Weak Convergence Analysis of Asymptotically Optimal Hypothesis Tests

Jayakrishnan Unnikrishnan 1 and Dayu Huang 2

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

2 University of Illinois at Urbana-Champaign, Illinois, United States, Email: dayuhuang@gmail.com

Abstract

In recent years solutions to various hypothesis testing problems in the asymptotic setting have been proposed using results from large deviations theory. Such tests are optimal in terms of appropriately defined error-exponents. For the practitioner, however, error probabilities in the finite sample size setting are more important. In this paper we show how results on weak convergence of the test statistic can be used to obtain better approximations for the error probabilities in the finite sample size setting. While this technique is popular among statisticians for common tests, we demonstrate its applicability for several recently proposed asymptotically optimal tests, including tests for robust goodness of fit, homogeneity tests, outlier hypothesis testing, and graphical model estimation.

I. INTRODUCTION

In hypothesis testing problems involving unknown distributions, one often has to rely on asymptotic notions for defining the optimality of a testing procedure. A simple example is the universal hypothesis testing problem, or the goodness of fit problem, wherein one seeks to test the hypothesis that a given set of observations is drawn according to a known probability law defined on a finite alphabet. In this problem one has to define the notion of an error exponent to be able to identify an optimal test. In this particular case, the test proposed by Hoeffding [3] is known to be optimal in an error exponent sense. Numerous
other similar examples of asymptotically optimal hypothesis tests are found in literature (see, e.g., [3]–[8]). Nevertheless, in a practical experiment, one has access only to a finite number of observations, and the metric of practical interest is the actual error probability with a finite number of samples rather than the error exponent.

In practice, statisticians typically use results on weak convergence of the test statistic to approximate the error probabilities of Hoeffding’s test. In this paper, we present a general approach for deriving such weak convergence results and demonstrate the usefulness of the technique by applying it to a wide range of asymptotically optimal hypothesis tests that have been recently proposed in the information theory literature. We consider problems in goodness of fit, robust goodness of fit, homogeneity testing, and a few recently proposed tests involving multiple hypotheses such as outlier hypothesis testing, and graphical model estimation. In all these examples, we show how the error probability can be approximately computed using the weak convergence result. These results can also be interpreted as formulas for evaluating the p-values [9] for the different tests. The p-value of a test is defined as the probability of obtaining a test statistic at least as extreme as the one that is actually observed, assuming that the null hypothesis is true.

Our results are based on a Taylor series expansion of the test statistic together with the central limit theorem applied to the empirical distribution of the observations. Similar techniques are used by statisticians [10] and have been known at least since Wilks’ original paper [11] in which the limiting distribution for composite likelihood ratio tests was identified. Similar results were also obtained in [12] for tests on exponential families of distributions. In our past work [13] we demonstrated the usefulness of this technique for choosing thresholds in a mismatched version of Hoeffding’s test. However, such results are rarely discussed in information theory literature and numerous recent works [4]–[7] on optimal tests fail to address the important issue of selecting thresholds for finite sample size experiments. We use these tests as examples to demonstrate how the weak convergence method can be applied to provide approximate guarantees for finite sample size problems. The accuracy of the approximations are demonstrated via simulations.

The rest of the paper is organized as follows. The main body of the paper is in Section II, where we first present the generic technique and then demonstrate how it can be applied to the various problems considered. In Section III we present some simulation results demonstrating the accuracy of the approximations. We conclude in Section IV. Below we summarize the notations we use in this paper.

**Notation:** For a finite set $Z$ we use $|Z|$ to denote its cardinality and $\mathcal{P}(Z)$ to denote the set of probability
distributions on $Z$. We identify probability measures $\pi \in \mathcal{P}(Z)$ also as vectors in the probability simplex in Euclidean space $\mathbb{R}^{\lvert Z \rvert}$. We use $\Gamma_n \in \mathcal{P}(Z)$ to denote the empirical distribution or *type* of a finite set of observations $(Z_1, Z_2, \ldots, Z_n)$:

$$
\Gamma_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i = z\}, \quad z \in Z
$$

where $\mathbb{I}$ denotes the indicator function. For hypothesis $H$ we use $P_H$ to denote the probability measure under hypothesis $H$ and for probability distribution $\pi \in \mathcal{P}(Z)$ we use $P_\pi$ to denote the probability measure when observations $Z_1, Z_2, \ldots$ are drawn i.i.d. under distribution $\pi$.

For any function $f : Z \mapsto \mathbb{R}$ and for any distribution $\pi \in \mathcal{P}(Z)$, we use the following notation for the expected value of function $f$:

$$
\pi(f) := \sum_{z \in Z} \pi(z) f(z).
$$

Note that for $z \in Z$, $\pi(z)$ denotes the probability mass at $z$ under measure $\pi$. The meaning will be clear from context.

For a sequence of random variables $X_1, X_2, \ldots$, we use $X_n \xrightarrow{d.} X$ to indicate weak convergence (i.e., convergence in distribution) to a random variable $X$. We use $F_d(\eta)$ to denote the cumulative density function (CDF) of a chi-square random variable with $d$ degrees of freedom evaluated at $\eta$.

II. WEAK CONVERGENCE RESULTS AND ERROR PROBABILITY ESTIMATES

All the results in this paper are concerned with variations of the following problem. Consider a hypothesis testing problem wherein one seeks to decide among a set of two or more hypotheses based on a sequence of observations of the form $Z = Z_1, Z_2, \ldots$ drawn from a set $Z$. For this problem let $\hat{H}_n$ be the output of a sequence of tests that decide in favor of or against a particular hypothesis $H$ based on the first $n$ observations in the sequence. Suppose $\hat{H}_n$ can be expressed by the rule

$$
\hat{H}_n = H \iff f_n(Z_1, Z_2, \ldots, Z_n) \leq \tau.
$$

where $f_n(.)$ is some function of the observations $Z_1, Z_2, \ldots, Z_n$ and $\tau$ is a threshold. Under hypothesis $H$, the error probability of this test sequence as a function of $n$ is defined by

$$
\beta_n := P_H\{\hat{H} \neq H\}
$$

$$
= P_H\{f_n(Z_1, Z_2, \ldots, Z_n) > \tau\}.
$$

Asymptotically optimal tests (e.g., [3]–[8]) typically guarantee inequalities on the error exponent, i.e., inequalities of the form

$$
\liminf_{n \to \infty} -\frac{1}{n} \log \beta_n \geq \eta.
$$
A simple example is the Hoeffding test for universal hypothesis testing. In this case under the null hypothesis $H$ the observations in $Z$ are drawn i.i.d. from a known distribution $π^0$. The test is given by

$$\hat{H}_n = H \iff D(Γ_n∥π^0) ≤ τ.$$  

(7)

Error exponent guarantees of the form (6) are based on large deviations analysis [14] applied to the test statistic $T_n := f_n(Z_1, Z_2, \ldots, Z_n)$. For instance, Sanov’s theorem implies that Hoeffding’s test satisfies (6) with $η = τ$. In this paper, we are concerned with alternate approaches to estimate the error probability based on a weak convergence analysis of the test statistic. In particular we will establish results of the form

$$g(n)T_n \xrightarrow{\text{d.}} T, \quad \text{under hypothesis } H$$

where $g(n)$ is some non-negative function of $n$ and $T$ is a random variable with a known distribution. We will then use this result to approximate the error probability as

$$β_n = \mathbb{P}_H\{g(n)T_n > g(n)τ\} \approx \mathbb{P}\{T > g(n)τ\}.$$  

(8)

We will also see through simulations that such approximations give accurate predictions of the actual error probabilities. Variations of such techniques are popularly used by statisticians [10] for common tests. For the ease of illustration, we will first discuss the application of the technique to Hoeffding’s test, and then proceed to other recently proposed asymptotically optimal tests.

