n}(a)])^2}{\mathbb{E}_{P_n}[\Lambda_{Y^n}(a)] + \mathbb{E}_{P_n}[\Lambda_{Z^n}(a)]} \\
&= \sum_a \frac{(p_n(a) - q_n(a))^2}{p_n(a) + q_n(a)}.
\end{aligned}$$

Now by Cauchy-Schwarz we have

$$\|p_n - q_n\|_1^2 = \Big(\sum_a \frac{|p_n(a) - q_n(a)|}{\sqrt{p_n(a) + q_n(a)}}\sqrt{p_n(a) + q_n(a)}\Big)^2 \leq 2\chi^2(p_n, q_n),$$

therefore Efron-Stein implies the random variable $\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})$ is concentrated around something strictly greater than $\frac{1}{2}\|p_n - q_n\|_1^2$, which is not tending to zero. ∎

We also have a corresponding result about inconsistency of the chi-squared test when $\alpha = 1$.

Lemma 14. *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x, y)$ the shadow (see [17]), i.e. the distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x, y)$ converges weakly to $\mu^2(x, y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to 2\int \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\, d\mu^2(x, y)$$

and

$$\begin{aligned}
\mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to &\int \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-y)y^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\frac{y}{x}\, d\mu^2(x, y) \\
&+ \int_{C^2} \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^k}{k!}\frac{(j-k)}{j+k}\, d\mu^2(x, y).
\end{aligned}$$

Proof: See Appendix B. ■

Theorem 7. *There exists a sequence of alphabets having linear growth for which the chi-squared test* (16) *is not* α*-universally consistent.*

Proof: Using the distributions from the proof of Theorem 5, applying Lemma 14, and numerically evaluating the resulting integrals, we see that under hypothesis $\mathcal{H}_0$,

$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.57$$
$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.49$$

whereas under hypothesis $\mathcal{H}_1$,

$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.49$$
$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.14.$$

By a similar argument as used in the proof of Theorem 5, it follows that under hypothesis $\mathcal{H}_0$, the test incorrectly declares $\mathcal{H}_1$. ■

C. Understanding the Inconsistency

The inconsistency of both the GLRT and chi-squared test for linear alphabets can be explained neatly by relating these tests to the L_2-norm test $nF \lessgtr 0$, where

$$F = \sum_a n(\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - \sum_a n(\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2.$$

Recall, from Lemmas 1 and 2 we know that the random variable nF concentrates around values which guarantee consistent detection, i.e. asymptotically $-\mathbb{E}_0[nF] = \mathbb{E}_1[nF] > 0$. But unlike our L_2-norm test, which weights all terms equally (by n), the χ^2 test weights the terms in the first sum of F by $(\Lambda_{X^n}(a) + \Lambda_{Z^n}(a))^{-1}$ and those in the second sum by $(\Lambda_{Y^n}(a) + \Lambda_{Z^n}(a))^{-1}$. 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.

For the case of the GLRT the same reasoning applies by reducing the GLRT to a chi-squared test via a Taylor series expansion, see Lemma 8. For these distributions, numerical calculations show it suffices to restrict attention to the case where the symbol count is zero in the training string and is positive in the test string or vice versa (in fact with high probability $N(a|X^n) = 0$ and $N(a|Z^n) \in \{1, 2, 3\}$ or vice-versa). This observation about the counts combined with Lemma 8 implies

$$G(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}) \approx \log(2)\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}),$$

(Lemma 23 in Appendix B makes this slightly more rigorous).

Another frequently used test in statistics is the Hellinger metric, $h(p, q)$, which for two mass functions p and q is defined via

$$h^2(p, q) = \frac{1}{2}\sum_{a\in\mathcal{A}}(\sqrt{p(a)} - \sqrt{q(a)})^2. \tag{17}$$

At first glance one may be tempted to think that the test

$$h^2(\Lambda_{X^n}, \Lambda_{Z^n}) \lessgtr h^2(\Lambda_{Y^n}, \Lambda_{Z^n}) \tag{18}$$

would not suffer from the same problems as the chi-squared test and GLRT since it does not involve divisions by empirical distributions. However since $(p-q)^2 = (\sqrt{p}-\sqrt{q})^2(\sqrt{p}+\sqrt{q})^2$, $h(p,q)$ may also be written as

$$h^2(p,q) = \frac{1}{2}\sum_{a\in\mathcal{A}} \frac{(p-q)^2}{(\sqrt{p}+\sqrt{q})^2},$$

and again the test involves divisions by counts. We conjecture (for evidence see the next sub-section) that the Hellinger test is not universally consistent for $\alpha = 1$[5].

D. Simulation ($\alpha = 1$ case)

In Figure 1 we show the empirical performance (over 10000 trials) of the L_2-norm classifier (3), the GLRT classifier (13), the chi-squared classifier (16) and the Hellinger classifier (18) for increasing n and a uniform prior on the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. The alphabet is $\mathcal{A}_n = \{1,\ldots,9n\}$; Example A refers to the distributions p_n, q_n appearing in the proof of Theorem 5; Example B is the same sequence p_n versus $r_n = 1/(9n)$, the uniform distribution. We see that in Example A the average error probability of the GLRT and chi-squared classifier tends to 1/2, as predicted by Theorems 5 and 7; we also notice the apparent inconsistency of the Hellinger test previously mentioned. In Example B, even though all tests seem to be consistent, the fraction of errors for our new classifier converges to zero more quickly than does the GLRT.

V. Testing with Infinite Training Data

In this section we suppose that the tester is given access to an "infinite" amount of training data, i.e. for each n he or she knows $(p_n, q_n, \mathcal{A}_n)$, the underlying distributions and alphabets. The following theorem answers the question for a sequence $\{p_n, q_n, \mathcal{A}_n\}$ satisfying

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

what, if any, are the conditions on the growth rate of the alphabet guaranteeing consistent testing between

$$\begin{aligned} &\mathcal{H}_0 : Z^n \sim p_n^n \\ &\mathcal{H}_1 : Z^n \sim q_n^n \quad \text{for all } n. \end{aligned}$$

Theorem 8. *For any sequence of alphabets $\{\mathcal{A}_n\}$ and sequence of distributions $\{p_n\}, \{q_n\}$ satisfying*

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

the likelihood ratio test

$$p_n^n(X^n) \lessgtr_{\mathcal{H}_0}^{\mathcal{H}_1} q_n^n(X^n)$$

is exponentially consistent, i.e. if

$$P_{e,n} = p_n^n(p_n^n(X^n) < q_n^n(X^n)) + q_n^n(p_n^n(X^n) > q_n^n(X^n))$$

denotes the sum of the type I and type II errors, then

$$\liminf -\frac{1}{n}\log\big(P_{e,n}\big) \geq \liminf \frac{1}{8}\Big(\sum_{a\in\mathcal{A}_n} |p_n - q_n|\Big)^2.$$

[5]The missing ingredient is the concentration of the random variable $h^2(\Lambda_{X^n}, \Lambda_{Z^n})$ about its mean. Once this is established one can readily verify using a calculation similar to Lemma 14 that the numerical values of the means imply the inconsistency. Concentration would also establish the consistency of the Hellinger test for sub-linear alphabet growth, since the inequality $\chi^2(p,q) \geq 2h^2(p,q)$ [24, Ch.4 §2] implies a proof along the lines of Theorem 6.

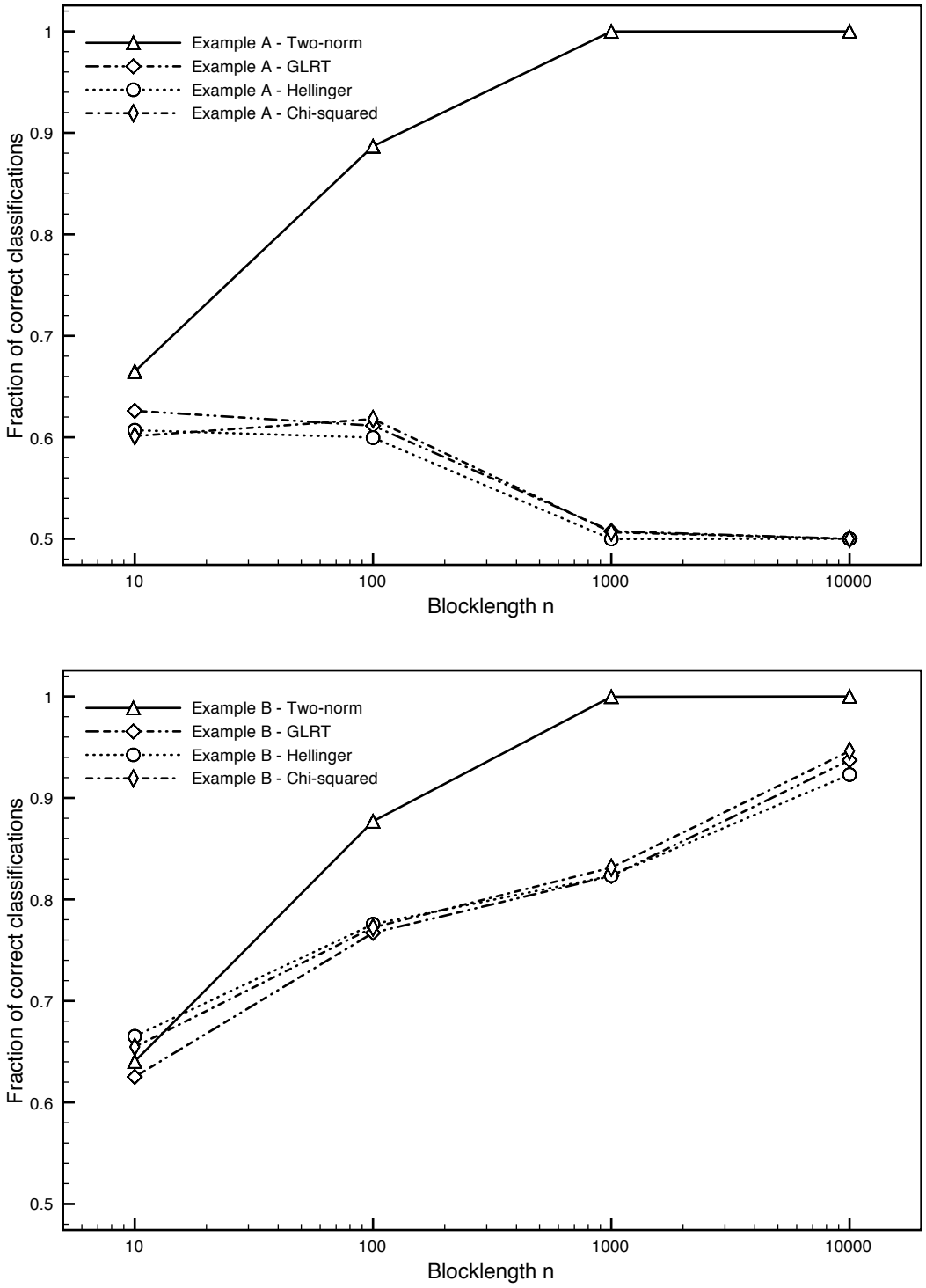

Fig. 1. Simulation of the performance of L_2-norm versus statistical tests. Example A illustrates the inconsistency of GLRT and Chi-squared (Theorems 5 and 7) and suggests inconsistency of Hellinger test.

