On Optimal Two Sample Homogeneity Tests for Finite Alphabets

Jayakrishnan Unnikrishnan
Audiovisual Communications Laboratory, School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Email: jay.unnikrishnan@epfl.ch

***Abstract*—Suppose we are given two independent strings of data from a known finite alphabet. We are interested in testing the null hypothesis that both the strings were drawn from the same distribution, assuming that the samples within each string are mutually independent. Among statisticians, the most popular solution for such a *homogeneity test* is the two sample chi-square test, primarily due to its ease of implementation and the fact that the limiting null hypothesis distribution of the associated test statistic is known and easy to compute. Although tests that are asymptotically optimal in error probability have been proposed in the information theory literature, such optimality results are not well-known and such tests are rarely used in practice. In this paper we seek to bridge the gap between theory and practice. We study two different optimal tests proposed by Shayevitz [1] and Gutman [2]. We first obtain a simplified structure of Shayevitz's test and then obtain limiting distributions of the test statistics used in both the tests. These results provide guidelines for choosing thresholds that guarantee an approximate false alarm constraint for finite length observation sequences, thus making these tests easy to use in practice. The approximation accuracies are demonstrated using simulations. We argue that such homogeneity tests with provable optimality properties could potentially be better choices than the chi-square test in practice.**

I. Introduction

Suppose we are given two independent strings of data $x^m := (x_1, x_2, \ldots, x_m)$ and $y^n := (y_1, y_2, \ldots, y_n)$ drawn from the same known finite alphabet $\mathsf{Z} := \{z_1, z_2, \ldots, z_N\}$. We are interested in testing *homogeneity*, i.e., whether or not both these strings are drawn i.i.d. from the same distribution in $\mathcal{P}(\mathsf{Z})$, the collection of all multinomial probability distributions on Z. In other words this hypothesis testing problem fundamentally aims to identify whether or not the two collections of samples are drawn from the same population. This is a fundamental problem in statistics with various practical applications [3].

This problem can be interpreted as a binary hypothesis testing problem with a composite null hypothesis representing the situation where both strings are drawn from identical distributions and a composite alternate hypothesis where both the strings are drawn from distinct distributions. Thus one can define two different probabilities of error, viz., the probability of false alarm under the null hypothesis and the probability of missed detection under the alternate hypothesis. A reasonable approach to solve this problem is to identify some testing procedure that optimizes the trade-off between these two quantities. Although it is intractable to solve this problem exactly, two different results satisfying two different notions of asymptotic optimality are known in the information theory literature [1] [2]. However, the most commonly used solution for this homogeneity testing problem is the two sample chi-square test [4] originally proposed by Pearson [5]. In this paper we first obtain a simplified structure for the test proposed in [1]. We then proceed to identify the limiting behavior of the test statistics used in the optimal tests of [1] and [2]. Such limiting results can be used to approximate the thresholds for these tests for a target false alarm probability, thus providing a practical alternative to the popular two sample chi-square test which does not have any known optimality properties.

A. Notation

For any probability mass function $\pi \in \mathcal{P}(\mathsf{Z})$ we use $\pi(z)$ to denote the probability of symbol $z \in \mathsf{Z}$. We sometimes also use the notation π to denote the vector of probabilities $(\pi(z_1), \pi(z_2), \ldots, \pi(z_N))$ and $\langle \pi, f \rangle$ to denote the inner product $\sum_{i=1}^{N} \pi(z_i) f(z_i)$ for any function f defined on Z. For two distributions $\pi, \nu \in \mathcal{P}(\mathsf{Z})$ the Kullback-Leibler divergence between π and ν is given by

$$D(\pi \| \nu) = \sum_{z \in \mathsf{Z}} \pi(z) \log \frac{\pi(z)}{\nu(z)}.$$

We use Γ_m^x to denote the empirical distribution of x^m and Γ_n^y to denote the empirical distribution of y^n. We use $\mathsf{P}_{\pi^1, \pi^2}$ to denote the probability measure when the first string is drawn from distribution π^1 and the second string is drawn from distribution π^2. When both strings are drawn from the same distribution μ we use P_μ for the probability measure. We use $\mathcal{N}(a, B)$ to denote a Gaussian random vector with mean a and covariance matrix B and χ_d^2 to denote a chi-square random variable with d degrees of freedom.