The main tool we use in this paper is a result on weak convergence of functions of empirical distributions of i.i.d. sequences contained in the following lemma.

**Lemma II.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 $μ ∈ \mathcal{P}(Z)$ such that $μ$ has full support over $Z$. Let $Γ^x_{λn}$ denote the empirical distribution of the observations in $x$ and $Γ^y_n$ denote the empirical distribution of the observations in $y$. Let $g : \mathcal{P}(Z) \mapsto \mathbb{R}$ and $h : \mathcal{P}(Z) \times \mathcal{P}(Z) \mapsto \mathbb{R}$ be continuous real-valued functions whose gradients and Hessians are continuous in the neighborhood of $μ$ and $(μ, μ)$ respectively. If the directional derivatives satisfy $∇g(μ)^T(ν_1 − μ) = 0$ and $∇h(μ, μ)^T(ν_1 − μ, ν_2 − μ) = 0$
for all $\nu_1, \nu_2 \in \mathcal{P}(Z)$, then

$$2n(g(\Gamma_{n}^\nu) - g(\mu)) \xrightarrow{d, n \to \infty} W^TM_g W$$

(9)

$$2n(h(\Gamma_{n}^\nu, \Gamma_{n}^\mu) - h(\mu, \mu)) \xrightarrow{d, n \to \infty} [W^T_{\lambda}, W^T_{\mu}]M_h \begin{bmatrix} W_{\lambda} \\ W \end{bmatrix}$$

(10)

where $M_g = \nabla^2 g(\mu)$, $M_h = \nabla^2 h(\mu, \mu)$ and $W_{\lambda}$ and $W$ are independent random vectors distributed as $W_{\lambda} \sim N(0, \Sigma_{\lambda})$ and $W \sim N(0, \Sigma)$ with $\Sigma = \text{diag}(\mu) - \mu\mu^T$.

Later in the paper we also need the following result.

**Lemma II.2.** Suppose that $V$ is an $m$-dimensional, $\mathcal{N}(0, I_m)$ random vector, and $D : \mathbb{R}^m \to \mathbb{R}^m$ is a rank $K$ matrix such that $D^TD$ is a projection matrix, i.e.,

$$D^TD = D^TD.$$

Then $\xi := ||DV||^2$ is a chi-square random variable with $K$ degrees of freedom.

Proofs for both the above lemmas are presented in the appendix. Below we consider several specific hypothesis tests that can be analyzed using these lemmas.

**A. Goodness of fit tests**

Consider the problem of universal hypothesis testing. We are given a sequence of observations $Z = Z_1, Z_2, \ldots$ from a finite alphabet $Z$ and we are interested in testing the null hypothesis that all the observations were drawn i.i.d. according to some known probability law $\pi^0$ belonging to the set $\mathcal{P}(Z)$ of all probability measures on $Z$. For this problem, Hoeffding [3] proposed the following generalized likelihood ratio test sequence (see also [13])

$$\phi_n^H = \mathbb{I}\{D(\Gamma_n||\pi^0) \geq \tau\}$$

(11)

where $\phi_n^H = 0$ indicates a decision in favor of the null hypothesis. This test sequence is known to be optimal in the sense of error exponents. Let $\phi := (\phi_1, \phi_2, \ldots)$ be any sequence of tests such that $\phi_n : Z^n \mapsto \{0, 1\}$. The error exponents under such test sequence are given by

$$J_0^H := \lim_{n \to \infty} \inf -\frac{1}{n} \log(P_{\pi^0}\{\phi_n(Z_1, Z_2, \ldots, Z_n) = 1\}),$$

$$J_1^H := \lim_{n \to \infty} \inf -\frac{1}{n} \log(P_{\pi^1}\{\phi_n(Z_1, Z_2, \ldots, Z_n) = 0\}).$$

(12)
Then the Hoeffding test sequence of (11) solves the following problem:

$$\sup_{\phi} \{ J^1_\phi : \text{subject to } J^0_\phi \geq \tau \}. \quad (13)$$

Although this is an optimality result in terms of error exponents, the error exponent does not give a good estimate for the false alarm probability, i.e., error probability under the null hypothesis:

$$P_{\pi^n}\{\phi_n(Z_1, Z_2, \ldots, Z_n) = 1 \}.$$  

Practitioners are interested in knowing how to choose $\tau$ for a target false alarm probability. A potential solution is to try to estimate the error probabilities by simulating the test. However, a quicker method for an approximate answer (see [13]) is obtained by studying the weak-convergence behavior of the test statistic $D(\Gamma_n\|\pi^0)$ under the null hypothesis. It is known that [11] under the null hypothesis we have

$$nD(\Gamma_n\|\pi^0) \xrightarrow{n \to \infty} \frac{1}{2} \chi^2_d |Z| - 1 \quad (14)$$

where $\chi^2_d$ denotes the chi-square distribution with $d$ degrees of freedom. This result can be obtained by applying the conclusion of (9) to the function $g(\pi) := D(\pi\|\pi^0)$ with $\mu = \pi^0$. This convergence result can be used to approximate the distribution of the test statistic for finite $n$, which can then be used to approximately compute the error probabilities as in (8).

**B. Robust goodness of fit tests**

One of the commonly faced problems in using goodness of fit tests is that often the underlying probability law may not be known accurately. This can lead to poor error performances in practice. A possible solution to such problems is to use a robust version of the goodness of fit test, wherein we expand the null hypothesis to include a wider class of distributions in place of a single distribution. In such robust goodness of fit problems, one is again faced with the problem of how to choose the threshold for a target false alarm probability under the null hypothesis. However, unlike in the ordinary goodness of fit problem, it is not possible to estimate the threshold by simulations since the null hypothesis is composite. In such settings, weak convergence results offer a quick method for an approximate solution. Below we consider a robust version of the Hoeffding test of (11). Similar approximate bounds can also be provided for a robust version of the Kolmogorov-Smirnov test [2], which is a goodness of fit test for continuous alphabets.

Let $Z := Z_1, Z_2, \ldots$ denote a i.i.d. sequence of observations from finite alphabet $Z$. A robust goodness of fit test was studied in [4] for testing a composite null hypothesis of the form

$$\mathcal{H}^{\text{rob}}_0 : Z_i \sim \pi \text{ for some } \pi \in \mathbb{P}$$
where
\[ P := \{ \pi \in \mathcal{P}(Z) : \pi(\psi_i) = 0, \ 1 \leq i \leq d \} \tag{15} \]

with \( \pi(\psi_i) \) defined as in (2). They studied a sequence of tests of the form
\[ \phi^\text{ROB}_n(Z) = \mathbb{I}\{ D^\text{ROB}(\Gamma_n||P) > \tau \} \tag{16} \]

where \( \phi^\text{ROB}_n = 0 \) indicates a decision in favor of the null hypothesis. The function \( D^\text{ROB} \) is a robust version of the divergence defined as
\[ D^\text{ROB}(\mu||P) := \inf_{\pi \in P} D(\mu||\pi). \]

This test sequence was shown to be optimal in solving the following error-exponent optimization problem:
\[ \sup_{\phi} \{ J_1^\phi : \text{subject to } J_0^\phi \geq \tau, \ \forall \pi^0 \in P \} \tag{17} \]

where the error exponents are as defined in (12). As before, the performance guarantees of this test are expressed in terms of error exponents. From a practitioner’s perspective, however, the exact error probabilities are more important. For this purpose, we study the weak convergence behavior of the statistic \( D^\text{ROB}(\Gamma_n||P) \) under distributions from the null hypothesis, \( P \). Specifically, we evaluate the limiting value of
\[ \mathbb{P}_\pi\{ nD^\text{ROB}(\Gamma_n||P) > \eta \} \tag{18} \]

for \( \pi \in P \).