Proof: By the Neyman Pearson theory the optimum test is the likelihood ratio test. Invoking Lemma 4 with the point sets $A_n = \{p_n^n\}, B_n = \{q_n^n\}$, we find the minimum error probability for this problem is

$$\mathfrak{R}(A_n, B_n) = 1 - \frac{1}{2}\|p_n^n - q_n^n\|_1.$$

To bound this probability, we follow [30, Cor. 13.1.1] and again make use of the Hellinger metric (17). First we recall the inequality (see [31, Ch.3])

$$h^2(p,q) \leq \frac{1}{2}\|p-q\|_1 \leq \sqrt{2}h(p,q). \tag{19}$$

For product measures it is well known that the Hellinger metric factorizes (see [31, Ch.3]). Thus in the i.i.d. case

$$h^2(p^n, q^n) = 1 - (1 - h^2(p,q))^n.$$

Applying these results allows us to write the following chain of inequalities

$$\begin{aligned}\mathfrak{R}(A_n, B_n) &= 1 - \frac{1}{2}\|p_n^n - q_n^n\|_1 \\ &\leq 1 - h^2(p_n^n, q_n^n) \\ &= (1 - h^2(p_n, q_n))^n \\ &\leq \exp(-nh^2(p_n, q_n)),\end{aligned}$$

where on the previous line we used the inequality $1 + x \leq \exp(x)$. Finally we use the right side of inequality (19) to give

$$\mathfrak{R}(A_n, B_n) \leq \exp(-n\frac{1}{8}\|p_n - q_n\|_1^2).$$

But by hypothesis

$$\liminf_{n\to\infty} \|p_n - q_n\|_1 > 0,$$

which gives the result. ■

Note that this result extends the classical fixed distribution, fixed alphabet i.i.d. case which states that the testing error, $\mathfrak{R}(\{p^n\}, \{q^n\})$, decays exponentially fast with the blocklength n when $p \neq q$ [30, Cor. 13.1.1]. In fact examining the proof we see that $nh^2(p_n, q_n) \to \infty$ is sufficient.

VI. Conclusions and Future work

We conclude with some comments on the general-source triangular array hypothesis testing problem (i.e. removing the α-source assumption). Firstly, Theorems 4 and 6 show that the GLRT and chi-squared tests are *universally* consistent (i.e. can handle non-homogeneous sources) provided that the underlying alphabet grows sub-linearly. Using Lemma 10 and bounding the L_2 distances by relative entropies (via Pinkser's inequality), one can also show that the L_2-test (3) is also universally consistent with sub-linear alphabet growth, provided that the asymptotic separation occurs in L_2, i.e. the assumption (2) is replaced by

$$\liminf_{n\to\infty} \|p_n - q_n\|_2^2 > 0.$$

The counterexample from the proof of Theorem 5 shows that neither the GLRT nor chi-squared test are universally consistent with linear alphabet growth. The following Lemma shows that the L_2-test (3) is also inconsistent for inhomogeneous sources with linear alphabet growth.

Lemma 15. *Let $\tilde{p}_n$ and $\tilde{q}_n$ be a sequence of $\alpha = 1$ large alphabet sources, defined on alphabet $\tilde{\mathcal{A}}_n$ such that $n\|\tilde{p}_n - \tilde{q}_n\|_2^2 = \epsilon$ for every n. Denote by ω a special symbol that does not occur in any of $\tilde{\mathcal{A}}_n$ and define*

$$\mathcal{A}_n = \tilde{\mathcal{A}}_n \cup \{\omega\}.$$

Let δ_x denote a point-mass at x and define $p_n = \frac{1}{2}\tilde{p}_n + \frac{1}{2}\delta_\omega$ and $q_n = \frac{1}{2}\tilde{q}_n + \frac{1}{2}\delta_\omega$. Then the test

$$\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2^2 \lessgtr \|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2^2$$

is inconsistent.

Proof: See Appendix C. ■

Roughly speaking the proof uses the fact that the L_2 distance for the $\alpha = 1$ component converges in probability to either $\frac{\epsilon}{4}$ or $-\frac{\epsilon}{4}$, but the variance for the symbol ω is order 1 in probability, and so reliable detection is impossible. Here the problem is that the L_2 test relies on the *unnormalized* counts, and a symbol with large probability can dominate the overall statistic. The GLRT and chi-squared test avoid this problem by using normalized counts, but as we have seen, this normalization can eliminate the bias necessary to ensure consistency.

Clearly it is desirable to know whether tests exist for the general-source triangular array hypothesis testing problem that are universally consistent for (super)-linear alphabet growth rates, or whether a converse along the lines of Theorem 2 can be proven. Theorem 2 provides an upper bound on the allowed growth rate, stating that quadratic alphabet growth cannot be handled.

Another line of investigation could be to distinguish between consistency and exponential consistency, following [9], and determine limits under the stricter exponential consistency requirement. A natural extension of the present problem is to suppose that we are given training sequences of length $N = N(n)$ and a test sequence of length n and ask for which alphabet growth rates and relation between N and n is universally consistent classification possible; for this problem, our Theorem 8 says that for any alphabet growth rate, $N = \infty$ and $n \to \infty$ can be handled.

APPENDIX A
PROOFS: SECTION III

This appendix is dedicated to the proof of Lemma 2, showing that the variance of F, the test random variable used for α-sources tends to zero when $0 < \alpha < 2$. (Note, such a result does not follow via other means, say Efron-Stein or bounded differences conditions.) We start by reproducing the moments of the binomial distribution.

Lemma 16 (Higher Moments of the Binomial)**.** *Suppose $N \sim$ Binomial(n, p)*

$$\begin{aligned}
\mathbb{E}[N^2] &= n^2p^2 + np(1-p) \\
\mathbb{E}[N^3] &= n^3p^3 + 3n^2p^2 - 3n^2p^3 + np - 3np^2 + 2np^3 \\
\mathbb{E}[N^4] &= n^4p^4 + 6n^3p^3 - 6n^3p^4 + 7n^2p^2 - 18n^2p^3 + 11n^2p^4 + np \\
&\quad - 7np^2 + 12np^3 - 6np^4
\end{aligned}$$

Proof: Direct calculation. ■

Computing the variance of F will require that the following results on covariance of multinomial vectors.

Lemma 17. *Suppose $X^n \sim p^n$. For $a \neq b$*

$$\begin{aligned}
\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] &= n(n-1)(n-2)(n-3)p^2(a)p^2(b) \\
&\quad + n(n-1)(n-2)[p(a)p^2(b) + p^2(a)p(b)] \\
&\quad + n(n-1)p(a)p(b)
\end{aligned}$$

Proof: Start by writing

$$\begin{aligned}
\mathbb{E}[N^2(a|X^n)]N^2(b|X^n)] &= \mathbb{E}\Big[\Big(\sum_{i=1}^n \mathbf{1}\{X_i = a\}\Big)^2\Big(\sum_{i=1}^n \mathbf{1}\{X_i = b\}\Big)\Big] \\
&= \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{l=1}^n \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_j = a\}\mathbf{1}\{X_k = b\}\mathbf{1}\{X_l = b\}]
\end{aligned}$$

Now observe that only certain cases have positive expectation these are

1) $i \neq j \neq k \neq l$ which occurs $n(n-1)(n-2)(n-3)$ times.
2) $i = j$ and $k \neq l$ with $i \neq k$ and $i \neq l$, which occurs $n(n-1)(n-2)$ times
3) $k = l$ and $i \neq j$ with $k \neq i$ and $k \neq j$, which occurs $n(n-1)(n-2)$ times.
4) $i = j$ and $k = l$ with $i \neq k$ which occurs $n(n-1)$ times.

■

Lemma 18. *Suppose* $X^n \sim p^n$. *For* $a \neq b$

$$\begin{aligned}\mathbb{E}[N^2(a|X^n)N(b|X^n)] &= n(n-1)(n-2)p^2(a)p(b) + n(n-1)p(a)p(b) \\ &= (n^3 - 3n^2 + 2n)p^2(a)p(b) + (n^2 - n)p(a)p(b)\end{aligned}$$

Proof: We have

$$\mathbb{E}[N^2(a|X^n)N(b|X^n)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_j = a\}\mathbf{1}\{X_k = b\}]$$

As in the proof of Lemma 17 only cases $i = j \neq k$ and $i \neq j \neq k$ yield a positive expectation. ■

To simplify the analysis we will use the following lemma to discard terms that vanish in the limit.

Lemma 19 (Discarding Rule)**.** *Suppose* $0 < \alpha < 2$ *and for all* $a, b \in \mathcal{A}_n$ *that* $p(a) = O(n^{-\alpha})$ *and* $q(b) = O(n^{-\alpha})$. *For integers* i, j *such that* $4 \geq j \geq i \geq 2$

$$1.\ n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} n^i p^j(a) \to 0 \text{ as } n \to \infty.$$

For positive integers i, j, k *such that* $4 \geq j + k > i \geq 2$ *or* $j + k = 4$ *and* $i = 1$

$$2.\ n^{2\alpha-4}\sum_{a,b\in\mathcal{A}_n} n^i p^j(a)q^k(b) \to 0 \text{ as } n \to \infty.$$

Proof: For the first property

$$\begin{aligned}n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} n^i p^j &\leq n^{2\alpha-4}\frac{3n^\alpha}{\check{c}}n^i\Big(\frac{\hat{c}}{n^\alpha}\Big)^j \\ &= n^{\alpha(3-j)-4+i}\frac{3\hat{c}^j}{\check{c}}\end{aligned}$$

Since $\alpha < 2$, examining the exponent alone we have for $3 \geq j$

$$\begin{aligned}\alpha(3-j) - 4 + i &< 2 - 2j + i \\ &\leq 2 - 2i + i \\ &\leq 2 - i \\ &\leq 0\end{aligned}$$

i.e $\alpha(3-j) - 4 + i < 0$. When $j = 4$ we have

$$-\alpha - 4 + i,$$

so for $i = 3$ the exponent is $-\alpha - 1 < 0$ and for $i = 4$ it is $-\alpha < 0$.