B. Outline

In Section II we describe the mathematical problem statement and describe the known results. We provide a simplified version of a known optimal test [1] in Section III. We then present our new results on the weak convergence of the various test statistics in Section IV. We discuss how these results can be used for selecting test thresholds in Section V and conclude in Section VI.

II. Problem description

Suppose we are given two independent strings of data $x^m := (x_1, x_2, \ldots, x_m)$ and $y^n := (y_1, y_2, \ldots, y_n)$ drawn from the same known finite alphabet Z. We think of x^m as the first m observations from an i.i.d. sequence x and y^n as the first n observations from another i.i.d. sequence y. We are interested in testing whether or not both these sequences are drawn from the same distribution in $\mathcal{P}(\mathsf{Z})$. For each $m, n \in \mathbb{Z}$ let $\phi_{m,n} : \mathsf{Z}^m \times \mathsf{Z}^n \mapsto \{0, 1\}$ represent a test on the first m observations from x and first n observations from y. The test outcome $\phi_{m,n}(x^m, y^n) = 0$ represents a decision in favor of the null hypothesis that both x and y are drawn from the same distribution and the outcome $\phi_{m,n}(x^m, y^n) = 1$ represents a decision in favor of the alternate hypothesis that x and y are drawn from different distributions. For any distribution $\mu \in \mathcal{P}(\mathsf{Z})$ of the observations under the null hypothesis, the probability of false alarm is given by

$$p_{FA}(\phi_{m,n}|\mu) = \mathsf{P}_\mu(\phi_{m,n}(x^m, y^n) = 1).$$

Similarly for any distinct distributions $\pi^1, \pi^2 \in \mathcal{P}(\mathsf{Z})$ the probability of missed detection is given by

$$p_{MD}(\phi_{m,n}|\pi^1, \pi^2) = \mathsf{P}_{\pi^1,\pi^2}(\phi_{m,n}(x^m, y^n) = 0).$$

In the classical Neyman-Pearson formulation of hypothesis testing one seeks to minimize the probability of missed detection subject to an upper bound on the probability of false alarm. In our problem since we do not know the values of μ, π^1 or π^2, it is not possible to solve this problem exactly. Instead, we have to use an asymptotic version. For this purpose we consider the limit as $m, n \to \infty$. We further assume that m scales linearly in n as $m = \lambda n$ for some $\lambda \geq 1$. In this setting we use $\phi_n(x, y)$ to denote the test outcome $\phi_{\lambda n,n}(x^{\lambda n}, y^n)$. We define two kinds of error exponents. The false alarm error-exponent and the missed-detection exponent are defined respectively as

$$\begin{aligned} E_{FA}(\phi|\mu) &:= \liminf_{n\to\infty} -\frac{1}{n}\log p_{FA}(\phi_n|\mu) \\ E_{MD}(\phi|\pi^1, \pi^2) &:= \liminf_{n\to\infty} -\frac{1}{n}\log p_{MD}(\phi_n|\pi^1, \pi^2). \end{aligned}$$

Two versions of asymptotically optimal tests are known in literature.

Shayevitz [1] studied this problem in the context of a two-sensor network. The null hypothesis corresponds to the scenario in which both sensors observe noise and the alternate hypothesis corresponds to the scenario in which some phenomenon is present which leads to both sensors making observations from distinct distributions. One of the contributions of [1] is a solution to the following optimization problem:

$$\begin{aligned} \sup_\phi \quad & E_{MD}(\phi|\pi^1, \pi^2) \\ s.t. \quad & \lim_{n\to\infty} p_{FA}(\phi_n|\mu) = 0 \text{ for all } \mu \in \mathcal{P}(\mathsf{Z}). \end{aligned} \tag{1}$$

The following sequence of tests solves the problem (1):

$$\phi_n^A(x, y) = \mathcal{I}\left\{\inf_{\mu\in\mathcal{P}(\mathsf{Z})} \max\{D(\Gamma^x_{\lambda n}\|\mu), D(\Gamma^y_n\|\mu)\} \geq \delta_n\right\} \tag{2}$$

where $\delta_n = \frac{|\mathsf{Z}|\log n}{n}$ and $\mathcal{I}$ denotes the indicator function.