In Theorem II.3 we answer this question under the following assumptions on the functions \( \psi_i \) used in the definition of \( P \) in (15). Let \( \psi \) denote the vector of functions \( (\psi_1, \psi_2, \ldots, \psi_d)^T \) and \( Z_\pi \) denote the support of distribution \( \pi \).

(A1) There is some distribution \( \pi^0 \in P \) such that the functions \( \{ \psi_i : 0 \leq i \leq d \} \) are linearly independent over \( Z_{\pi^0} \), where \( \psi_0 \equiv 1 \).

(A2) The origin \( 0 \in \mathbb{R}^d \) is an interior point of the set of feasible moment vectors, defined as
\[ \Delta := \{ x \in \mathbb{R}^d : x_i = \mu(\psi_i), i = 1, \ldots, d, \ \text{for some } \mu \in \mathcal{P}(Z) \}. \]

The main result of this section is the following theorem: For any \( \pi \in P \), let \( d_\pi \) denote the number that is one lower than the maximal number of functions in \( \{ \psi_i : 0 \leq i \leq d \} \) that are linearly independent over \( Z_\pi \). In other words, \( d_\pi + 1 \) is the dimension of the span of the functions \( \{ \psi_i : 0 \leq i \leq d \} \) when restricted to \( Z_\pi \).
Theorem II.3. Suppose assumptions (A1) and (A2) hold. Then, the following weak convergence result holds under \( \pi \):

\[
2nD^{\text{ROB}}(\Gamma_n \| \mathbb{P}) \xrightarrow{d} \chi^2_d.
\]

Hence we have

\[
\sup_{\pi \in \mathbb{P}} \lim_{n \to \infty} P_\pi \{2nD^{\text{ROB}}(\Gamma_n \| \mathbb{P}) > \eta\} = 1 - F_d(\eta)
\]

where \( F_d(\eta) \) is the CDF of a chi-square distribution with \( d \) degrees of freedom evaluated at \( \eta \).

The second conclusion in the theorem above follows directly from the first, using the fact that chi-square distributions are stochastically ordered according to the number of degrees of freedom. We provide a proof in the appendix.

C. Homogeneity testing

Two sample homogeneity testing refers to the problem of determining whether or not two strings of data follow the same distribution. Consider for instance two strings of data \( X = (X_1, X_2, \ldots, X_m) \) and \( Y = (Y_1, Y_2, \ldots, Y_n) \) drawn i.i.d. according to some unknown distributions on a finite alphabet \( Z \). The objective is to determine whether or not both strings are drawn from identical distributions in \( \mathcal{P}(Z) \). As before we work in the regime where \( m \) and \( n \) are linearly related as \( m = \lambda n \). We use \( \Gamma^n_\lambda \) and \( \Gamma^n_\eta \) respectively to denote the empirical distributions of \( X \) and \( Y \). Let \( \phi_n : Z^{\lambda n} \times Z^n \mapsto \{0,1\} \) represent a sequence of tests with a test outcome of \( \phi_n(X, Y) = 0 \) denoting a decision in favor of the null hypothesis that \( X \) and \( Y \) are drawn from the same distributions and \( \phi_n(X, Y) = 1 \) denoting a decision in favor of the alternate hypothesis that \( X \) and \( Y \) are drawn from different distributions. We define two kinds of error probabilities and corresponding error exponents - for false alarms and missed detections. For any distribution \( \mu \in \mathcal{P}(Z) \) of the observations under the null hypothesis, the probability of false alarm is given by

\[
p_{FA}(\phi_n | \mu) = P_\mu(\phi_n(X, Y) = 1).
\]

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

\[
p_{MD}(\phi_n | \pi^1, \pi^2) = P_{\pi^1, \pi^2}(\phi_n(X, Y) = 0).
\]

The corresponding false alarm error-exponent and missed-detection exponent are defined respectively as

\[
E_{FA}(\phi | \mu) := \liminf_{n \to \infty} -\frac{1}{n} \log p_{FA}(\phi_n | \mu)
\]

\[
E_{MD}(\phi | \pi^1, \pi^2) := \liminf_{n \to \infty} -\frac{1}{n} \log p_{MD}(\phi_n | \pi^1, \pi^2).
\]
Two versions of asymptotically optimal tests are known in literature. The following sequence of tests were first proposed in [5].

\[
\phi_n^A(X, Y) = \mathbb{I}\left\{ \inf_{\mu \in \mathcal{P}(Z)} \max\{D(\Gamma_n^x||\mu), D(\Gamma_n^y||\mu)\} \geq \delta_n \right\}
\]  

(21)

where \( \delta_n = \frac{|Z| \log n}{n} \). It was shown in [5] that this test sequence solves the following optimization problem:

\[
\sup_{\phi} E_{MD}(\phi|\pi^1, \pi^2)
\]

s.t.

\[
\lim_{n \to \infty} p_{FA}(\phi_n|\mu) = 0 \text{ for all } \mu \in \mathcal{P}(Z).
\]

(22)

However, the structure of the test can be simplified using the following lemma.

**Lemma II.4.** For any distributions \( \mu^1, \mu^2 \in \mathcal{P}(Z) \), the infimum in

\[
\inf_{\nu \in \mathcal{P}(Z)} \max\{D(\mu^1||\nu), D(\mu^2||\nu)\}
\]

(23)

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

A proof is provided in the appendix. The above result implies that the test of (21) can be simplified to the following form:

\[
\phi_n^A(X, Y) = \mathbb{I}\left\{ D(\Gamma_n^x||\alpha_n \Gamma_n^x + (1 - \alpha_n) \Gamma_n^y) \geq \delta_n \right\}
\]

(24)

where \( \alpha_n \in [0, 1] \) satisfies

\[
D(\Gamma_n^x||\alpha_n \Gamma_n^x + (1 - \alpha_n) \Gamma_n^y) = D(\Gamma_n^y||\alpha_n \Gamma_n^x + (1 - \alpha_n) \Gamma_n^y).
\]

(25)

Another optimal test was proposed in [6]. The test sequence is given by

\[
\phi_n^B(X, Y) = \mathbb{I}\left\{ \lambda D(\Gamma_n^x||\frac{1}{1 + \lambda}(\lambda \Gamma_n^x + \Gamma_n^y)) + D(\Gamma_n^y||\frac{1}{1 + \lambda}(\lambda \Gamma_n^x + \Gamma_n^y)) \geq \tilde{\delta}_n \right\}
\]

(26)

where \( \tilde{\delta}_n = \eta + O\left(\frac{\log n}{n}\right) \) was shown to be optimal for the following problem

\[
\sup_{\phi} E_{MD}(\phi|\pi^1, \pi^2)
\]

s.t.

\[
E_{FA}(\phi_n|\mu) \geq \eta \text{ for all } \mu \in \mathcal{P}(Z).
\]

(27)

A third possible solution to the homogeneity testing problem is given by the chi-square test that is used commonly by statisticians. The chi-square distance between two distributions is defined as

\[
\chi^2(\pi, \nu) := \sum_{z \in Z} \frac{2(\pi(z) - \nu(z))^2}{\pi(z) + \nu(z)}, \quad \pi, \nu \in \mathcal{P}(Z).
\]
The two sample chi-square test is given by

$$
\phi_n^C(X, Y) = \mathbb{I}\left\{ \chi^2(\Gamma_{\hat{\delta}_n}^x, \Gamma_{\hat{\delta}_n}^y) \geq \hat{\delta}_n \right\}
$$

(28)

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

We now determine the weak convergence behaviour of the test statistics used in the tests described above. These results are based on the following lemma which follows from Lemma II.1.