For the second property, argue with cases:

$$n^{2\alpha-4}\sum_{a,b\in\mathcal{A}_n} n^i p^j(a)q^k(b) \leq n^{4\alpha-4-(j+k)\alpha+i}\frac{9\hat{c}^{j+k}}{\check{c}^2}$$

when $i = 2, j + k = 3$ the sum behaves like $n^{\alpha-2}$, for $i = 2, (j + k) = 4$ it behaves like n^{-2} and for $i = 3, (j + k) = 4$ it behaves like n^{-1}, thus in all three cases the sum goes to zero when $0 < \alpha < 2$ as $n \to \infty$. For $j + k = 4$ and $i = 1$ the sum behaves like n^{-3}, which again goes to zero as $n \to \infty$. ■

Lemma (2)**.** *For* $i = 0, 1$

$$\mathrm{Var}_i[n^\alpha F] \to 0$$

for all $0 < \alpha < 2$*.*

Proof: Throughout we suppose hypothesis $\mathcal{H}_1$ is in effect and simply write $\mathbb{E}$ for $\mathbb{E}_1$, the other case is handled analogously.

$$\begin{aligned}
\mathbb{E}[F^2] = \mathbb{E}[\|\Lambda_{X^n}\|_2^4] &- 2\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2] \\
&- 4\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] + \mathbb{E}[\|\Lambda_{Y^n}\|_2^4] \\
&+ 4\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] \\
&+ 4\mathbb{E}[\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle^2].
\end{aligned}$$

Notice that every term in the expansion above has a common factor n^{-4} and therefore we will be dealing with terms such as

$$\begin{aligned}
n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^4] &= n^{2\alpha-4}\mathbb{E}\Big[\Big(\sum_{a\in\mathcal{A}_n} N^2(a|X^n)\Big)^2\Big] \\
&= n^{2\alpha-4}\mathbb{E}\Big[\sum_{a,b\in\mathcal{A}_n} N^2(a|X^n)N^2(b|X^n)\Big] \\
&= n^{2\alpha-4}\Big[\sum_{a\in\mathcal{A}_n}\mathbb{E}[N^4(a|X^n)] + \sum_{a\neq b\in\mathcal{A}_n}\mathbb{E}[N^2(a|X^n)N^2(b|X^n)]\Big]. \qquad (20)
\end{aligned}$$

Using the discarding rule (Lemma 19) we can safely ignore terms that vanish in the limit. For example since $N(a|X^n)$ is binomial, recalling Fact 16 we see

$$\begin{aligned}
n^{2\alpha-4}\sum_{a\in\mathcal{A}_n}\mathbb{E}[N^4(a|X^n)] &= n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} n^4p_n^4(a) + 6n^3p_n^3(a) - 6n^3p_n^4(a) + 7n^2p_n^2(a) - 18n^2p_n^3(a) \\
&\quad + 11n^2p_n^4(a) + np_n(a) - 7np_n^2(a) + 12np_n^3(a) - 6np_n^4(a) \\
&\simeq n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} np_n(a) = n^{2\alpha-4}n, \qquad (21)
\end{aligned}$$

where the notation

$$a_n \simeq b_n \text{ means } \lim_{n\to\infty} a_n - b_n = 0.$$

For the "cross-terms", by Lemma 17 we have

$$\begin{aligned}
\sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] &= \sum_{a\neq b} n^4p_n^2(a)p_n^2(b) - 6n^3p_n^2(a)p_n^2(b) + 11n^2p_n^2(a)p_n^2(b) - 6np_n^2(a)p_n^2(b) \\
&\quad + (n^3 - 3n^2 + 2n)p_n^2(a)p_n(b) + (n^3 - 3n^2 + 2n)p_n^2(b)p_n(a) \\
&\quad + (n^2 - n)p_n(a)p_n(b)
\end{aligned}$$

Note that

$$\sum_{a\neq b} p(a)q^i(b) = \sum_b q^i(b)(1 - p(b)) = \sum_a q^i(a) - q^i(a)p(a)$$

therefore

$$\sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] = \sum_{a\neq b} n^4p_n^2(a)p_n^2(b) - 6n^3p_n^2(a)p_n^2(b) + 11n^2p_n^2(a)p_n^2(b) - 6np_n^2(a)p_n^2(b)$$
$$+ 2(n^3-3n^2+2n)\Big[\sum_a p_n^2(a) - p_n^3(a)\Big]$$
$$+ (n^2-n) - (n^2-n)\sum_a p_n^2(a).$$

Applying the discarding rule we see the terms

$$\sum_{a\neq b} n^4p_n^2(a)p_n^2(b) + 2n^3\sum_a p_n^2(a) + n^2 - n$$

are significant in the limit. Therefore combining the previous display and (21) calculations gives

$$n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^4] \simeq n^{2\alpha-4}\Big(\sum_{a\neq b} n^4p_n^2(a)p_n^2(b) + 2n^3\sum_a p_n^2(a) + n^2\Big).$$

An analogous argument tells us that

$$n^{2\alpha}\mathbb{E}[\|\Lambda_{Y^n}\|_2^4] \simeq n^{2\alpha-4}\Big(\sum_{a\neq b} n^4q_n^2(a)q_n^2(b) + 2n^3\sum_a q_n^2(a) + n^2\Big).$$

We now turn our attention to

$$\begin{aligned}
&n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2]\\
&= n^{2\alpha-4}\Big(\sum_a n^2p_n^2(a) + np_n(a) - np_n^2(a)\Big)\Big(\sum_a n^2q_n^2(a) + nq_n(a) - nq_n^2(a)\Big)\\
&= n^{2\alpha-4}\Big(n + \sum_a n^2p_n^2(a) - np_n^2(a)\Big)\Big(n + \sum_a n^2q_n^2(a) - nq_n^2(a)\Big)\\
&= n^{2\alpha-4}\Big(n^2 + \sum_a (n^3q_n^2(a) - n^2q_n^2(a)) + \sum_a (n^3p_n^2(a) - n^2p_n^2(a))\\
&\quad + \sum_{a,b}(n^2p_n^2(a) - np_n^2(a))(n^2q_n^2(b) - nq_n^2(b))\Big)
\end{aligned}$$

In the final sum, the expansion starts with $n^4p_n^2(a)q_n^2(b)$ plus terms of lower order in n (still with a product of 4 probabilities), therefore applying our discarding rule we see

$$\begin{aligned}
&- 2n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2]\\
&\simeq -2n^{2\alpha-4}\Big(n^2 + \sum_a n^3q_n^2(a) + \sum_a n^3p_n^2(a) + \sum_{a,b} n^4p_n^2(a)q_n^2(b)\Big).
\end{aligned}$$

Now we turn to

$$\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] = \mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n}\rangle] - \mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\langle\Lambda_{Z^n}, \Lambda_{Y^n}\rangle] \tag{22}$$

The first term on the right is

$$\begin{aligned}
&n^{-4}\mathbb{E}\Big[\Big(\sum_a N^2(a|X^n)\Big)\Big(\sum_a N(a|Z^n)N(a|X^n)\Big)\Big]\\
&= n^{-4}\sum_{a,b}\mathbb{E}[N^2(a|X^n)N(b|X^n)]\mathbb{E}[N(b|Z^n)]\\
&= n^{-4}\sum_a \mathbb{E}[N^3(a|X^n)]\mathbb{E}[N(a|Z^n)] + n^{-4}\sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N(b|X^n)]\mathbb{E}[N(b|Z^n)].
\end{aligned} \tag{23}$$

Applying Fact 1 and the discarding rule to the first sum in (23) gives

$$n^{2\alpha-4}\sum_a \mathbb{E}[N^3(a|X^n)]\mathbb{E}[N(a|Z^n)] = n^{2\alpha-4}\Big[\sum_a (n^3p_n^3(a) + 3n^2p_n^2(a) - 3n^2p_n^3(a) + np_n(a) - 3np_n^2(a) + 2np_n^3(a))nq_n(a)\Big] \simeq 0.$$

For the second sum of (23), applying Lemma 18 gives

$$n^{-4}\Big[\sum_{a\neq b}(n^4 - 3n^3 + 2n^2)p_n^2(a)p_n(b)q_n(b) + (n^3 - n^2)p_n(a)p_n(b)q_n(b)\Big]$$

and we see only terms

$$n^{-4}\Big[\sum_{a\neq b} n^4p_n^2(a)p_n(b)q_n(b) + n^3p_n(a)p_n(b)q_n(b)\Big]$$

are significant. Turning to the second term of the right of (22) we have

$$\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\langle\Lambda_{Z^n},\Lambda_{Y^n}\rangle] = n^{-4}\sum_{a,b}[np_n(a) + n^2p_n^2(a) - np_n^2(a)]n^2q_n^2(b)$$

and it follows that the significant terms are

$$n^{-4}\sum_{a\neq b} n^3p_n(a)q_n^2(b) + n^4p_n^2(a)q_n^2(b).$$

Therefore

$$\begin{aligned}&-4n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]\\ &\simeq -4n^{2\alpha-4}\Big(\sum_{a\neq b} n^4p_n^2(a)p_n(b)q_n(b) + n^3p_n(a)p_n(b)q_n(b) - n^3p_n(a)q_n^2(b) - n^4p_n^2(a)q_n^2(b)\Big).\end{aligned}$$

The term

$$\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]$$

can be handled as above and we see that

$$\begin{aligned}&4n^{2\alpha}\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]\\ &\simeq 4n^{2\alpha-4}\Big(\sum_{a\neq b} n^4q_n^2(a)q_n(b)p_n(b) + n^3q_n(a)q_n(b)p_n(b) - n^3q_n(a)q_n^2(b) - n^4q_n^2(a)q_n^2(b)\Big).\end{aligned}$$

The final term is

$$
\begin{aligned}
&\mathbb{E}[\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle^2] \\
&= n^{-4}\mathbb{E}\Big[\Big(\sum_a N(a|Z^n)\big(N(a|X^n) - N(a|Y^n)\big)\Big)^2\Big] \\
&= n^{-4}\sum_{a,b}\mathbb{E}\Big[N(a|Z^n)\big(N(a|X^n) - N(a|Y^n)\big)N(b|Z^n)\big(N(b|X^n) - N(b|Y^n)\big)\Big] \\
&= n^{-4}\sum_{a,b}\Big(\mathbb{E}[N(a|Z^n)N(b|Z^n)]\mathbb{E}[N(a|X^n)N(b|X^n)] - \mathbb{E}[N(a|Z^n)N(b|Z^n)]\mathbb{E}[N(a|X^n)]\mathbb{E}[N(b|Y^n)] \\
&\qquad - \mathbb{E}[N(a|Z^n)N(b|Z^n)]\mathbb{E}[N(a|Y^n)\mathbb{E}[N(b|X^n)] + \mathbb{E}[N(a|Z^n)N(b|Z^n)]\mathbb{E}[N(a|Y^n)N(b|Y^n)]\Big) \\
&= n^{-4}\sum_a [nq_n(a) + (n^2-n)q_n^2(a)][np_n(a) + (n^2-n)p_n^2(a)] - [nq_n(a) + (n^2-n)q_n^2(a)]n^2p_n(a)q_n(a) \\
&\qquad - [nq_n(a) + (n^2-n)q_n^2(a)]n^2q_n(a)p_n(a) + [nq_n(a) + (n^2-n)q_n^2(a)][nq_n(a) + (n^2-n)q_n^2(a)] \\
&\quad + \sum_{a\neq b}(n^2-n)^2q_n(a)q_n(b)p_n(a)q_n(b) - (n^4-n^3)q_n(a)q_n(b)p_n(a)q_n(b) \\
&\qquad - (n^4-n^3)q_n(a)q_n(b)q_n(a)p_n(b) + (n^2-n)^2q_n(a)q_n(b)q_n(a)q_n(b).
\end{aligned}
$$

In the last line of the previous display, terms in the summation over a are such that every probability is accompanied by an n of the same or lesser power and therefore these terms vanish in the limit. In the summation over $a \neq b$ every term involves four probabilities so we only keep the n^4 terms. Hence

$$
\begin{aligned}
&\mathbb{E}[\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle^2] \\
&\sim 4n^{2\alpha-4}\Big(n^4\sum_{a\neq b} q_n(a)q_n(b)p_n(a)q_n(b) - q_n(a)q_n^2(b)p_n(a) - q_n^2(a)q_n(b)p_n(b) + q_n^2(a)q_n^2(b)\Big).
\end{aligned}
$$