Gutman [2] studied this problem in the context of multihypothesis testing with training sequences. He used the following optimality criterion

$$\begin{aligned} \sup_\phi \quad & E_{MD}(\phi|\pi^1, \pi^2) \\ s.t. \quad & E_{FA}(\phi_n|\mu) \geq \eta \text{ for all } \mu \in \mathcal{P}(\mathsf{Z}) \end{aligned} \tag{3}$$

and showed that the following sequence of likelihood ratio tests solves problem (3):

$$\begin{aligned} \phi_n^B(x, y) \quad = \quad & \mathcal{I}\left\{\lambda D(\Gamma^x_{\lambda n}\|\tfrac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n)) \right. \\ & \left. + D(\Gamma^y_n\|\tfrac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n)) \geq \tilde{\delta}_n\right\} \end{aligned} \tag{4}$$

where $\tilde{\delta}_n = \eta + \mathrm{O}(\frac{\log n}{n})$. Interestingly, both the optimal sequences of tests of (2) and (4) do not depend on the true values of π^1 and π^2.

Although these optimal solutions are known, the test usually used by statisticians is the two sample chi-square test. The chi-square distance between two distributions is defined as

$$\chi^2(\pi, \nu) := \sum_{z\in\mathsf{Z}} \frac{2(\pi(z) - \nu(z))^2}{(\pi(z) + \nu(z))}, \quad \pi, \nu \in \mathcal{P}(\mathsf{Z}).$$

The two sample chi-square test is given by

$$\phi_n^C(x, y) = \mathcal{I}\left\{\chi^2(\Gamma^x_{\lambda n}, \Gamma^y_n) \geq \hat{\delta}_n\right\} \tag{5}$$

where $\hat{\delta}_n$ is chosen to approximately meet the false alarm constraint based on the weak convergence of the test statistic.

The main reason for popularity of the chi-square test is the fact that the test statistic is easy to compute and that the limiting behavior of the test statistic is known, which makes it possible to set an approximate threshold for a target false alarm probability, or to compute the *p-value* of the test statistic [3]. We observe that in both the optimal tests (2) and (4), the guarantees on the false alarm probability (p_{FA}) hold only in the asymptotic sense as the sequence length goes to infinity. In (2) we are guaranteed that p_{FA} will eventually go to zero and in (4) we are guaranteed that p_{FA} decays as $\exp(-n\eta)$. However, in practice one has access to only a finite number of data points and is interested in guaranteeing a constant upper bound on the false alarm probability. For this, we need to be able to choose the thresholds of these tests to meet a given target false alarm level for a given sequence length. In the following sections we derive weak-convergence results for the test statistics used in (2) and (4) and demonstrate that these results give good approximations to the actual false alarm probabilities in these tests. We first obtain a simplified form of the test of (2).

III. Simplified form of Test ϕ^A

The test statistic used in the test ϕ^A of (2) can be simplified a great deal via the following lemma.

Lemma III.1. *For any distributions $\mu^1, \mu^2 \in \mathcal{P}(\mathsf{Z})$, the infimum in*

$$\inf_{\nu\in\mathcal{P}(\mathsf{Z})} \max\{D(\mu^1\|\nu), D(\mu^2\|\nu)\} \tag{6}$$

is achieved at a point ν^* that satisfies $D(\mu^1\|\nu^*) = D(\mu^2\|\nu^*)$. Furthermore, ν^* can be expressed in the form $\nu^* = \alpha\mu^1 + (1-\alpha)\mu^2$ for some $0 \le \alpha \le 1$.