**Lemma II.5.** Suppose we are given a string \( X \) of observations of length \( \lambda n \) and another independent string \( Y \) of length \( n \) both drawn i.i.d. from the same distribution \( \mu \in \mathcal{P}(Z) \) such that \( \mu \) has full support over \( Z \). Let \( Z_1^n := D(\Gamma_{\lambda n}^x \parallel \frac{1}{2}(\Gamma_{\lambda n}^x + \Gamma_{\lambda n}^y)) \), \( Z_2^n := D(\Gamma_{\lambda n}^y \parallel \frac{1}{2}(\Gamma_{\lambda n}^x + \Gamma_{\lambda n}^y)) \), \( Y_1^n := D(\Gamma_{\lambda n}^x \parallel \frac{1}{1+\lambda} \Gamma_{\lambda n}^x + \Gamma_{\lambda n}^y) \) and \( Y_2^n := D(\Gamma_{\lambda n}^y \parallel \frac{1}{1+\lambda} \Gamma_{\lambda n}^x + \Gamma_{\lambda n}^y) \). Then the following results hold:

$$
\begin{align*}
\frac{8n\lambda}{1+\lambda} Z_1^n & \xrightarrow{d, n \to \infty} \chi^2_{|Z|-1}, \\
\frac{8n\lambda}{1+\lambda} Z_2^n & \xrightarrow{d, n \to \infty} \chi^2_{|Z|-1}, \\
2n(1 + \lambda)Y_1^n & \xrightarrow{d, n \to \infty} \chi^2_{|Z|-1}, \\
2n(Y_1^n + Y_2^n) & \xrightarrow{d, n \to \infty} \chi^2_{|Z|-1}.
\end{align*}
$$

(29) (30) (31) (32)

**Proof:** To prove (29) we apply Lemma II.1 to the function \( h(\pi, \nu) := D(\pi \parallel \frac{1}{2}(\pi + \nu)) \). It is easy to verified that the gradient and Hessian satisfy the necessary regularity conditions. Computing the Hessian at \( (\mu, \mu) \) we obtain

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

(33)

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 II.1 we obtain

$$
2n Z_1^n \xrightarrow{d, n \to \infty} (W_{\lambda} - W)^T \text{diag} \left( \frac{1}{4\mu} \right) (W_{\lambda} - W).
$$

Equivalently we can write \( \frac{8n\lambda}{1+\lambda} Z_1^n \xrightarrow{d, n \to \infty} W^T \text{diag} \left( \frac{1}{\mu} \right) W \).

Now \( W^T \text{diag} \left( \frac{1}{\mu} \right) W \) can be written as \( \|DV\|^2 \) where \( D = \text{diag} \left( \frac{1}{\mu} \right)^{\frac{1}{2}} \Sigma^{\frac{1}{2}} \) where \( \Sigma = \text{diag}(\mu) - \mu \mu^T \) as in Lemma II.1 and \( V \sim \mathcal{N}(0, I_{|Z|}) \). It can easily be verified that \( D \) satisfies the conditions of Lemma II.2 and hence the conclusion follows by applying Lemma II.2.

Now if we apply Lemma II.1 to the function \( h(\pi, \nu) := D(\pi \parallel \frac{1}{2}(\pi + \nu)) - D(\nu \parallel \frac{1}{2}(\pi + \nu)) \), we see that the Hessian at \( (\mu, \mu) \) vanishes. Hence (31) follows.
To prove (31) we follow the same steps as before and apply Lemma II.1 to the function \( h(\pi, \nu) := D(\pi \| \frac{1}{1+\lambda}(\lambda \pi + \nu)) \). It is easily verified that the gradient and Hessian satisfy the necessary regularity conditions. Computing the Hessian at \((\mu, \mu)\) we obtain
\[
M' = \begin{bmatrix}
\text{diag} \left( \frac{1}{1+\lambda} \mu \right) & -\text{diag} \left( \frac{1}{1+\lambda} \mu \right) \\
-\text{diag} \left( \frac{1}{1+\lambda} \mu \right) & \text{diag} \left( \frac{1}{1+\lambda} \mu \right)
\end{bmatrix}
\]
which is just a scaled version of \( M \) in (33). Thus, by following the same steps as in the proof of (29) we get (31).

Similarly for proving (32) we apply Lemma II.1 to the function \( h(\pi, \nu) := \lambda D(\pi \| \frac{1}{1+\lambda}(\lambda \pi + \nu)) + D(\nu \| \frac{1}{1+\lambda}(\lambda \pi + \nu)) \). Computing the Hessian at \((\mu, \mu)\) we see that the new Hessian is just \((\lambda + \lambda^2)\) times \( M' \). Thus the result of (32) follows by a similar argument as before.

Using the conclusion of the above lemma we can identify the weak convergence behaviour of the two sample tests discussed above.

**Theorem II.6.** Assume that the data strings \( X \) of length \( m \) and \( Y \) of length \( n \) are drawn i.i.d. according to some fixed distribution \( \mu \in \mathcal{P}(\mathbb{Z}) \) such that \( \mu \) has full support on \( \mathbb{Z} \). Further assume that \( m \) grows linearly in \( n \) as \( m = \lambda n \). Let \( Z_n = D(\Gamma_{\lambda n}^{x} \| \alpha_{n} \Gamma_{\lambda n}^{x} + (1 - \alpha_{n}) \Gamma_{n}^{y}) \) with \( \alpha_{n} \) given by (25). Similarly, let \( Y_n := \lambda D(\Gamma_{\lambda n}^{x} \| \frac{1}{1+\lambda}(\lambda \Pi_{\lambda n}^{x} + \Pi_{n}^{y})) + D(\Pi_{n}^{y} \| \frac{1}{1+\lambda}(\lambda \Pi_{\lambda n}^{x} + \Pi_{n}^{y})) \) and \( X_n := \chi^{2}(\Gamma_{\lambda n}^{x}, \Gamma_{n}^{y}) \). Then if \( m \) grows linearly in \( n \) as \( m = \lambda n \), we have

\[
\begin{align*}
\frac{8n\lambda}{1+\lambda} Z_n & \xrightarrow{d.} n \to \infty \lambda_{|Z|-1}^2 \quad (34) \\
2nY_n & \xrightarrow{d.} n \to \infty \lambda_{|Z|-1}^2 \quad (35) \\
\frac{n\lambda}{1+\lambda} X_n & \xrightarrow{d.} n \to \infty \lambda_{|Z|-1}^2 \quad (36)
\end{align*}
\]

**Proof:** We first note that the quantity \( D(\Gamma_{\lambda n}^{x} \| \alpha \Gamma_{\lambda n}^{x} + (1 - \alpha) \Gamma_{n}^{y}) \) is a decreasing function of \( \alpha \) and that \( D(\Gamma_{n}^{y} \| \alpha \Pi_{\lambda n}^{x} + (1 - \alpha) \Pi_{\lambda n}^{x}) \) is an increasing function of \( \alpha \). Thus, if we define \( Z_{1n}^{n} := D(\Gamma_{\lambda n}^{x} \| \frac{1}{1+\lambda}(\lambda \Pi_{\lambda n}^{x} + \Pi_{n}^{y})) \) and \( Z_{2n}^{n} := D(\Gamma_{n}^{y} \| \frac{1}{1+\lambda}(\lambda \Pi_{\lambda n}^{x} + \Pi_{n}^{y})) \), it follows by the definition of \( \alpha_{n} \) in (25) that

\[
\min\{Z_{1n}^{n}, Z_{2n}^{n}\} \leq Z_n \leq \max\{Z_{1n}^{n}, Z_{2n}^{n}\}. \quad (37)
\]

Now if \( W_n := \min\{Z_{1n}^{n}, Z_{2n}^{n}\} \), then we have \( |W_n - Z_{1n}^{n}| \leq |Z_{1n}^{n} - Z_{2n}^{n}| \). Hence by (31) we have \( n(W_n - Z_{1n}^{n}) \xrightarrow{d.} n \to \infty 0 \). Combining with (31) we get \( \frac{8n\lambda}{1+\lambda} W_n \xrightarrow{d.} n \to \infty \lambda_{|Z|-1}^2 \). By a similar argument it also follows that \( V_n := \max\{Z_{1n}^{n}, Z_{2n}^{n}\} \) satisfies \( \frac{8n\lambda}{1+\lambda} V_n \xrightarrow{d.} n \to \infty \lambda_{|Z|-1}^2 \). Thus by (37) 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.