Combining all the above we have shown that

$$
\begin{aligned}
E[n^{2\alpha}F^2] \simeq n^{2\alpha-4}\Bigg[&\Big(n^2 + 2\sum_a n^3p_n^2(a) + \sum_{a\neq b} n^4p_n^2(a)p_n^2(b)\Big) \\
&- 2\Big(n^2 + \sum_a n^3q_n^2(a) + \sum_a n^3p_n^2(a) + \sum_{a\neq b} n^4p_n^2(a)q_n^2(b)\Big) \\
&- 4\Big(\sum_{a\neq b} n^4p_n^2(a)p_n(b)q_n(b) + n^3p_n(a)p_n(b)q_n(b) - n^3p_n(a)q_n^2(b) - n^4p_n^2(a)q_n^2(b)\Big) \\
&+ \Big(\sum_{a\neq b} n^4q_n^2(a)q_n^2(b) + 2\sum_a n^3q_n^2(a) + n^2\Big) \\
&+ 4\Big(\sum_{a\neq b} n^4q_n^2(a)q_n(b)p_n(b) + n^3q_n(a)q_n(b)p_n(b) - n^3q_n(a)q_n^2(b) - n^4q_n^2(a)q_n^2(b)\Big) \\
&+ 4\Big(\sum_{a\neq b} n^4q_n(a)q_n(b)p_n(a)q_n(b) - n^4q_n(a)q_n^2(b)p_n(a) \\
&\qquad - n^4q_n^2(a)q_n(b)p_n(b) + n^4q_n^2(a)q_n^2(b)\Big)\Bigg].
\end{aligned}
$$

In the above there are several simplifications, for example all of the n^3 terms self-cancel (note

$$
n^3\sum_{a\neq b} p_n(a)q_n^2(b) = n^3\sum_a q_n^2(a) - q_n^3(a) \sim n^3\sum_a q_n^2(a)).
$$

After performing the cancellations we have

$$E[n^{2\alpha}F^2] \simeq n^{2\alpha}\Big(\sum_{a\neq b} p_n^2(a)p_n^2(b) - 4p_n^2(a)p_n(b)q_n(b) + 2p_n^2(a)q_n^2(b) + q_n^2(a)q_n^2(b) + 4q_n(a)q_n(b)p_n(a)q_n(b) - 4q_n(a)q_n^2(b)p_n(a)\Big).$$

We now compute

$$\begin{aligned} n^{2\alpha}\mathbb{E}[F]^2 &= n^{2\alpha}\Big(\sum_{a\in\mathcal{A}_n}(p_n(a)-q_n(a))^2 + n^{-1}(q_n^2(a)-p_n^2(a))\Big)^2 \\ &= n^{2\alpha}\sum_{a,b\in\mathcal{A}_n}\big((p_n(a)-q_n(a))^2 + n^{-1}(q_n^2(a)-p_n^2(a))\big)\big((p_n(b)-q_n(b))^2 + n^{-1}(q_n^2(b)-p_n^2(b))\big) \end{aligned}$$

Since every term in the above sum involves a quartic product of probabilities it follows that

$$\begin{aligned} n^{2\alpha}\mathbb{E}[F]^2 &\simeq n^{2\alpha}\sum_{a\neq b}((p_n(a)-q_n(a))^2((p_n(b)-q_n(b))^2 \\ &= n^{2\alpha}\sum_{a\neq b}(p_n^2(a) - 2p_n(a)q_n(a) + q_n^2(a))(p_n^2(b) - 2p_n(b)q_n(b) + q_n^2(b)) \\ &= n^{2\alpha}\sum_{a\neq b} p_n^2(a)p_n^2(b) - 2p_n^2(a)p_n(b)q_n(b) + p_n^2(a)q_n^2(b) \\ &\quad - 2p_n(a)q_n(a)p_n^2(b) + 4p_n(a)q_n(a)p_n(b)q_n(b) - 2p_n(a)q_n(a)q_n^2(b) \\ &\quad + q_n^2(a)p_n^2(b) - 2q_n^2(a)p_n(b)q_n(b) + q_n^2(a)q_n^2(b) \\ &= n^{2\alpha}\sum_{a\neq b} p_n^2(a)p_n^2(b) - 4p_n^2(a)p_n(b)q_n(b) + 2p_n^2(a)q_n^2(b) \\ &\quad + 4p_n(a)q_n(a)p_n(b)q_n(b) - 4p_n(a)q_n(a)q_n^2(b) + q_n^2(a)q_n^2(b). \end{aligned}$$

Therefore we have shown for $0 < \alpha < 2$

$$n^{2\alpha}\mathbb{E}[F^2] \simeq n^{2\alpha}\mathbb{E}[F]^2$$

giving the result. ■

We note that concentration results sharper than those obtained with Chebyshev's inequality and the variance calculation can be obtained in some cases using Martingale techniques. For $\alpha = 1$ one such result is as follows.

Theorem 9. *For $\alpha = 1$ and any $\gamma > 0$*

$$\Pr\Big(|F - \mathbb{E}[F]| > \gamma\Big) \leq 2\exp\Big(-\frac{\epsilon^2\gamma^2 n}{96(n^{1/3}+\Theta(1))^2}\Big) \tag{24}$$

$$+ \Big(1 + \frac{\Theta(1)}{\gamma(1-\epsilon)}\Big)3n\exp\Big(-\frac{(n^{1/3}-\Theta(1))^2}{2(\hat{c} + (n^{1/3}-\Theta(1))/3)}\Big). \tag{25}$$

Proof:

$$t_y(j) = \begin{cases} j-n & \text{if } j \in \{n+1,\ldots,2n\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } t_z(j) = \begin{cases} j-2n & \text{if } j \in \{2n+1,\ldots,3n\} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\{\mathcal{F}_j\}_{j=1}^{3n}$ be a filtration defined as

$$\mathcal{F}_j = \sigma(X_1^j, Y_1^{t_y(j)}, Z_1^{t_z(j)})$$

and define a Doob Martingale $\{W_j\}_{j=0}^{3n}$ as follows

$$W_j = \begin{cases} \mathbb{E}[F(X^n, Y^n, Z^n)] & \text{if } j = 0 \\ \mathbb{E}[F(X^n, Y^n, Z^n)|\mathcal{F}_j] & j \in \{1, \dots, 3n\}. \end{cases}$$

Let $D_j = W_j - W_{j-1}$ be the resulting martingale difference sequence (MDS) and $a^* \in \mathcal{A}_n$ be a most likely symbol over the measures p_n, q_n. Using the bounds established in Lemma 20 we have for $j \in \{1, \dots, n\}$

$$\begin{aligned} \Pr(|D_j| > \alpha) &\leq \Pr\Big(\frac{2}{n}(N(X_j|X_1^{j-1}) + \Theta(1)) > \alpha\Big) \\ &\leq \Pr(N(a^*|X^n) + \Theta(1) > (n/2)\alpha). \end{aligned}$$

Taking $\alpha = \frac{2}{n}(n^{1/3} + \Theta(1))$ gives

$$\Pr\Big(|D_j| > \frac{2}{n}(n^{1/3} + \Theta(1))\Big) \leq \Pr(N(a^*|X^n) > n^{1/3})$$

We now make use of the following 'Chernoff Inequality' [32], which states that if X is binomial, then

$$\Pr(X \geq \mathbb{E}[X] + \lambda) \leq \exp\Big(-\frac{\lambda^2}{2(\mathbb{E}[X] + \lambda/3)}\Big).$$

Now using $\lambda = n^{1/3} - \hat{c}$ in Chernoff Inequality, we have[6] for $j \in \{1...n\}$

$$\Pr(|D_j| > \frac{2}{n}(n^{1/3} + \Theta(1))) \leq \exp\Big(-\frac{(n^{1/3} - \hat{c})^2}{2(\hat{c} + (n^{1/3} - \hat{c})/3)}\Big),$$

similar bounds apply for $j \in \{n+1, \dots, 3n\}$. A result of [33], [34] states for any MDS (D_j), for every $\gamma > 0$ and each sequence of positive numbers (w_j) and any $0 < \epsilon < 1$,

$$\begin{aligned} \Pr\Big(|\sum_j D_j| > \gamma\Big) \leq 2\exp\Big(\frac{-\epsilon^2\gamma^2}{8\sum_{j=1}^n w_j^2}\Big) \\ + \Big(1 + \frac{\|D^*\|_\infty}{\gamma(1-\epsilon)}\Big)\sum_{j=1}^n \Pr(|D_j| > w_j), \end{aligned}$$

where $\|D^*\|_\infty = \sup_i \|D_i\|_\infty$. For our particular set of D_i, it follows from Lemma 20 that the worst case jump is only $\Theta(1)$, therefore $\|D_i\|_\infty \leq \Theta(1)$. Choosing $w_j = \frac{2}{n}(n^{1/3} + \Theta(1))$, $j = 1, \dots, 3n$ gives

$$\begin{aligned} \Pr\Big(|\sum_j D_j| > \gamma\Big) \leq 2\exp\Big(-\frac{\epsilon^2\gamma^2 n}{96(n^{1/3} + \Theta(1))^2}\Big) \\ + \Big(1 + \frac{\Theta(1)}{\gamma(1-\epsilon)}\Big)3n\exp\Big(-\frac{(n^{1/3} - \Theta(1))^2}{2(\hat{c} + (n^{1/3} - \Theta(1))/3)}\Big). \end{aligned}$$

■

[6]Recall $\Pr(\mathcal{B}(n,p) > x)$ is monotonic increasing in p and for rare events sources $p_n(a) \leq \frac{\hat{c}}{n}$

Lemma 20. *Let $\{D_j\}$ be the martingale difference sequence appearing in the proof of Theorem 1 and t be the function defined there, then*

$$|D_j| \leq \begin{cases} \frac{2}{n}(N(X_j|X_1^{j-1}) + \Theta(1)) & j \in \{1,\dots,n\} \\ \frac{2}{n}(N(Y_{t_y(j)}|Y_1^{t_y(j)-1}) & \\ \quad + \Theta(1)) & j \in \{n+1,\dots,2n\} \\ \frac{2}{n}(N(Z_{t_z(j)}|Y^n) & \\ \quad + N(Z_{t_z(i)}|X^n) + \Theta(1)) & j \in \{2n+1,\dots,3n\}. \end{cases}$$

Proof: We will only do the third case, the others are similar. Let $j \in \{2n+1,\dots,3n\}$, let $\tilde{Z}_{t_z(j)}$ be an independent copy of $Z_{t_z(j)}$ define $\tilde{Z}^n = (Z_1,\dots,\tilde{Z}_{t_z(j)},\dots,Z_n)$, then

$$\begin{aligned} |D_j| = \frac{1}{n}\Big| \sum_{a\in\mathcal{A}_n} & \mathbb{E}[(N(a|X^n) - N(a|Z^n))^2 \\ & - (N(a|Y^n) - N(a|Z^n))^2 - N((a|X^n) - N(a|\tilde{Z}^n))^2 \\ & + (N(a|Y^n) - N(a|\tilde{Z}^n))^2|\mathcal{F}_j]\Big|. \end{aligned}$$

Expanding the squares and cancelling gives

$$\begin{aligned} |D_j| &= \frac{1}{n}\Big| \sum_{a\in\mathcal{A}_n} \mathbb{E}[2N(a|X^n)(N(a|\tilde{Z}^n) - N(a|Z^n)) \\ &\quad + 2N(a|Y^n)(N(a|Z^n) - N(a|\tilde{Z}^n))|\mathcal{F}_j]\Big| \\ &= \frac{2}{n}\Big| \sum_{a\in\mathcal{A}_n} \mathbb{E}[N(a|X^n)(\mathbf{1}\{\tilde{Z}_{t_z(j)} = a\} - \mathbf{1}\{Z_{t_z(j)} = a\}) \\ &\quad + N(a|Y^n)(\mathbf{1}\{Z_{t_z(j)} = a\} - \mathbf{1}\{\tilde{Z}_{t_z(j)} = a\})|\mathcal{F}_j]\Big| \\ &= \Big|\frac{2}{n} \sum_{a\in\mathcal{A}_n} (N(a|Y^n) - N(a|X^n))\mathbf{1}\{Z_{t_z(j)} = a\}) \\ &\quad + (N(a|X^n) - N(a|Y^n))\mathbb{E}[\mathbf{1}\{\tilde{Z}_{t_z(j)} = a\})] \end{aligned}$$

where on the previous line we used the fact that $X^n, Y^n, Z_{t_z(j)}$ are measurable with respect to $\mathcal{F}_j$. Applying the triangle inequality and the bound $p_n(a) \leq \hat{c}/n$ for all $a \in \mathcal{A}_n$ gives

$$\begin{aligned} |D_j| &\leq \frac{2}{n}\Big(N(Z_{t_z(j)}|X^n) + N(Z_{t_z(j)}|Y^n) \\ &\qquad + \sum_{a\in\mathcal{A}_n} (N(a|X^n) + N(a|Y^n))\frac{\hat{c}}{n}\Big) \\ &= \frac{2}{n}(N(Z_{t_z(j)}|X^n) + N(Z_{t_z(j)}|Y^n) + 2\hat{c}). \end{aligned}$$

■

APPENDIX B
PROOFS: SECTION IV

Lemma (8)**.** *Suppose p and q are distributions on an alphabet $\mathcal{A}$, then*

$$G(p,q,\mathcal{A}) = \sum_{a\in\mathcal{A}} \sum_{i:\text{even}} \frac{1}{i(i-1)} \frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}.$$

Further,

$$p(a)\log\frac{2p(a)}{p(a)+q(a)}+q(a)\log\frac{2q(a)}{p(a)+q(a)}\geq 0.$$

For another proof along the same lines see [28, Th. 1].