Proof: Define the function $f_{12}(\nu) := \max\{D(\mu^1\|\nu), D(\mu^2\|\nu)\}$. Since $f_{12}(u)$ is finite when u is the uniform distribution on Z we see that the value of the optimization problem in (6) is finite. It is also easy to see that without loss of optimality we can restrict the infimum to $\mathcal{P}_{12}(\mathsf{Z}) := \{\nu \in \mathcal{P}(\mathsf{Z}) : \text{supp}(\nu) \subseteq \text{supp}(\mu^1) \cup \text{supp}(\mu^2)\}$. This is because for any $\nu \in \mathcal{P}(\mathsf{Z})$ its restriction ν_{12} onto $\text{supp}(\mu^1) \cup \text{supp}(\mu^2)$ satisfies $D(\mu^1\|\nu_{12}) \le D(\mu^1\|\nu)$ and $D(\mu^2\|\nu_{12}) \le D(\mu^2\|\nu)$. Now $\mathcal{P}_{12}(\mathsf{Z})$ is a compact set, the function $f_{12}(.)$ is bounded below by 0 on $\mathcal{P}_{12}(\mathsf{Z})$, and moreover the function $f_{12}(.)$ is continuous in the relative interior of the set $\mathcal{P}_{12}(\mathsf{Z})$. Thus the infimum in $\inf_{\nu\in\mathcal{P}_{12}(\mathsf{Z})} f_{12}(\nu)$ is achieved since the optimal value is finite by the argument above.

Now (6) can be equivalently written as a convex problem:

$$\begin{aligned} \min_{\tau,\nu} \quad & \tau \\ \text{s.t.} \quad & D(\mu^1\|\nu) \le \tau, \quad D(\mu^2\|\nu) \le \tau, \\ & \sum_{x\in\mathsf{Z}} \nu(x) = 1, \quad \nu(x) \ge 0, \text{ for all } x \in \mathsf{Z}. \end{aligned}$$

Let $\hat{\nu}$ represent the optimizer of this problem. Considering the first order condition for optimality in a Lagrange-relaxed version of this problem it follows that there exists scalars ℓ_1, ℓ_2, and κ such that

$$\ell_1\mu^1(x) + \ell_2\mu^2(x) = \kappa\hat{\nu}(x), \text{ for all } x \in \mathbb{Z}$$

which implies that the optimizer $\hat{\nu}$ can be expressed as an affine combination of μ^1 and μ^2. Now by the definition of $f_{12}(.)$ it further follows that $\hat{\nu}$ can be expressed as a convex combination of μ^1 and μ^2. ∎

From the above lemma it follows that the test (2) can equivalently be written as

$$\phi_n^A(x,y) = \mathcal{I}\{D(\Gamma_{\lambda n}^x\|\alpha_n\Gamma_{\lambda n}^x + (1-\alpha_n)\Gamma_n^y) \ge \delta_n\} \qquad (7)$$

where $\alpha_n \in [0,1]$ satisfies

$$D(\Gamma_{\lambda n}^x\|\alpha_n\Gamma_{\lambda n}^x + (1-\alpha_n)\Gamma_n^y) = D(\Gamma_n^y\|\alpha_n\Gamma_{\lambda n}^x + (1-\alpha_n)\Gamma_n^y).$$

Furthermore, it is obvious that given $\Gamma_{\lambda n}^x$ and Γ_n^y the value of α_n can be easily computed by binary search since the function $g(\alpha) := D(\Gamma_{\lambda n}^x\|\alpha\Gamma_{\lambda n}^x + (1-\alpha)\Gamma_n^y) - D(\Gamma_n^y\|\alpha\Gamma_{\lambda n}^x + (1-\alpha)\Gamma_n^y)$ is a monotonically decreasing function of α, and α_n can be approximated by the value of α at which the function $g(.)$ is approximately zero. Now let $Z_n := D(\Gamma_{\lambda n}^x\|\alpha_n\Gamma_{\lambda n}^x + (1-\alpha_n)\Gamma_n^y)$. Thus the test of (2) is just a threshold test on Z_n. Although the test (2) looks complicated, the discussion above implies that the test statistic is in fact quite easy to compute.