The weak convergence of $Y_n$ is exactly the result of (32) in Lemma II.1. For the result on $X_n$, we apply Lemma II.1 to the function $h(\pi, \nu) = \chi^2(\pi, \nu)$. It is easily verified that the gradient and Hessian satisfy the necessary regularity conditions. Computing the Hessian at $(\mu, \mu)$ we obtain

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

Following the same steps as in the proof of (31) in Lemma II.1, we obtain (36).

The results of Theorem II.6 can be used to choose the threshold to meet approximate guarantees on the false alarm probability for the three tests of (24), (26) and (28) by following the same steps as in (8). The accuracy of these approximations are illustrated by simulations in Section III.

**D. Maximum likelihood estimates for multiple hypotheses**

Asymptotic optimality results are also known for some multiple hypothesis testing problems. Most optimality results in such problems are for generalized likelihood ratio tests based on the maximum likelihood estimate of the hypothesis. Even for such problems, it is possible to obtain weak convergence results for improving accuracy in estimating error probabilities in hypothesis tests and significance levels for the ML estimates. We consider three specific examples.

1) **Classification with Empirically Observed Statistics:** In this problem studied in [6], the objective is to classify a test string as having been generated by one of $M$ sources, using the knowledge of training strings drawn from each of the $M$ sources. The observations from each source $i$ are assumed to be drawn i.i.d. from a finite set $Z$ according to some fixed but unknown distributions. Let $X^i = (X^i_1, X^i_2, \ldots X^i_m)$ denote the sequence of training observations drawn from source $i$, and let $Z = (Z_1, \ldots, Z_n)$ denote the test string. This is a multiple hypothesis testing problem where hypothesis $H_i$ corresponds to the situation in which the test string was drawn from the $i^{th}$ source. The test proposed in [6] is based on the empirical distributions of the training and test strings. Let $\Gamma_m^i$ denote the empirical distribution of the $i^{th}$ training string $X^i$ and let $\Gamma_n^i$ denote the empirical distribution of the test string $Z$. We assume that $m$ grows linearly in $n$ as $m = \lambda n$ for some $\lambda > 0$. In this regime we use $\Delta_n^i$ to denote the empirical distribution of the string of length $(\lambda + 1)n$ obtained by appending the $i^{th}$ training string $X^i$ to the test string $Z$. The test accepts hypothesis $H_i$ if and only the following event occurs:

$$E_i \equiv \{D(\Gamma_n^i \| \Delta_n^i) + \lambda D(\Gamma_n^i \| \Delta_n^j) < \frac{\tau_n}{n} \} \cap \{D(\Gamma_n^i \| \Delta_n^i) + \lambda D(\Gamma_n^i \| \Delta_n^j) \geq \frac{\tau_n}{n} \}.$$  (38)
If all events $E_i : i = 1,2,\ldots,M$ do not occur, then the test rejects all $M$ hypotheses. We say that this test leads to an error under hypothesis $H_i$ if it outputs some $H_j$ with $j \neq i$ when hypothesis $H_i$ is indeed true. Note that rejection is not considered an error event under any of the $M$ hypotheses.

This test is known to be asymptotically optimal (see [6, Thm 2]) in the sense that it minimizes the rejection event subject to the constraint that the error exponent under every hypothesis is no less than $\eta$. However, in order to use this test in practice we need to have an idea of the exact error probability resulting from using this test in an experiment with a finite number of observations. We now compute an approximate bound on the error probability. From the definition of the error event it follows that the error probability $p_i^e$ under hypothesis $H_i$ satisfies

$$p_i^e = P_{H_i} \left( \bigcup_{j \neq i} E_j \right) \leq P_{H_i} \{ D(\Gamma_n \| \Delta_i^n) + \lambda D(\Gamma_{\lambda n}^i \| \Delta_i^n) < \frac{\tau}{n} \}. \quad (39)$$

We now obtain an approximate formula for $p_i^e$ from the weak convergence behavior of the test statistic $D(\Gamma_n \| \Delta_i^n) + \lambda D(\Gamma_{\lambda n}^i \| \Delta_i^n)$ under hypothesis $H_i$. This is in fact immediate from Theorem II.6. It is easy to see that $\Delta_i^n = \frac{1}{\lambda + 1} (\lambda \Gamma_{\lambda n}^i + \Gamma_n)$. Thus by the result of (35) we have

$$2n(D(\Gamma_n \| \Delta_i^n) + \lambda D(\Gamma_{\lambda n}^i \| \Delta_i^n)) \xrightarrow{d.}{n \to \infty} \chi^2_{|Z|-1}$$

provided the true distribution of the observations under hypothesis $H_i$ has full support over $Z$. Thus, we can approximately lower bound $p_i^e$ as

$$p_i^e \leq p(n, \tau) \approx 1 - F_{|Z|-1}(2\tau)$$

where $p(n, \tau) = P_{H_i} \{ D(\Gamma_n \| \Delta_i^n) + \lambda D(\Gamma_{\lambda n}^i \| \Delta_i^n) < \frac{\tau}{n} \}$ is the right hand side of (39). Thus, we can use this approximate formula for setting $\tau$ such that we meet a desired upper bound on the error probability under the different hypotheses.

2) Universal outlier hypothesis testing: A related multiple hypothesis testing problem was studied in [7]. Suppose one is given $K \geq 3$ strings $X^1, X^2, \ldots, X^K$ of length $n$ each drawn i.i.d. from some distribution on $P(Z)$. The objective is to test the $K$ hypothesis $H_1, H_2, \ldots, H_K$ where $H_i$ corresponds to the setting in which all strings except the $i$-th string are drawn under identical distributions. In other words, under hypothesis $H_i$, the string $X^i$ is an outlier, while all other strings are drawn from identical distributions. The objective is to identify the index outlier string without any prior knowledge about the distributions or the hypothesis. Let $\Gamma_n^i$ denote the empirical distribution of $X_i$. The test proposed in [7]...
is to output the hypothesis \( H_i^* \) given by

\[
i^* = \arg \min_i T_n^i
\]

where \( T_n^i := \sum_{j \neq i} D \left( \Gamma_j^n \| \frac{1}{K-1} \sum_{k \neq i} \Gamma_k^n \right). \tag{40}\]

This test is shown to have asymptotic optimality properties in terms of error exponents. However, in practice, if one were to use this test, it is important to obtain an idea of the accuracy of the test. In other words, we would like to know whether the sample size is large enough to be confident of the result of the test. One way to do this is to compare the test statistics \( T_n^i \) to a threshold and accept a hypothesis \( H_i \) if and only if \( T_n^i \) is the sole test statistic that exceeds the threshold. Equivalently, the test accepts hypothesis \( H_i \) if and only if the following event occurs

\[
E_i = \{ T_n^i < \frac{T}{n} \} \cap \bigcap_{j \neq i} \{ T_n^j \geq \frac{T}{n} \}. \tag{41}\]

The threshold \( \frac{T}{n} \) can be wisely chosen if the weak convergence behavior of the test statistic is known. Below we characterize the weak convergence of the statistic \( T_n^K \) under hypothesis \( H_K \). Similar results also hold under the other hypotheses. By following the same steps as in Section II-D1 this result can be used to approximate the error probability under each hypothesis \( H_i \).