Proof: Suppose first that $\operatorname{supp}p=\operatorname{supp}q=\mathcal{A}$, then

$$\begin{aligned}
D\Big(p\Big|\Big|\frac{p+q}{2}\Big)&=\sum_a p(a)\log\Big(\frac{2p(a)}{p(a)+q(a)}\Big)\\
&=\sum_a p(a)\log\Big(1+\frac{p(a)-q(a)}{p(a)+q(a)}\Big)\\
&=\sum_a\Big[\frac{p(a)+q(a)}{2}+\frac{p(a)-q(a)}{2}\Big]\log\Big(1+\frac{p(a)-q(a)}{p(a)+q(a)}\Big)\\
&=\sum_a\Big[\frac{p(a)+q(a)}{2}+\frac{p(a)-q(a)}{2}\Big]\sum_{i=1}^{\infty}(-1)^{i+1}\Big(\frac{p(a)-q(a)}{p(a)+q(a)}\Big)^i\frac{1}{i}\\
&=\sum_a\sum_{i=1}^{\infty}(-1)^{i+1}\frac{1}{2i}\left(\frac{(p(a)-q(a))^i}{(p(a)+q(a))^{i-1}}+\frac{(p(a)-q(a))^{i+1}}{(p(a)+q(a))^i}\right).
\end{aligned}$$

Similarly

$$D\Big(q\Big|\Big|\frac{p+q}{2}\Big)=\sum_a\sum_{i=1}^{\infty}(-1)^{i+1}\frac{1}{2i}\left(\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}+\frac{(q(a)-p(a))^{i+1}}{(p(a)+q(a))^i}\right).$$

Combining the terms and using the fact that for i odd $(x-y)^i+(y-x)^i=0$, we get

$$\begin{aligned}
D\Big(p\Big|\Big|\frac{p+q}{2}\Big)+D\Big(q\Big|\Big|\frac{p+q}{2}\Big)&=\sum_a\sum_{i:\text{odd}}\frac{1}{i}\frac{(q(a)-p(a))^{i+1}}{(p(a)+q(a))^i}-\sum_{i:\text{even}}\frac{1}{i}\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}\\
&=\sum_a\sum_{i:\text{even}}\frac{1}{i(i-1)}\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}.
\end{aligned}$$

Turning to mismatched supports. Firstly whenever $p(a)>0$ and $q(a)=0$, by continuity conventions

$$\begin{aligned}
D(p(a)||(p(a)+q(a))/2)+D(q(a)||(p(a)+q(a))/2)&=D(p(a)||p(a)/2)\\
&=p(a)\log(2)
\end{aligned}$$

where $D(p(a)||q(a))=p(a)\log(p(a)/q(a))$, but since in this case

$$\begin{aligned}
\sum_{i:\text{even}}\frac{1}{i(i-1)}\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}&=p(a)\sum_{i:\text{even}}\frac{(-1)^i}{i(i-1)}\\
&=p(a)\log(2)
\end{aligned}$$

the expansion is valid. An analogous argument holds for $q(a)>0$ and $p(a)=0$ concluding the proof. ■

Lemma (10)**.** *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 $\epsilon>0$*

$$p_n^n(D(\Lambda_{X^n}||p_n)>\epsilon)\leq e^{-n(\epsilon-\delta_n)},$$

where $\delta_n(|\mathcal{A}_n|)\to 0$ as $n\to\infty$.

Proof:

$$\begin{aligned} p_n^n(D(\Lambda_{X^n}||p_n) > \epsilon) &= \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} \sum_{\mathbf{x}\in T(Q)} p_n^n(\mathbf{x}) \\ &= \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} |T(Q)|e^{-n[D(Q||p_n)+H(Q)]} \\ &\le \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} e^{-n\epsilon} \\ &\le |\mathcal{P}^n(\mathcal{A}_n)|e^{-n\epsilon}. \end{aligned}$$

Applying Lemma 9 gives the result. ■

Lemma 21.

$$\begin{aligned} &\sup_{j\in[0,n),k\in[0,n]} \Big|\frac{j+1}{n}\log\frac{2(j+1)}{j+1+k} - \frac{j}{n}\log\frac{2j}{j+k}\Big| \\ &\qquad \le \frac{1}{n}(1+\log 2+\log(1+n)) \end{aligned} \tag{26}$$

$$\textit{and} \sup_{j\in[0,n),k\in[0,n]} \Big|\frac{k}{n}\log\frac{2k}{k+j+1} - \frac{k}{n}\log\frac{2k}{k+j}\Big| \le \frac{1}{n}. \tag{27}$$

Proof: First we prove (26). Suppose $j \neq 0$, then

$$\begin{aligned} &\Big|\frac{j+1}{n}\log\frac{2(j+1)}{j+1+k} - \frac{j}{n}\log\frac{2j}{j+k}\Big| \\ &= \frac{1}{n}\Big|j\log\frac{2(j+1)}{j+1+k}\frac{j+k}{2j} + \log\frac{2(j+1)}{j+1+k}\Big| \\ &\le \frac{1}{n}\Big(\Big|j\log\frac{j^2+jk+j+k}{(j+1+k)j}\Big| + \log 2 + \Big|\log\frac{j+1}{j+1+k}\Big|\Big) \\ &\le \frac{1}{n}\Big(\frac{k}{(j+1+k)} + \log 2 + \log\Big(1+\frac{k}{j+1}\Big)\Big). \end{aligned}$$

Using the monotonicity of $\log(1+x)$ gives the bound of the lemma. For $j = 0$, continuity gives

$$\begin{aligned} &\Big|\frac{j+1}{n}\log\frac{2(j+1)}{j+1+k} - \frac{j}{n}\log\frac{2j}{j+k}\Big| \\ &= \frac{1}{n}\Big|\log\frac{2}{1+k}\Big| \\ &\le \frac{1}{n}(\log 2+\log(1+k)) \\ &\le \frac{1}{n}(\log 2+\log(1+n)), \end{aligned}$$

but since the bound of the lemma is larger, we have the result. To show (27), observe for $k \neq 0$ we have

$$\begin{aligned}\Big|\frac{k}{n}\log\frac{2k}{k+j+1} - \frac{k}{n}\log\frac{2k}{k+j}\Big| &= \frac{1}{n}\Big|k\log\frac{2k}{k+j+1}\frac{k+j}{2k}\Big| \\ &= \frac{1}{n}\Big|k\log\frac{k+j+1}{k+j}\Big| \\ &\leq \frac{1}{n}\frac{k}{k+j} \\ &\leq \frac{1}{n},\end{aligned}$$

where the previous line follows from $k \leq k+j$. The case $k=0$ is handled by continuity. ■

Lemma (12)**.** *The quantity*

$$D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2)$$

viewed as a real-valued function of the vector $(\mathbf{x},\mathbf{z}) = (x_1,\ldots,x_n,z_1,\ldots,z_n)$ *has the bounded differences property with constant*

$$\frac{2}{n}(1+\log 2+\log(1+n)).$$

Proof: Consider the difference

$$|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}'}||(\Lambda_{\mathbf{x}'}+\Lambda_{\mathbf{z}})/2)|$$

where $\mathbf{x}'$ is identical to $\mathbf{x}$ except for one position. Without loss of generality suppose the change from $\mathbf{x}$ to $\mathbf{x}'$ replaced an occurrence of $a \in \mathcal{A}_n$ with $b \in \mathcal{A}_n$ where $a \neq b$. It follows from the definition of relative entropy that

$$\begin{aligned}&|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}'}||(\Lambda_{\mathbf{x}'}+\Lambda_{\mathbf{z}})/2)| \\ &\leq \Big|\frac{N(a|\mathbf{x})}{n}\log\frac{2N(a|\mathbf{x})}{N(a|\mathbf{x})+N(a|\mathbf{z})} \\ &\qquad - \frac{N(a|\mathbf{x}')}{n}\log\frac{2N(a|\mathbf{x}')}{N(a|\mathbf{x}')+N(a|\mathbf{z})}\Big| \\ &\quad + \Big|\frac{N(b|\mathbf{x})}{n}\log\frac{2N(b|\mathbf{x})}{N(b|\mathbf{x})+N(b|\mathbf{z})} \\ &\qquad - \frac{N(b|\mathbf{x}')}{n}\log\frac{2N(b|\mathbf{x}')}{N(b|\mathbf{x}')+N(b|\mathbf{z})}\Big|. \qquad (28)\end{aligned}$$

Let

$$j+1 = N(a|\mathbf{x}) \text{ and } k = N(a|\mathbf{z}), \text{ then } j = N(a|\mathbf{x}'),$$

then the first absolute value in the righthand side of (28) is of the form

$$\Big|\frac{j+1}{n}\log\frac{2(j+1)}{(j+1)+k} - \frac{j}{n}\log\frac{2j}{j+k}\Big|$$

which is bounded by $\frac{1}{n}(1+\log 2+\log(1+n))$ from Lemma 21. For the second summand, suppose

$$j = N(b|\mathbf{x}) \text{ and } k = N(b|\mathbf{z}), \text{then } (j+1) = N(b|\mathbf{x}'),$$

and it follows the same bound holds. Now instead consider the difference

$$|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}'})/2)|$$

where $\mathbf{z}'$ is identical to $\mathbf{z}$ except for one position. Again, without loss of generality suppose that the change replaced an occurrence of $a \in \mathcal{A}_n$ with $b \in \mathcal{A}_n$ where $a \neq b$. It follows that

$$\begin{aligned}
&|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}'})/2)| \\
&\leq \Big| N(a|\mathbf{x}) \log \frac{2N(a|\mathbf{x})}{N(a|\mathbf{x})+N(a|\mathbf{z})} \\
&\qquad - N(a|\mathbf{x}) \log \frac{2N(a|\mathbf{x})}{N(a|\mathbf{x})+N(a|\mathbf{z}')} \Big| \\
&\quad + \Big| N(b|\mathbf{x}) \log \frac{2N(b|\mathbf{x})}{N(b|\mathbf{x})+N(b|\mathbf{z})} \\
&\qquad - N(b|\mathbf{x}) \log \frac{2N(b|\mathbf{x})}{N(b|\mathbf{x})+N(b|\mathbf{z}')} \Big|. \qquad (29)
\end{aligned}$$