IV. Weak convergence results

In the classical Neyman-Pearson hypothesis testing problem, one chooses the threshold that guarantees some bound on the false alarm probability of the test. Although it is not tractable to obtain an exact evaluation of the false alarm probability as a function of the threshold, we will now show that in the asymptotic regime, it is possible to obtain weak-convergence results on the test statistics that can be used to approximate the false alarm probability. All our results are based on the following basic lemma.

Lemma IV.1. *Suppose we are given a string x of observations of length λn and another independent string y of length n both drawn i.i.d. from the same distribution $\mu \in \mathcal{P}(\mathsf{Z})$ such that μ has full support over Z. Let $\Gamma_{\lambda n}^x$ denote the empirical distribution of the observations in x and Γ_n^y denote the empirical distribution of the observations in y. Let $h : \mathcal{P}(\mathsf{Z})\times\mathcal{P}(\mathsf{Z}) \mapsto \mathbb{R}$ be a continuous real-valued function whose gradient and Hessian are continuous in the neighborhood of (μ,μ). If the directional derivative satisfies $\nabla h(\mu,\mu)^T(\nu_1-\mu, \nu_2-\mu) = 0$ for all $\nu_1, \nu_2 \in \mathcal{P}(\mathsf{Z})$, then*

$$2n(h(\Gamma_{\lambda n}^x, \Gamma_n^y) - h(\mu,\mu)) \xrightarrow[n\to\infty]{d.} [W_\lambda^T, W^T]M \begin{bmatrix} W_\lambda \\ W \end{bmatrix}$$

where $M = \nabla^2 h(\mu,\mu)$ and W_λ and W are independent random vectors distributed as $W_\lambda \sim \mathcal{N}(0, \frac{\Sigma}{\lambda})$ and $W \sim \mathcal{N}(0,\Sigma)$ with $\Sigma = diag(\mu) - \mu\mu^T$.

Proof: Let $G_{n,x} := n^{\frac{1}{2}}(\Gamma_{\lambda n}^x - \mu)$ and $G_{n,y} := n^{\frac{1}{2}}(\Gamma_n^y - \mu)$. We know that $\Gamma_{\lambda n}^x$ and Γ_n^y can be written as empirical averages of i.i.d. vectors. Hence, they satisfy the central limit theorem which says that,

$$G_{n,x} = n^{\frac{1}{2}}(\Gamma_{\lambda n}^x - \mu) \xrightarrow[n\to\infty]{d.} W_\lambda \qquad (8)$$

$$G_{n,y} = n^{\frac{1}{2}}(\Gamma_n^y - \mu) \xrightarrow[n\to\infty]{d.} W \qquad (9)$$

where the distributions of W and W_λ are as defined in the statement of the lemma. Considering a second-order Taylor's expansion and using the condition on the directional derivative, we have, for n large enough,

$$\begin{aligned} & 2n(h(\Gamma_{\lambda n}^x, \Gamma_n^y) - h(\mu,\mu)) \\ &= [G_{n,x}^T, G_{n,y}^T]\nabla^2 h(\tilde{\Gamma}_n^x, \tilde{\Gamma}_n^y) \begin{bmatrix} G_{n,x} \\ G_{n,y} \end{bmatrix} \end{aligned}$$

where $\tilde{\Gamma}_n^x = \gamma\Gamma_{\lambda n}^x + (1-\gamma)\mu$ and $\tilde{\Gamma}_n^y = \gamma\Gamma_n^y + (1-\gamma)\mu$ for some $\gamma = \gamma(n) \in [0,1]$. We also know by the strong law of large numbers that $\Gamma_{\lambda n}^x$ and Γ_n^y and hence $\tilde{\Gamma}_n^x$ and $\tilde{\Gamma}_n^y$ converge to μ almost surely. By the continuity of the Hessian, we have

$$\nabla^2 h(\tilde{\Gamma}_n^x, \tilde{\Gamma}_n^y) \xrightarrow[n\to\infty]{a.s.} \nabla^2 h(\mu,\mu). \qquad (10)$$