**Proposition II.7.** Assume that the data strings \( X^1, X^2, \ldots, X^{K-1} \) are of equal length \( n \) and are drawn i.i.d. according to some fixed distribution \( \mu \in \mathcal{P}(Z) \) such that \( \mu \) has full support on \( Z \). Then we have

\[
\frac{2n(K-1)}{K-2} T_n^K \xrightarrow{d.} \chi^2_{(K-1)(|Z|-1)}. \tag{42}\]

This proposition is proved in the appendix using a variation of Lemma II.1.

3) **Test for identifying variable dependency:** The Chow-Liu algorithm \[15\] is a procedure to study statistical dependencies between variables in a multivariate distribution. In this approach, the dependencies among the variables are assumed to be modeled by a probabilistic graphical model \[16\] in the form of a tree. The variables are represented by vertices in the tree graph and the dependencies are represented by the edges in the tree. Suppose the variables are stacked up to form a random vector \( Z \). Let \( Z_{(j)} \) denote the \( j \)th element of the random vector \( Z \), and \( Z_S \) denote the set of elements of \( Z \) whose indices are in \( S \). If \( \mu \) is the distribution of \( Z \), we say that \( \mu \) is Markov on a graph \( G = (V, E) \) if the following holds for all \( j \in V \),

\[
P_\mu(Z_{(j)}|Z_{V\setminus\{j\}}) = P_\mu(Z_{(j)}|Z_{N_{(j)}})
\]
where $V$ is the set of vertices of $G$, $E$ is the set of edges of $G$, and $\mathcal{N}(j)$ is the set of vertices adjacent to $j$ in $G$. Let $\mathcal{P}(G)$ denote the set of all distributions that are Markov over $G$. Suppose one is given a sequence of observations $Z = Z_1, Z_2, \ldots$ where $Z_i \in Z^{|V|}$. Let $Z_i,\{j\}$ denote the $j$th element of the vector $Z_i$. The Chow-Liu algorithm provides a procedure to identify the graph structure representing the dependencies among the variables in $Z$ assuming that the graph is a tree.

Estimating the tree structure is a multiple hypothesis testing problem where each hypothesis is a tree structure $T_i$ in the set of all possible trees $\mathcal{T}$ with vertex set $V$. The Chow-Liu algorithm outputs the hypothesis $T_i$ given by:

$$\hat{T} = \arg\min_{T_i \in \mathcal{T}} \min_{Q \in \mathcal{P}(T_i)} D(\Gamma_n || Q).$$

The error exponent for this test has been examined in [8]. In practice, it is usually desirable, as suggested in [6], to have a no-match decision. In this case, the test accepts hypothesis $T_i$ if and only if the following occurs,

$$E_i = \{\hat{D}(T_i) < \frac{\tau}{n}\} \cap \bigcap_{j \neq i, T_j \in \mathcal{T}} \{\hat{D}(T_j) \geq \frac{\tau}{n}\}.$$  \hspace{1cm} (43)

where $\hat{D}(T) = \min_{Q \in \mathcal{P}(T)} D(\Gamma_n || Q)$ A no-match decision is made if $\hat{D} := \min_{T \in \mathcal{T}} \hat{D}(T)$ satisfies:

$$\hat{D} \geq \frac{\tau'}{n}.$$ \hspace{1cm} (44)

The weak convergence of the test statistic is derived in the following proposition, and can be used in choosing the threshold based on the requirement on the probability of error by following the same steps as in Section II-D1.

**Proposition II.8.** Assume the data $(Z_1, Z_2, \ldots, Z_n)$ are drawn i.i.d. according to a distribution $\mu$ that has full support over $Z^{|V|}$ and is Markov over the tree $T$. Then we have

$$2n\hat{D}_T \xrightarrow{n \to \infty} \chi^2_d$$ \hspace{1cm} (45)

where

$$d = |Z|^{|V|} - |Z|(|Z| - 1)(|V| - 1) - |Z|. $$ \hspace{1cm} (46)

When $\mu$ does not have full support, the chi-square approximation holds with $d \leq |Z|^{|V|} - |Z|(|Z| - 1)(|V| - 1) - |Z|$. This proposition is proved in the appendix using a variation of Lemma II.1.
FIG. 1. False alarm probabilities of the various homogeneity tests of Section II-C shown along with the $\chi^2$ approximation of these error probabilities obtained using the weak convergence result. A uniform distribution over an alphabet of size 8 was chosen for the distribution $\mu$.

III. SIMULATIONS

The weak convergence results obtained in this paper are useful for identifying thresholds and approximately estimating error probabilities for the various tests studied. For this purpose it is of interest to study the accuracy of the approximations obtained. The approximation accuracy of the simple goodness of fit test of Section II-A are discussed in [13]. For the other tests studied in this paper we now compare the accuracies of the error probability approximations obtained via the weak convergence results with the error probabilities computed using simulations.

A. Homogeneity testing

In order to estimate the accuracy of the approximation obtained from the weak convergence, we simulated the three tests considered in Section II-C using a uniform distribution over an alphabet of size 8 for $\mu$. 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.
B. Outlier hypothesis testing

We simulated the distribution of $T^K_n$ from (40) to estimate $\Pr\{T^K_n > \eta\}$ for fixed $\eta$ with $K = 4$ and $n$ ranging from 10 to 190. This gives an estimate of the error probability of the test proposed in (41) under hypothesis $H_K$. The results of Figure 2 indicate the accuracy of the approximation predicted by (42) assuming that $\mu$ is a uniform distribution on an alphabet of size 8. Clearly, the approximation is quite accurate in this regime.

C. Tests for identifying variable dependency.

We simulated the test given in (43), with a random generated distribution over a star-shaped tree with 4 vertices and $|Z| = 3$. We estimate both the error probability of not making a correct decision the test as well as the probability of making a no-match decision. The results in Figure 3 indicate the accuracy of the approximation predicted by (45).

IV. Conclusion and Future Work

In this paper we have identified weak convergence behavior of various test statistics that give rise to tests that are optimal in an error-exponent sense for several hypothesis testing problems. Although large deviations analysis for optimality in an error-exponent sense provides a convenient criterion for
identifying optimal hypothesis tests, they do not give an immediate solution to the problem of selecting thresholds in a practical experiment. In practice, the error probabilities in a finite sample size setting is more important than the error exponents. The weak convergence technique presented in this paper can be easily used to obtain the weak convergence behavior of the test statistic in many recently proposed tests. Moreover, our simulations results showed that such results on the weak convergence behavior of the test statistics can lead to good approximations of error probabilities for moderate sample lengths.

The technique used here is quite general and can be easily generalized to other tests with optimality properties guaranteed via large deviations analyses. This is evidenced by the fact that the results on the wide variety of tests presented here can all be derived based on Lemmas II.1 and II.2 or minor variations thereof. A different use of such results is in model-fitting problems. Weak convergence analysis of the test statistic can be used to quantify the effect of outliers in such problems as shown in [17]. One could also use other approaches for obtaining estimates of error probabilities of asymptotically optimal tests. Here we have analyzed only weak convergence behavior of the test statistics. However, it may be possible to obtain error probability estimates with even higher accuracy by studying Rao-Bahadur type refinements [18] to the large deviations optimality results. However, such results are generally more difficult to obtain.