Let

$$j+1 = N(a|\mathbf{z}) \text{ and } k = N(a|\mathbf{x}), \text{ then } j = N(a|\mathbf{z}'),$$

then by way of Lemma 21 the first absolute value of (29) is bounded by $\frac{1}{n}$. The second term is handled analogously. Since $\frac{2}{n} < \frac{2}{n}(1+\log 2+\log(1+n))$, the bounded differences property is established. ■

Lemma (13)**.** *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x,y)$ the shadow (see [17]), i.e. distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x,y)$ converges weakly to $\mu^2(x,y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\begin{aligned}
\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int_{C^2} \Big[&\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) \\
&- \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^{k}}{k!} \log(j+k) \Big]\ d\mu^2(x,y)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \to \int_{C^2} \Big[&\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) \\
&- \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^{k}}{k!} \log(j+k) \Big]\ d\mu^2(x,y).
\end{aligned}$$

Proof: For notational convenience let

$$\begin{aligned}
g_k^n(x) &= \binom{n}{k}\Big(\frac{x}{n}\Big)^k\Big(1-\frac{x}{n}\Big)^{n-k} \\
\text{and } g_k(x) &= \frac{x^k \exp(-x)}{k!},
\end{aligned}$$

and note for all sequences $x_n \to x$, $g_k^n(x_n) \to g_k(x)$. Now we compute

$$\begin{aligned}
\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] = n^{-1} &\sum_{a\in\mathcal{A}_n} \mathbb{E}[N(a|Z^n)\log 2N(a|Z^n)] \\
&- \mathbb{E}[N(a|Z^n)\log(N(a|X^n)+N(a|Z^n))]. \qquad (30)
\end{aligned}$$

Starting with the second term on the righthand side (recalling the convention that $0 \log 0 = 0$)

$$\begin{aligned}
&n^{-1} \sum_{a \in \mathcal{A}_n} \mathbb{E}[N(a|Z^n) \log(N(a|X^n) + N(a|Z^n))] \\
&= \sum_{a \in \mathcal{A}_n} \sum_{j=1}^{n} \frac{j}{n} \binom{n}{j} p_n(a)^j (1 - p_n(a))^{n-j} \\
&\quad \times \sum_{k=0}^{n} \binom{n}{k} p_n(a)^k (1 - p_n(a))^{n-k} \log(j + k) \\
&= \sum_{j=1}^{n} \sum_{k=0}^{n} \left[\sum_{a \in \mathcal{A}_n} p_n(a) g_{j-1}^{n-1}((n-1)p_n(a)) \times g_k^n(np_n(a)) \right] \log(j + k).
\end{aligned}$$

Using $\mathcal{B}(n, p)$ to denote a Binomial(n, p) random variable we have for all $n \geq \check{c}$

$$\begin{aligned}
1 &= \sum_{a \in \mathcal{A}_n} n^{-1} \mathbb{E}[\mathcal{B}(n, p_n(a))] \\
&= \sum_{a \in \mathcal{A}_n} \sum_{j=0}^{n} \frac{j}{n} \binom{n}{j} p_n(a)^j (1 - p_n(a))^{n-j} \\
&= \sum_{j=1}^{n} \sum_{k=0}^{n} \sum_{a \in \mathcal{A}_n} p_n(a) g_{j-1}^{n-1}((n-1)p_n(a)) g_k^n(np_n(a)) \\
&= \sum_{j=1}^{n} \sum_{k=0}^{n} \int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n} x\Big) g_k^n(x) d\mu_n(x),
\end{aligned}$$

where $\mu_n(\cdot) = \int_C \mu(\cdot, y) dy$. Thus it follows there exist a pair of random variables (J_n, K_n) taking values in $\{1, \ldots, n\} \times \{0, \ldots, n\}$,

$$\Pr(J_n = j, K_n = k) = \begin{cases} \int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n} x\Big) g_k^n(x) d\mu_n(x) & j \in \{1, \ldots, n\}, \\ & k \in \{0, \ldots, n\}. \\ 0 & \text{otherwise.} \end{cases}$$

Since $np_n(X_n)$ converges in distribution to W with distribution $\mu(\cdot) = \int_C \mu^2(\cdot, y) dy$, we can create a sequence of random variables $\{W_n\}$ such that $W_n =^d np_n(X_n)$ and converges to W almost surely. Then since $g_k^n(W_n) \to g_k(W)$ almost surely and g_k^n is bounded,

$$\lim_{n \to \infty} \mathbb{E}[g_{j-1}^{n-1}\Big(\frac{n-1}{n} W_n\Big) g_k^n(W_n)] = \mathbb{E}[g_{j-1}(W) g_k(W)],$$

and there are random variables (J, K) taking values in $\{1, \ldots\} \times \{0, \ldots\}$ with joint distribution so that

$$\Pr(J = j, K = k) = \begin{cases} \mathbb{E}[g_{j-1}(W) g_k(W)] & j, k \in \{1, \ldots\} \times \{0, \ldots\} \\ 0 & \text{otherwise,} \end{cases}$$

and (J_n, K_n) converge in distribution to the pair (J, K). Now,

$$\begin{aligned}\mathbb{E}[(J_n + K_n)] &= \sum_{j=1}^{n}\sum_{k=0}^{n}(j+k)\int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big)g_k^n(x)d\mu_n(x)\\ &= \int_C \sum_{j=1}^{n}\sum_{k=0}^{n}(j+k)g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big)g_k^n(x)d\mu_n(x)\\ &= \int_C (1+2x-\frac{x}{n})\ d\mu_n(x)\\ &\to 1+\int_C 2x\ d\mu(x)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[J+K] &= \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\int_C (j+k)\frac{e^{-x}x^{j-1}}{(j-1)!}\frac{e^{-x}x^k}{k!}\ d\mu(x)\\ &= \int_C \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}(j+k)\frac{e^{-x}x^{j-1}}{(j-1)!}\frac{e^{-x}x^k}{k!}\ d\mu(x)\\ &= \int_C (1+2x)\ d\mu(x).\end{aligned}$$

Hence $\mathbb{E}[(J_n+K_n)] \to \mathbb{E}[J+K]$ implying that J_n+K_n is uniformly integrable. It follows that $\log(J_n+K_n)$ is uniformly integrable and by way of monotone convergence

$$\mathbb{E}[\log(J_n+K_n)] \to \mathbb{E}[\log(J+K)].$$

which gives the convergence of the second term on the right of (30). A similar argument applies to the first term. Therefore

$$\begin{aligned}\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int_C \Big[\sum_{j=1}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\log(2j)\\ -\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\log(j+k)\Big]d\mu(x).\end{aligned}$$

An analogous argument establishes the second claim of the lemma. ■

Lemma (14)**.** *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x,y)$ the shadow (see [17]), i.e. distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x,y)$ converges weakly to $\mu^2(x,y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\mathbb{E}[\chi_2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to 2\int_{C^2}\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\ d\mu^2(x,y)$$

and

$$\begin{aligned}\mathbb{E}[\chi_2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to \int_{C^2}\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-y)y^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\frac{y}{x}\ d\mu^2(x,y)\\ +\int_{C^2}\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^k}{k!}\frac{(j-k)}{j+k}\ d\mu^2(x,y)\end{aligned}$$

Proof: The proof immediately follows that of Lemma 13, once one notices that

$$\begin{aligned}
\mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] &= n^{-1} \sum_{a \in \mathcal{A}_n} \mathbb{E}\Big[\frac{(N(a|X^n) - N(a|Z^n))^2}{N(a|X^n) + N(a|Z^n)}\Big] \\
&= n^{-1} \sum_{a \in \mathcal{A}_n} \mathbb{E}\Big[\frac{N(a|X^n)(N(a|X^n) - N(a|Z^n))}{N(a|X^n) - N(a|Z^n)}\Big] \\
&\quad + \mathbb{E}\Big[\frac{N(a|Z^n)(N(a|Z^n) - N(a|X^n))}{N(a|X^n) + N(a|Z^n)}\Big] \\
&= 2 \sum_{a \in \mathcal{A}_n} \sum_{j=1}^{n} \frac{j}{n} \binom{n}{j} p_n(a)^j (1 - p_n(a))^{n-j} \\
&\quad \times \sum_{k=0}^{n} \binom{n}{k} p_n(a)^k (1 - p_n(a))^{n-k} \frac{(j-k)}{j+k}.
\end{aligned}$$

■

Lemma 22. *For all $k \in [0, n]$ and $j \in [0, n]$*

$$\left| \frac{\left(\frac{j+1}{n} - \frac{k}{n}\right)^2}{\frac{j+1}{n} + \frac{k}{n}} - \frac{\left(\frac{j}{n} - \frac{k}{n}\right)^2}{\frac{j}{n} + \frac{k}{n}} \right| \leq \frac{4}{n}$$

Proof:

$$\begin{aligned}
\left| \frac{\left(\frac{j+1}{n} - \frac{k}{n}\right)^2}{\frac{j+1}{n} + \frac{k}{n}} - \frac{\left(\frac{j}{n} - \frac{k}{n}\right)^2}{\frac{j}{n} + \frac{k}{n}} \right| &= \frac{1}{n} \Big| \frac{(j+1-k)^2}{j+1+k} - \frac{(j-k)^2}{j+k} \Big| \\
&= \frac{1}{n} \Big| \frac{((j-k)^2 + 2(j-k) + 1)(j+k)}{(j+1+k)(j+k)} - \frac{(j-k)^2(j+1+k)}{(j+k)(j+1+k)} \Big| \\
&= \frac{1}{n} \Big| \frac{-(j-k)^2 + (2(j-k)+1)(j+k)}{(j+1+k)(j+k)} \Big| \\
&\leq \frac{1}{n} \Big| \frac{(j-k)^2}{(j+1+k)(j+k)} + \frac{(2j+2k+1)}{j+k+1} \Big| \\
&\leq \frac{1+2+1}{n}
\end{aligned}$$

Where the final inequality uses the triangle inequality and the fact that $(j-k)^2 \leq (j+k)^2$. ■

Lemma 23. *Define the sets*

$$\mathcal{Z}'_n(p, q, i) = \Big\{ a : p(a) = 0 \text{ and } q(a) = \frac{i}{n} \Big\}$$

$$\text{and } \mathcal{Z}_n(p, q, j) = \bigcup_{i=1}^{j} \mathcal{Z}'_n(p, q, i) \cup \mathcal{Z}'_n(q, p, i).$$

For all $j \geq 1$

$$G(p, q, \mathcal{A}) \geq \log(2) \chi^2(p, q, \mathcal{Z}_n(p, q, j))$$

Proof: Note that from the proof of Lemma 8 we know that the summand in the definition of $G(p, q, \mathcal{A})$ is non-negative, therefore

$$G(p, q, \mathcal{A}) \geq G(p, q, \mathcal{Z}_n(p, q)).$$

On the set $\mathcal{Z}_n(p,q,j)$ either $q(a)=0$ or $p(a)=0$ and when $q(a)=0$ we have that

$$p(a)\log\Big(\frac{2p(a)}{p(a)+q(a)}\Big)+q(a)\log\Big(\frac{2q(a)}{p(a)+q(a)}\Big)=p(a)\log(2)$$

and analogously the summand is $q(a)\log(2)$ when $p(a)=0$. Therefore

$$\begin{aligned}G(p,q,\mathcal{Z}_n(p,q,j)) &= \sum_{a\in\mathcal{Z}_n(p,q,j)} p(a)\log\Big(\frac{2p(a)}{p(a)+q(a)}\Big)+q(a)\log\Big(\frac{2q(a)}{p(a)+q(a)}\Big)\\ &= \log(2)\sum_{a\in\mathcal{Z}_n(p,q,j)}\frac{(p(a)-q(a))^2}{p(a)+q(a)}\\ &= \log(2)\chi^2(p,q,\mathcal{Z}_n(p,q,j))\end{aligned}$$

■

APPENDIX C
PROOFS: SECTION VI

In this appendix we prove the following result.