By applying the vector-version of Slutsky's theorem [7], together with (8), (9) and (10), we conclude that

$$\begin{aligned} & [G_{n,x}^T, G_{n,y}^T]\nabla^2 h(\tilde{\Gamma}_n^x, \tilde{\Gamma}_n^y) \begin{bmatrix} G_{n,x} \\ G_{n,y} \end{bmatrix} \\ & \xrightarrow[n\to\infty]{d.} [W_\lambda^T, W^T]\nabla^2 h(\mu,\mu) \begin{bmatrix} W_\lambda \\ W \end{bmatrix}, \end{aligned}$$

thus establishing the lemma. ∎

As an immediate consequence of the above lemma we have the following result:

Lemma IV.2. *Suppose we are given a string x of observations of length λn and another independent string y of length n both drawn i.i.d. from the same distribution $\mu \in \mathcal{P}(\mathsf{Z})$ such that μ has full support over Z. Let $Y_1^n := D(\Gamma^x_{\lambda n} \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n))$ and $Y_2^n := D(\Gamma^y_n \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n))$. Then the following results hold:*

$$\frac{8n\lambda}{1+\lambda} Y_1^n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1}, \tag{11}$$

$$\frac{8n\lambda}{(1+\lambda)^2}(\lambda Y_1^n + Y_2^n) \xrightarrow[n\to\infty]{d.} \chi^2_{N-1}, \tag{12}$$

$$n(Y_1^n - Y_2^n) \xrightarrow[n\to\infty]{d.} 0. \tag{13}$$

Proof: To prove (11) we apply Lemma IV.1 to the function $h(\pi, \nu) := D(\pi \| \frac{1}{2}(\pi + \nu))$. It is easily verified that the gradient and Hessian satisfy the necessary regularity conditions. Computing the Hessian at (μ, μ) we obtain

$$M = \begin{bmatrix} \text{diag}\left(\frac{1}{4\mu}\right) & -\text{diag}\left(\frac{1}{4\mu}\right) \\ -\text{diag}\left(\frac{1}{4\mu}\right) & \text{diag}\left(\frac{1}{4\mu}\right) \end{bmatrix}$$

where $\text{diag}\left(\frac{1}{4\mu}\right)$ denotes a diagonal matrix with the i-th diagonal entry given by $\frac{1}{4\mu_i}$. Applying the conclusion of Lemma IV.1 we obtain

$$2nY_1^n \xrightarrow[n\to\infty]{d.} (W_\lambda - W)^T \text{diag}\left(\frac{1}{4\mu}\right)(W_\lambda - W).$$

Equivalently we can write $\frac{8n\lambda}{1+\lambda} Y_1^n \xrightarrow[n\to\infty]{d.} W^T \text{diag}\left(\frac{1}{\mu}\right) W$. It can be shown using the result of [8, Lemma III.7] that $W^T \text{diag}\left(\frac{1}{\mu}\right) W$ has a χ^2_{N-1} distribution thus proving (11).

Similarly for proving (12) we apply Lemma IV.1 to the function $h(\pi, \nu) := \lambda D(\pi \| \frac{1}{2}(\pi + \nu)) + D(\pi \| \frac{1}{2}(\pi + \nu))$. Computing the Hessian at (μ, μ) we see that the new Hessian is just $(1 + \lambda)$ times M. Thus the result of (12) by a similar argument as before.

Now if we apply Lemma IV.1 to the function $h(\pi, \nu) := D(\pi \| \frac{1}{2}(\pi + \nu)) - D(\nu \| \frac{1}{2}(\pi + \nu))$, we see that the Hessian at (μ, μ) vanishes. Hence (13) follows. ∎

We are now ready to obtain the weak convergence behavior of the test statistic Z_n used in the test of (7).