Fig. 3. Probabilities that the test (43) does not accept the true hypothesis and the probability of claiming a no match (44), shown along with the estimate obtained via $\chi^2$ approximation.
compared to the weak convergence approximations.

ACKNOWLEDGEMENTS

This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no 247006.

APPENDIX

A. Proof of Lemma II.1

Proof: We prove only the result of (10) since result of (9) is the special case when \( h(x, y) = g(x) \).

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 \( \Gamma_{n}^{y} \) can be written as empirical averages of i.i.d. vectors. Hence, they satisfy the central limit theorem which says that,

\[
G_{n,x} = n^{\frac{1}{2}}(\Gamma_{\lambda n}^{x} - \mu) \xrightarrow{d} W_{\lambda}
\]

\[
G_{n,y} = n^{\frac{1}{2}}(\Gamma_{n}^{y} - \mu) \xrightarrow{d} W
\]

where the distributions of \( W \) and \( W_{\lambda} \) are as defined in the statement of the lemma. Moreover, since strings \( x \) and \( y \) are drawn independently, we have

\[
\begin{bmatrix}
G_{n,x} \\
G_{n,y}
\end{bmatrix} \xrightarrow{d} 
\begin{bmatrix}
W_{\lambda} \\
W
\end{bmatrix}
\]

with \( W \) and \( W_{\lambda} \) mutually independent. Considering a second-order Taylor’s expansion and using the condition on the directional derivative, we have, for \( n \) large enough,

\[
2n(h(\Gamma_{\lambda n}^{x}, \Gamma_{n}^{y}) - h(\mu, \mu)) = [G_{n,x}^{T}, G_{n,y}^{T}] \nabla^2 h(\tilde{\Gamma}_{n}^{x}, \tilde{\Gamma}_{n}^{y})
\]

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 \( \Gamma_{n}^{y} \) and hence \( \tilde{\Gamma}_{n}^{x} \) and \( \tilde{\Gamma}_{n}^{y} \) converge to \( \mu \) almost surely. By the continuity of the Hessian, we have

\[
\nabla^2 h(\tilde{\Gamma}_{n}^{x}, \tilde{\Gamma}_{n}^{y}) \xrightarrow{a.s.} \nabla^2 h(\mu, \mu).
\]

By applying the vector-version of Slutsky’s theorem [19], together with (49) and (50), we conclude that

\[
[G_{n,x}^{T}, G_{n,y}^{T}] \nabla^2 h(\tilde{\Gamma}_{n}^{x}, \tilde{\Gamma}_{n}^{y})
\]

\[
\begin{bmatrix}
G_{n,x} \\
G_{n,y}
\end{bmatrix} \xrightarrow{d} [W_{\lambda}^{T}, W^{T}] \nabla^2 h(\mu, \mu)
\]

\[
\begin{bmatrix}
W_{\lambda} \\
W
\end{bmatrix}
\]

thus establishing the lemma.

October 28, 2013 DRAFT
B. Proof of Lemma II.2

Since $D^T D$ is a projection matrix with rank $K$, it follows that $K$ eigenvalues of $D^T D$ are 1 and others 0. Let $\{u^1, \ldots, u^m\}$ denote an orthonormal basis, chosen so that the first $K$ vectors span the range space of $D^T D$. Hence $D^T D u^i = u^i$ for $1 \leq i \leq K$, and $D^T D u^i = 0$ for all other $i$.

Let $U$ denote the unitary matrix whose $m$ columns are $\{u^1, \ldots, u^m\}$. Then $\tilde{V} = U V$ is also an $N(0, I_m)$ random variable, and hence $D^T D V$ and $D^T \tilde{D} V$ have the same Gaussian distribution.

By construction the vector $\tilde{Y} = D^T D \tilde{V}$ has components given by

$$\tilde{Y}_i = \begin{cases} 
\tilde{V}_i & 1 \leq i \leq K \\
0 & K < i \leq m 
\end{cases}$$

It follows that $\|\tilde{Y}\|^2 = \|D^T D \tilde{V}\|^2 = \tilde{V}_1^2 + \cdots + \tilde{V}_K^2$ has a chi-squared distribution with $K$ degrees of freedom. Thus we conclude that $\|D^T D V\|^2$ also has a chi-squared distribution with $K$ degrees of freedom. The conclusion of the lemma follows from the observation that

$$\|D^T D V\|^2 = V^T D^T DD^T D V = V^T D^T D V = \|DV\|^2.$$

C. Proof of Theorem II.3

We know from the results of [4] that the robust divergence with respect to $\mathbb{P}$ can be expressed as the solution to the following optimization problem:

$$D^{\text{ROB}}(\mu \| \mathbb{P}) = \sup_{\mathcal{R}} \mu(\log(1 + r^T \psi)).$$

where the supremum is taken over

$$\mathcal{R} := \{r \in \mathbb{R}^d : 1 + r^T \psi(z) \geq 0 \text{ for all } z \in \mathbb{Z}\}$$

and $r^T \psi$ denotes the function $\sum_{i=1}^d r_i \psi_i$.

To prove Theorem II.3 We need the following lemma.

**Lemma A.1.** Suppose the functions $\{\psi_i : 0 \leq i \leq d\}$ are linearly independent over the support of $\pi$. If $Z_\pi$ denotes the support of $\pi \in \mathbb{P}$, then there exists an open neighborhood $B \subset \mathcal{P}(Z_\pi)$ such that for all $\mu \in B$, the supremum in (51) is achieved at a unique point $r(\mu)$.

**Proof:** We verify that the proposition in the lemma holds for $B = \{\mu \in \mathcal{P}(\mathbb{Z}) : |\mu(z) - \pi(z)| < \epsilon \text{ for all } z \in Z_\pi, \mu(y) = 0 \text{ for all } y \in \mathbb{Z} \setminus Z_\pi\}$ where $\epsilon = \frac{1}{2} \min_{z \in Z_\pi} \pi(z)$. Clearly, for all $\mu \in B$, the support of $\mu$ is equal to $Z_\pi$ and hence $0 \leq D^{\text{ROB}}(\mu \| \mathbb{P}) \leq D(\mu \| \pi) < \infty$. 

October 28, 2013
Now since the functions $\psi_i$ are linearly independent over $\mathbb{Z}$, and the value of the the optimization problem in (51) is finite, it follows that we can restrict the constraint set in (51) to a bounded subset of the closed set $\mathcal{R}$. Thus we can restrict the optimization in (51) to a compact set. Furthermore, the objective function is strictly concave by the linear independence assumption and hence the conclusion of the lemma follows.

It follows from $0 \leq D^{\text{ROB}}(\mu\|P) \leq D(\mu\|\pi)$ that $D^{\text{ROB}}(\pi\|P) = 0$. This is achieved by $r(\pi)$ satisfying $1 + r(\pi)^T\psi(z) = 1$ for all $z \in \mathbb{Z}$. Since the objective function in (51) is continuously differentiable in a neighborhood $B$ of $r(\pi)$, it follows from the uniqueness of $r(\pi)$ and Inverse Function Theorem [20, Theorem 9.2.1] that $r(\mu)$ is continuous in a neighborhood $B'$ of $\pi$. Therefore, the constraints $r \in \mathbb{R}$ are inactive in a neighborhood $B''$ of $\pi$. Define

$$g(\mu) = \sup_r \mu(\log(1 + r^T\psi)),$$

We have argued that $g(\mu) = D^{\text{ROB}}(\mu\|P)$ in an open neighborhood $B''$ of $\pi$. It is easy to check that the conditions of Lemma II.1 are satisfied for $g(\mu)$. The rest of the work to prove Theorem II.3 is to derive the Hessian of $g$ by applying the Implicit Function Theorem: It follows from first order necessary condition of optimality that

$$\mu \left( \frac{\psi_i}{1 + r^T\psi} \right) = 0 \text{ for all } i.$$

Consequently,

$$(\nabla r(\mu))_z = \left[ \mu \left( \frac{\psi\psi^T}{(1 + r^T\psi)^2} \right) \right]^{-1} \frac{\psi(z)}{1 + r^T\psi(z)}.$$