Lemma (15)**.** *Let $\tilde{p}_n$ and $\tilde{q}_n$ be a sequence of $\alpha=1$ large alphabet sources, defined on alphabet $\tilde{\mathcal{A}}_n$ such that $n\|\tilde{p}_n-\tilde{q}_n\|_2^2=\epsilon$ for every n. Denote by ω a special symbol that does not occur in any of $\tilde{\mathcal{A}}_n$ and define*

$$\mathcal{A}_n \;=\tilde{\mathcal{A}}_n\cup\{\omega\}.$$

Let δ_a denote a point-mass at a and define $p_n=\frac{1}{2}\tilde{p}_n+\frac{1}{2}\delta_\omega$ and $q_n=\frac{1}{2}\tilde{q}_n+\frac{1}{2}\delta_\omega$. Then the test

$$\|\Lambda_{X^n}-\Lambda_{Z^n}\|_2^2 \lessgtr \|\Lambda_{Y^n}-\Lambda_{Z^n}\|_2^2 \tag{31}$$

is inconsistent.

Throughout this appendix we assume the setup of Lemma 15, i.e. $X^n\sim p_n^n$, $Y^n\sim q_n^n$ and we will see it suffices to consider the case $Z^n\sim p_n^n$, i.e. hypothesis $\mathcal{H}_0$ is in effect.

We use the notation $X^{n/i}$ to mean X^n without the ith component, i.e.

$$X^{n/i}=X_1,X_2,\ldots,X_{i-1},X_{i+1},\ldots,X_n.$$

Lemma 24. *For any $i\in\{1,\ldots,n\}$*

$$N^2(a|X^n)=\mathbf{1}\{X_i=a\}(1+2N(a|X^{n/i}))+N^2(a|X^{n/i}).$$

Proof:

$$\begin{aligned}N^2(a|X^n) &= \Big(\mathbf{1}\{X_i=a\}+N(a|X^{n/i})\Big)^2\\ &= \mathbf{1}\{X_i=a\}+2N(a|X^{n/i})\mathbf{1}\{X_i=a\}+N^2(a|X^{n/i})\\ &= \mathbf{1}\{X_i=a\}(1+2N(a|X^{n/i}))+N^2(a|X^{n/i}).\end{aligned}$$

■

Lemma 25.

$$\mathbb{E}[N(a|X^n)N(b|X^n)]=\begin{cases}(n^2-n)p_n(a)p_n(b) & \text{if } a\neq b\\ np_n(a)+(n^2-n)p_n^2(a) & \text{if } a=b.\end{cases}$$

Proof: The proof is similar to that of Lemma 17 and so is omitted. ■

Let T denote the restriction of the L_2-norm test (31) to $\tilde{\mathcal{A}}_n$, i.e.

$$T(X^n, Y^n, Z^n) = \frac{1}{n^2} \sum_{a \in \tilde{\mathcal{A}}_n} N^2(a|X^n) - N^2(a|Y^n) - 2N(a|Z^n)[N(a|X^n) - N(a|Y^n)].$$

Lemma 26. *Under distribution* $P_n = p_n^n \times q_n^n \times p_n^n$

$$\mathrm{Var}[nT(X^n, Y^n, Z^n)] \to 0.$$

Proof: Recall the Efron-Stein inequality, which states that

$$\mathrm{Var}(nT) = n^2 \mathrm{Var}(T) \leq \frac{1}{2} n^2 \sum_{i=1}^{3n} \mathbb{E}[(T - \tilde{T}_i)^2]$$

where $\tilde{T}_i$ is identical to T except that the ith argument of T is replaced with an independent copy having the same distribution. Thus we now investigate what happens when we replace one of the X_i, Y_i or Z_i.

Denote by $\tilde{X}_i^n = X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n$, where $\tilde{X}_i =^d X_i$. Then for $i \in \{1, \ldots, n\}$

$$\begin{aligned} T - \tilde{T}_i &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} N^2(a|X^n) - N^2(a|\tilde{X}_i^n) - 2N(a|Z^n)(N(a|X^n) - N(a|\tilde{X}_i^n)) \\ &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(1 + 2N(a|X^{n/i}) - 2N(a|Z^n)) \end{aligned}$$

where on the previous line we used Lemma 24.

Hence for $i \in \{1, \ldots, n\}$ we have

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2} \mathbb{E}\Big[\Big(\sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(1 + 2N(a|X^{n/i}) - 2N(a|Z^n))\Big)^2\Big] \\ &= n^{-2} \mathbb{E}\Big[\sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(\mathbf{1}\{X_i = b\} - \mathbf{1}\{\tilde{X}_i = b\}) \\ &\qquad \times (1 + 2N(a|X^{n/i}) - 2N(a|Z^n))(1 + 2N(b|X^{n/i}) - 2N(b|Z^n))\Big] \end{aligned}$$

Let $S(a, b) = (1 + 2N(a|X^{n/i}) - 2N(a|Z^n))(1 + 2N(b|X^{n/i}) - 2N(b|Z^n))$, so that

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] - \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)] \\ &\quad - \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] + \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)]. \end{aligned}$$

Because the indicators act like selectors the above display may be written as

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] + \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}S(a, a)] \\ &\quad - \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] + \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)]. \end{aligned}$$

Now because $X_i =^d \tilde{X}_i$, we may write

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2}\Big[\sum_{a \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] - 2 \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)]\Big] \\ &= n^{-2}\Big[\sum_{a \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] - 2 \sum_{a \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = a\}S(a, a)] \\ &\quad - 2 \sum_{a \neq b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)].\Big]. \end{aligned}$$

Since $S \perp\!\!\!\perp (X_i, \tilde{X}_i)$ it remains to compute $\mathbb{E}[S(a,b)]$.

Expanding S gives

$$\begin{aligned}S(a,b) = 1 &+ 2N(b|X^{n/i}) - 2N(b|Z^n) + 2N(a|X^{n/i}) + 4N(a|X^{n/i})N(b|X^{n/i}) \\ &- 4N(a|X^{n/i})N(b|Z^n) - 2N(a|Z^n) - 4N(a|Z^n)N(b|X^{n/i}) + 4N(a|Z^n)N(b|Z^n)\end{aligned}$$

For $a = b$ applying Lemma 25 gives

$$\begin{aligned}\mathbb{E}[S(a,b)] &= 1 + 2(n-1)p_n(a) - 2np_n(a) + 2(n-1)p_n(a) + 4(n^2-3n+2)p_n^2(a) \\ &\quad + 4(n-1)p_n(a) - 4(n-1)p_n(a)np_n(a) - 2np_n(a) \\ &\quad - 4np_n(a)(n-1)p_n(a) + 4(n^2-n)p_n^2(a) + 4np_n(a) \\ &= 1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a).\end{aligned}$$

Similarly for $a \neq b$ we get

$$\begin{aligned}\mathbb{E}[S(a,b)] &= 1 + 2(n-1)p_n(b) - 2np_n(b) + 2(n-1)p_n(a) + 4(n^2-3n+2)p_n(a)p_n(b) \\ &\quad - 4(n-1)p_n(a)np_n(b) - 2np_n(a) - 4np_n(a)(n-1)p_n(b) \\ &\quad + 4(n^2-n)p_n(a)p_n(b) \\ &= 1 - 2p_n(b) - 2p_n(a) - 8np_n(a)p_n(b) + 8p_n(a)p_n(b).\end{aligned}$$

Putting things together we can now evaluate to give

$$\begin{aligned}n^2\mathbb{E}[(T-\tilde{T}_i)^2] = n^{-2}\Big[&\sum_{a\in\tilde{\mathcal{A}}_n} 2p_n(a)(1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a)) \\ &- 2\sum_{a\in\tilde{\mathcal{A}}_n} p_n(a)p_n(a)(1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a)) \\ &- 2\sum_{a\neq b\in\tilde{\mathcal{A}}_n} p_n(a)p_n(b)(1 - 2p_n(b) - 2p_n(a) - 8np_n(a)p_n(b) + 8p_n(a)p_n(b))\Big].\end{aligned}$$

We can get a valid upper bound by keeping only those terms which are positive, i.e.

$$\begin{aligned}n^2\mathbb{E}[(T-\tilde{T}_i)^2] \leq n^{-2}\Big[&\sum_{a\in\tilde{\mathcal{A}}_n} 2p_n(a)(1 + 8np_n(a) + 8p_n^2(a)) \\ &- 2\sum_{a\in\tilde{\mathcal{A}}_n} p_n(a)p_n(a)(-8p_n(a) - 8np_n^2(a)) \\ &- 2\sum_{a\neq b\in\tilde{\mathcal{A}}_n} p_n(a)p_n(b)(-2p_n(b) - 2p_n(a) - 8np_n(a)p_n(b))\Big].\end{aligned}$$

Now summing each factor in the squares braces, and just writing the order of the resulting sum we have

$$\begin{aligned}n^2\mathbb{E}[(T-\tilde{T}_i)^2] &\leq n^{-2}[O(1) + O(1) + O(n^{-2}) + O(n^{-2}) + O(n^{-2}) + O(n^{-1}) + O(n^{-1}) + O(n^{-1})] \\ &= O(n^{-2})\end{aligned}$$

and therefore

$$\sum_{i=1}^{n} n^2\mathbb{E}[(T-\tilde{T}_i)^2] \leq O(n^{-1}).$$

When changing a Y_i, proceeding as before we get

$$\begin{aligned}
T - \tilde{T}_{i+n} &= T(X^n, Y^n, Z^n) - T(X^n, \tilde{Y}_i^n, Z^n) \\
&= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} N^2(a|\tilde{Y}_i^n) - N^2(a|Y^n) + 2N(a|Z^n)[N(a|Y^n) - N(a|\tilde{Y}_i^n)] \\
&= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} [\mathbf{1}\{Y_i = a\} - \mathbf{1}\{\tilde{Y}_i = a\}](2N(a|Z^n) - 1 - 2N(a|Y^{n/i})).
\end{aligned}$$