Proposition IV.3. *Assume that the data strings x^m and y^n are drawn i.i.d. according to some fixed distribution $\mu \in \mathcal{P}(\mathsf{Z})$ such that μ has full support on Z. Further assume that m grows linearly in n as $m = \lambda n$. Let α_n and Z_n be as before with $Z_n = D(\Gamma^x_{\lambda n} \| \alpha_n \Gamma^x_{\lambda n} + (1 - \alpha_n)\Gamma^y_n)$. Then if m grows linearly in n as $m = \lambda n$, we have*

$$\frac{8n\lambda}{1+\lambda} Z_n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \tag{14}$$

Proof: Let $Y_1^n := D(\Gamma^x_{\lambda n} \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n))$ and $Y_2^n := D(\Gamma^y_n \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n))$. From Lemma III.1 we have

$$\min\{Y_1^n, Y_2^n\} \le Z_n \le \max\{Y_1^n, Y_2^n\}. \tag{15}$$

Now if $W_n := \min\{Y_1^n, Y_2^n\}$, then we have $|W_n - Y_1^n| \le |Y_1^n - Y_2^n|$. Hence by (13) we have $n(W_n - Y_1^n) \xrightarrow[n\to\infty]{d.} 0$. Combining with (11) we get $\frac{8n\lambda}{1+\lambda} W_n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1}$. By a similar argument it also follows that $V_n := \max\{Y_1^n, Y_2^n\}$ satisfies $\frac{8n\lambda}{1+\lambda} V_n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1}$. Thus by (15) we see that nZ_n is sandwiched between two random quantities having the same weak convergence behavior. Thus nZ_n should also have the same weak convergence limit. ∎

We now consider the test of (4) proposed in [2]. The test statistic in this test can be expressed as:

$$Y_n := \lambda D(\Gamma^x_{\lambda n} \| \tfrac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n)) + D(\Gamma^y_n \| \tfrac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n)).$$

In the following theorem we characterize the limiting behavior of this test statistic.

Proposition IV.4. *Assume that the data strings x^m and y^n are drawn i.i.d. according to some fixed distribution $\mu \in \mathcal{P}(\mathsf{Z})$ such that μ has full support on Z. Let $Y_n := \lambda D(\Gamma^x_{\lambda n} \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n)) + D(\Gamma^y_n \| \frac{1}{2}(\Gamma^x_{\lambda n} + \Gamma^y_n))$. Then we have*

$$\frac{8n\lambda}{(1+\lambda)^2} Y_n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \tag{16}$$

Proof: This is exactly the result of (12) in Lemma IV.2. ∎

Similarly, we can also identify the limiting behavior of the chi-square test statistic used in (5) via the results of Lemma IV.1. Although this result is well known in statistics literature, we provide a simple proof for completeness.

Proposition IV.5. *Assume that the data strings x^m and y^n are drawn i.i.d. according to some fixed distribution $\mu \in \mathcal{P}(\mathsf{Z})$ such that μ has full support on Z. Let $X_n := \chi^2(\Gamma^x_{\lambda n}, \Gamma^y_n)$. Then we have*

$$\frac{n\lambda}{1+\lambda} X_n \xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \tag{17}$$

Proof: We apply Lemma IV.1 to the function $f(\pi, \nu) = \chi^2(\pi, \nu)$. It is easily verified that the gradient and Hessian satisfy the necessary regularity conditions. Computing the Hessian at (μ, μ) we obtain

$$M = \begin{bmatrix} \text{diag}\left(\frac{2}{\mu}\right) & -\text{diag}\left(\frac{2}{\mu}\right) \\ -\text{diag}\left(\frac{2}{\mu}\right) & \text{diag}\left(\frac{2}{\mu}\right) \end{bmatrix}.$$

Following the same steps as in the proof of (11) in Lemma IV.2, the conclusion follows. ∎

We observe from Propositions IV.3, IV.4 and IV.5 that the limiting distribution of the test statistics of all the three tests ϕ^A of (7), ϕ^B of (4) and ϕ^C of (5) under the null hypothesis depend only on the support size of the true distribution μ and not on the specific value of μ. In the following section we discuss how these weak convergence results can be used to select the test thresholds for a target false alarm probability.