Now define $U(a,b) = (2N(a|Z^n) - 1 - 2N(a|Y^{n/i}))(2N(b|Z^n) - 1 - 2N(b|Y^{n/i}))$, then

$$\begin{aligned}
n^2\mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}\Big[(\mathbf{1}\{Y_i = a\} - \mathbf{1}\{\tilde{Y}_i = a\})(\mathbf{1}\{Y_i = b\} - \mathbf{1}\{\tilde{Y}_i = b\})U(a,b)\Big] \\
&= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{Y_i = a\}\mathbf{1}\{Y_i = b\}U(a,b)] - \mathbb{E}[\mathbf{1}\{Y_i = a\}\mathbf{1}\{\tilde{Y}_i = b\}U(a,b)] \\
&\quad - \mathbb{E}[\mathbf{1}\{\tilde{Y}_i = a\}\mathbf{1}\{Y_i = b\}U(a,b)] + \mathbb{E}[\mathbf{1}\{\tilde{Y}_i = a\}\mathbf{1}\{\tilde{Y}_i = b\}U(a,b)] \\
&= n^{-2}\Bigg[\sum_{a \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{Y_i = a\}U(a,a)] - \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i = a\}\mathbf{1}\{Y_i = b\}U(a,b)] \Bigg] \\
&= n^{-2}\Bigg[\sum_{a \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{Y_i = a\}U(a,a)] - \sum_{a \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i = a\}\mathbf{1}\{Y_i = a\}U(a,a)] \\
&\quad - \sum_{a \neq b \in \tilde{\mathcal{A}}_n} 2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i = a\}\mathbf{1}\{Y_i = b\}U(a,b)] \Bigg].
\end{aligned}$$

Computing $\mathbb{E}[U(a,b)]$ yields

$$\begin{aligned}
\mathbb{E}[U(a,a)] &= 4np_n(a) + 4(n^2 - n)p_n^2(a) - 2np_n(a) - 4np_n(a)(n-1)q_n(a) \\
&\quad - 2np_n(a) + 1 + 2(n-1)q_n(a) - 4(n-1)q_n(a)np_n(a) \\
&\quad + 2(n-1)q_n(a) + 4(n-1)q_n(a) + 4(n-1)(n-2)q_n^2(a)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[U(a,b)] &= 4(n^2 - n)p_n(a)p_n(b) - 2np_n(a) - 4np_n(a)(n-1)q_n(b) \\
&\quad - 2np_n(b) + 1 + 2(n-1)q_n(b) - 4(n-1)q_n(a)np_n(b) \\
&\quad + 2(n-1)q_n(a) + 4(n-1)(n-2)q_n(a)q_n(b).
\end{aligned}$$

For any $a, b \in \tilde{\mathcal{A}}_n$ the absolute value of every term appearing in $U(\cdot,\cdot)$ is $O(1)$, and since $U(a,b) \perp\!\!\!\perp (Y_i, \tilde{Y}_i)$ it follows that

$$\sum_{i=1}^{n} n^2\mathbb{E}[(T - \tilde{T}_{i+n})^2] = O(n^{-1}).$$

When replacing a Z_i, we have

$$T - \tilde{T}_{i+2n} = n^{-2}2 \sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{\tilde{Z}_i = a\} - \mathbf{1}\{Z_i = a\})(N(a|X^n) - N(a|Y^n))$$

Thus for $i \in \{1, \dots, n\}$ we have

$$n^2\mathbb{E}[(T - \tilde{T}_{i+2n})^2] = n^{-2}\mathbb{E}\Big[\Big(\sum_{a\in\tilde{\mathcal{A}}_n}(\mathbf{1}\{\tilde{Z}_i = a\} - \mathbf{1}\{Z_i = a\})(N(a|X^n) - N(a|Y^n))\Big)^2\Big]$$
$$= n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}\Big[(\mathbf{1}\{Z_i = a\} - \mathbf{1}\{\tilde{Z}_i = a\})(\mathbf{1}\{Z_i = b\} - \mathbf{1}\{\tilde{Z}_i = b\})V(a,b)\Big]$$

where we defined

$$V(a,b) = (N(a|X^n) - N(a|Y^n))(N(b|X^n) - N(b|Y^n))$$
$$= N(a|X^n)N(b|X^n) - N(a|X^n)N(b|Y^n) - N(a|Y^n)N(b|X^n) + N(a|Y^n)N(b|Y^n).$$

Expanding the terms and using the selection property we get

$$n^2\mathbb{E}[(T - \tilde{T}_i)^2] = n^{-2}\Bigg[\sum_{a,b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{Z_i = b\}V(a,b)] - \mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)]$$
$$- \mathbb{E}[\mathbf{1}\{\tilde{Z}_i = a\}\mathbf{1}\{Z_i = b\}V(a,b)] + \mathbb{E}[\mathbf{1}\{\tilde{Z}_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)]\Bigg]$$
$$= n^{-2}\Bigg[2\sum_{a\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{Z_i = a\}V(a,a)] - 2\sum_{a,b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)]\Bigg]$$

On account of the independence of $(Z_i, \tilde{Z}_i)$ and $V(\cdot,\cdot)$ it remains to compute $\mathbb{E}[V(a,b)]$, yielding

$$\mathbb{E}[V(a,a)] = np_n(a) + (n^2 - n)p_n^2(a) - n^2p_n(a)q_n(a) - n^2p_n(a)q_n(a) + nq_n(a) + (n^2 - n)q_n^2(a)$$

and for $a \neq b$

$$\mathbb{E}[V(a,b)] = (n^2 - n)p_n(a)p_n(b) - n^2p_n(a)q_n(b) - n^2p_n(b)q_n(a) + (n^2 - n)q_n(a)q_n(b).$$

Each term appearing in $V(\cdot,\cdot)$ has absolute value $O(1)$ and so it follows that

$$\sum_{i=1}^{n} n^2\mathbb{E}[(T - \tilde{T}_{i+2n})^2] = O(n^{-1}).$$

Therefore we have shown

$$\mathrm{Var}(nT) \leq O(n^{-1}) \to 0.$$

■

Proof of Lemma 15: Suppose hypothesis $\mathcal{H}_0$ is in effect. Chebyshev's inequality combined with Lemma 26 imply that

$$n\Big[\sum_{a\in\tilde{\mathcal{A}}_n}(\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - (\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2\Big]$$

is close to its mean with high probability. Thus, using $\to_{P_n}$ to denote convergence in probability, we have

$$n\sum_{a\in\tilde{\mathcal{A}}_n}(\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - (\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2 \to_{P_n} -\epsilon/4.$$

Next we note that by the Central Limit Theorem,

$$2\sqrt{n}\Big(\Lambda_{X^n}(\omega) - \frac{1}{2}\Big) = 2\sqrt{n}\Big(\sum_{i=1}^{n}\frac{\mathbf{1}(X_i = \omega)}{n} - \frac{1}{2}\Big) \Rightarrow \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes a standard Normal random variable. Similarly $2\sqrt{n}(\Lambda_{Y^n}(\omega)-1/2) \Rightarrow \mathcal{N}(0,1)$ and $2\sqrt{n}(\Lambda_{Z^n}(\omega)-1/2) \Rightarrow \mathcal{N}(0,1)$. Furthermore the independence of the X^n, Y^n, Z^n sequences implies the independence of the limiting distributions. Let $\tilde{X}, \tilde{Y}, \tilde{Z}$ be independent $\mathcal{N}(0,1)$. Now by the continuous mapping theorem [35, Ch.1 §7] it follows that

$$\begin{aligned} 4n\Big[(\Lambda_{X^n}(\omega)-\Lambda_{Z^n}(\omega))^2-(\Lambda_{Y^n}(\omega)-\Lambda_{Z^n}(\omega))^2\Big] &\Rightarrow \tilde{X}^2+\tilde{Z}^2-2\tilde{X}\tilde{Z}-\tilde{Y}^2-\tilde{Z}^2+2\tilde{Y}\tilde{Z} \\ &= \tilde{X}^2-\tilde{Y}^2-2\tilde{Z}(\tilde{X}-\tilde{Y}). \end{aligned}$$

Finally, Slutsky's theorem [35, Ch.1 §5.4] tells us that if $X_n \Rightarrow X$ and $Y_n \to_P c$ then $X_n + Y_n \Rightarrow X + c$, therefore

$$4n\Big[\|\Lambda_{X^n}-\Lambda_{Z^n}\|_2^2-\|\Lambda_{Y^n}-\Lambda_{Z^n}\|_2^2\Big] \Rightarrow \tilde{X}^2-\tilde{Y}^2-2\tilde{Z}(\tilde{X}-\tilde{Y})-\epsilon.$$

This random variable has positive probability of being positive, and thus the test is inconsistent. ■

REFERENCES

[1] R. H. Baayen, *Word Frequency Distributions*. Kluwer Academic Press, 2001.

[2] J. Neyman and E. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part I," *Biometrika*, vol. 20A, no. 1/2, pp. 175–240, Jul 1928.

[3] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, Jan 1952.

[4] ——, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, Jan 1956.

[5] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Stat.*, Apr 1965.

[6] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401 – 408, Mar 1989.

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

[8] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1504 – 1517, 2002.

[9] A. Barron, "Uniformly powerful goodness of fit tests," *Ann. Stat.*, vol. 17, no. 1, pp. 107–124, Mar 1989.

[10] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct 2008.

[11] M. S. Ermakov, "Asymptotic minimaxity of chi-square tests," *Theory Probab. Appl.*, vol. 42, no. 4, pp. 589–610, 1998.

[12] L. Holst, "Asymptotic normality and efficiency for certain goodness-of-fit tests," *Biometrika*, vol. 59, no. 1, pp. 137–145, Apr 1972.

[13] M. Quine and J. Robinson, "Efficiencies of chi-square and likelihood ratio goodness-of-fit tests," *Ann. Stat.*, vol. 13, no. 2, pp. 727–742, Jun 1985.

[14] W. Kallenberg, "On moderate and large deviations in multinomial distributions," *Ann. Stat.*, vol. 13, no. 4, pp. 1554–1580, Dec 1985.

[15] T. R. Read and N. A. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, 1988.

[16] P. Harremoës and I. Vajda, "On Bahadur efficiency of power divergence statistics," 2010, submitted to IEEE Trans. Inf. Theory.

[17] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," 2009, submitted to IEEE Trans. Inf. Theory.

[18] N. Santhanam, A. Orlitsky, and K.Viswanathan, "New tricks for old dogs: Large alphabet probability estimation," in *Information Theory Workshop, 2007. ITW '07. IEEE*, 2007, pp. 638 – 643.

[19] J. Acharya, H. Das, A. Orlitsky, S. Pan, and N. P. Santhanam, "Classification using pattern probability estimators," in *IEEE International Symposium on Information Theory*, 2010, pp. 1493–1497.

[20] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Ndellec and C. Rouveirol, Eds. Springer Berlin / Heidelberg, 1998, vol. 1398, pp. 137–142, 10.1007/BFb0026683.

[21] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[22] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[23] E. V. Khmaladze, "Statistical analysis of large number of rare events," Centre for Mathematics and Computer Science, Netherlands, Tech. Rep. MS-R8804, 1988.

[24] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[25] Y. Ritov and P. Bickel, "Achieving information bounds in non and semiparametric models," *Ann. Stat.*, Jan 1990.

[26] B. Efron and C. Stein, "The jackknife estimate of variance," *Ann. Stat.*, vol. 9, no. 3, pp. 586 – 596, May 1981.

[27] J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *Ann. Stat.*, vol. 14, no. 2, pp. 753 – 758, Jun 1986.

[28] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602 – 1609, 2000.

[29] A. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

[30] E. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

[31] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, 1990.

[32] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79 – 127, Jan 2006.

[33] T. Chalker, A. Godbole, P. Hitczenko, and J. Radcliff, "On the size of a random sphere of influence graph," *Adv. Appl. Prob.*, Jan 1999.

[34] A. Godbole and P. Hitczenko, "Beyond the method of bounded differences," in *Microsurveys in Probability*, 1998, pp. 1528–32.

[35] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.