V. Approximate thresholds

The weak convergence behavior of the test statistics in the three tests we have considered can be used to approximately choose the test threshold for a target false alarm probability. For example in the chi-square test ϕ^C of (5) if under the null

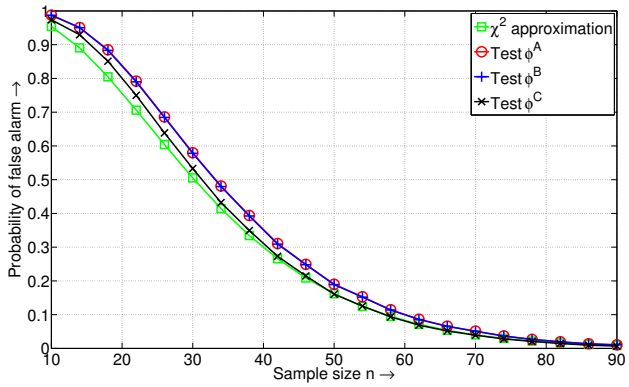

Fig. 1. False alarm probabilities of the various tests are shown along with the χ^2 approximation of these error probabilities obtained using the weak convergence result.

hypothesis the observations are drawn from some distribution $\mu \in \mathcal{P}(\mathsf{Z})$ with full support, then the test statistic X_n satisfies

$$\lim_{n\to\infty} \mathsf{P}_\mu\{\frac{n\lambda}{(1+\lambda)^2} X_n > x\} = 1 - F(x)$$

where $F(x)$ denotes the cdf of a chi-square random variable with $N-1$ degrees of freedom. This relation can be used to approximate the threshold to be used in (5) for a target false alarm probability, by approximating the true probability with the limiting probability. Similarly, the thresholds for the optimal tests ϕ^A of (7) and ϕ^B of (4) can be chosen using the weak convergence of their respective test statistics.

In order to estimate the accuracy of the approximation obtained from the weak convergence, we simulated the three tests using a uniform distribution over an alphabet of size 8 for μ. In Figure 1 we have plotted the false alarm probabilities of the three tests as a function of the sequence length n obtained by simulations. In the same figure we also have a plot of the approximate value of the false alarm probability computed using the weak convergence approximation suggested in the previous paragraph. Clearly, that the error predictions obtained via the weak-convergence approximations are quite accurate for values of n greater than 45.

VI. Summary and Future Work

We have studied the homogeneity testing problem for multinomial distributions. Although optimal results have been proposed for this problem in information theory literature, such results are not well-known among statisticians and such tests are rarely used in practice. In this paper, we have simplified the structure of one of these tests and also identified the limiting behavior of the test statistics used in both the tests. These results can be used to approximate the thresholds for these tests. Such homogeneity tests with provable optimality properties could potentially be better choices than the chi-square test in practice.

In terms of future work it would be of interest to identify the optimal tests for this problem in the setting in which the alphabet size is allowed to increase with the sample size. Such a setup is relevant in problems involving data from continuous alphabet distributions which could be first quantized and then tested. The literature (see, e.g., [9] and references therein) on the simpler single sample goodness-of-fit problem could be a good starting point for such an investigation. A different direction for future work is to extend these results on two-sample tests to more general k-sample tests and to evaluate the asymptotic efficiency of these tests.

Acknowledgements

The author thanks Dayu Huang for useful discussions. This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006.

References

[1] O. Shayevitz, "On Rényi measures and hypothesis testing," in *Proc. of International Symposium on Information Theory (ISIT 2011)*, St. Petersburg, 2011.

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

[3] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, 3rd ed., ser. Springer Texts in Statistics. New York: Springer, 2005.

[4] P. Greenwood and M. Nikulin, *A guide to chi-squared testing*, ser. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. New York: John Wiley & Sons, 1996.

[5] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, no. 1/2, pp. 250–254, 1911. [Online]. Available: http://www.jstor.org/stable/2331453

[6] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.

[7] P. Billingsley, *Convergence of Probability Measures*. New York: John Wiley & Sons, 1968.

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

[9] P. Harremoes and I. Vajda, "On the Bahadur-Efficient Testing of Uniformity by Means of the Entropy," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 321 –331, Jan. 2008.