It follows that

$$[M_g]_{z,z'} = \frac{\psi(z)}{1 + r^T\psi(z)} \left[ \mu \left( \frac{\psi\psi^T}{(1 + r^T\psi)^2} \right) \right]^{-1} \frac{\psi(z')}{1 + r^T\psi(z')}.$$

Applying Lemma II.1 leads to

$$2nD^{\text{ROB}}(\Gamma_n\|P) \xrightarrow{n \to \infty} V^T \Sigma\Sigma^\dagger M_g \Sigma^\dagger V$$

where $\Sigma = \text{diag}(\mu) - \mu^T\mu$ and $V \sim \mathcal{N}(0, I_{|Z|})$. Let $D = \Sigma^\dagger M_g \Sigma^\dagger$. It can be verified that the rank of $D$ is $d$, $D^T = D$ and $DD = D$. Consequently,

$$2nD^{\text{ROB}}(\Gamma_n\|P) \xrightarrow{n \to \infty} \|DV\|^2.$$

The theorem then follows by Lemma II.2.

As an alternative approach to prove Theorem II.3, it was also shown in [13, Section II.E] that the robust divergence can be expressed as a mismatched divergence defined with respect to a log-linear function class. Theorem II.3 is thus a special case of the weak convergence result of mismatched divergence statistic from [13, Theorem III.3(i)].
D. Proof of Lemma II.4

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

Now (23) can be equivalently written as a convex problem:

\[
\begin{align*}
\min_{\tau, \nu} & \quad \tau \\
\text{s.t.} & \quad D(\mu^1\|\nu) \leq \tau, \quad D(\mu^2\|\nu) \leq \tau, \\
& \quad \sum_{x \in Z} \nu(x) = 1, \quad \nu(x) \geq 0, \text{ for all } x \in Z.
\end{align*}
\]

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 \( \ell_1, \ell_2, \) and \( \kappa \) such that

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

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

E. Proof of Proposition II.7

We need the following lemma. We omit the proof because it follows via the same steps as that used for proving (10).

**Lemma A.2.** Assume that the data strings \( X^1, X^2, \ldots, X^{K-1} \) are strings of length \( n \) drawn i.i.d. according to some fixed distribution \( \mu \in \mathcal{P}(Z) \) such that \( \mu \) has full support on \( Z \). Let \( \Gamma_i^n \) denote the empirical distribution of the observations in \( X^i \). Let \( h : (\mathcal{P}(Z))^{K-1} \mapsto \mathbb{R} \) be a continuous real-valued function whose gradient and Hessian are continuous in the neighborhood of \( \mu^{K-1} := (\mu, \mu, \ldots, \mu) \in (\mathcal{P}(Z))^{K-1} \). If the directional derivative satisfies \( \nabla h(\mu^{K-1})^T(\psi - \mu^{K-1}) = 0 \) for all \( \psi \in (\mathcal{P}(Z))^{K-1} \),
then

\[ 2n(h(\Gamma_1^n, \Gamma_2^n, \ldots, \Gamma_{K-1}^n) - h(\mu^{K-1})) \rightarrow_{n \to \infty} [W^T_1, W^T_2, \ldots, W^T_{K-1}] M_h \]

(52)

where \( M_h = \nabla^2 h(\mu^{K-1}) \) and \( W_i \) are i.i.d. random vectors distributed as \( W_i \sim \mathcal{N}(0, \Sigma) \) with \( \Sigma = \text{diag}(\mu) - \mu \mu^T \).

We apply the conclusion of this lemma to the function

\[ h(\nu^1, \nu^2, \ldots, \nu^{K-1}) = \sum_{j=1}^{K-1} D \left( \nu^j \left| \frac{1}{K-1} \sum_{k \neq j} \nu^k \right. \right) \]

The Hessian is given by

\[
M_h = \begin{bmatrix}
\text{diag}(\frac{1}{\mu}) & 0 & 0 \\
0 & \ddots & 0 \\
0 & 0 & \text{diag}(\frac{1}{\mu})
\end{bmatrix} - \frac{1}{K-1} \begin{bmatrix}
\text{diag}(\frac{1}{\mu}) & \ldots & \text{diag}(\frac{1}{\mu}) \\
\vdots & \ddots & \vdots \\
\text{diag}(\frac{1}{\mu}) & \ldots & \text{diag}(\frac{1}{\mu})
\end{bmatrix}.
\]

Now right hand side of (52) can be expressed as \( \frac{K-2}{K-1} \|DV\|^2 \) where \( V \sim \mathcal{N}(0, I_{(K-1)|Z|}) \) and 

\[
D = \left( \frac{K-1}{K-2} \right)^\frac{1}{2} M_h \Sigma^\frac{1}{2}_w
\]

with

\[
\Sigma_w = \begin{bmatrix}
\text{diag}(\mu) - \mu \mu^T & 0 & 0 \\
0 & \ddots & 0 \\
0 & 0 & \text{diag}(\mu) - \mu \mu^T
\end{bmatrix}
\]

It can be verified that \( D \) satisfies the conditions of Lemma II.2 and that rank of \( D \) equals \( (K-1)(|Z| - 1) \).

The result then follows by Lemma II.2.

\( F. \) Proof of Proposition II.8

For \( T = (V, E) \),

\[
\hat{D}_T = \sum_{z \in Z^{|V|}} \mu(z) \log \mu(z)
\]

\[ - \sum_{e \in E} \sum_{z \in Z^{|V|}} \mu(z) \log \left( \sum_{z'' \in Z^{|V|} : z''|e| = z_e} \mu(z'') \right) \]

\[ + \sum_{v \in S} \sum_{z \in Z^{|V|}} \mu(z) \log \left( \sum_{z'' \in Z^{|V|} : z''|v| = z_v} \mu(z'') \right) \]
where $S$ is the multiset in which number of times an element $v \in V$ appears is equal to the total number of times $v$ appears as an node of edges in $E$ minus one. $z_e$ and $z_v$ are the components of $z$ corresponding to the edge $e$ and node $v$, respectively.

Let $M$ denote the hessian of $\hat{D}_T$, which is of dimension $|Z|^{|V|} \times |Z|^{|V|}$ and $M(z, z')$. We fix the ordering in the element of $Z^{|V|}$ so that the set of element in $Z^{|V|}$ with the chosen ordering is isomorphic to $\{1, \ldots, |Z|^{|V|}\}$ with the natural ordering. Let $M(z, z')$ denote the element on row $z$ and column $z'$ of the Hessian. Then

$$M(z, z') = \mathbb{I}\{z = z'\} \frac{1}{\mu(z)} - \sum_{e \in E} \mathbb{I}\{z_e = z'_e\} \frac{1}{\sum_{z'' \in Z^{|V|}: z''_e = z_e} \mu(z'')} + \sum_{v \in V} \mathbb{I}\{z_v = z'_v\} \frac{1}{\sum_{z'' \in Z^{|V|}: z''_v = z_v} \mu(z'')}$$

Let $\Sigma = \text{diag}(\mu) - \mu \mu^T$, i.e., $\Sigma$ is the $|Z|^{|V|}$ by $|Z|^{|V|}$ dimension matrix whose element at $(z, z')$ is given by

$$\Sigma(z, z') = \mathbb{I}\{z = z'\} \mu(z) - \mu(z) \mu(z').$$

Let $D = \Sigma^{1/2} M \Sigma^{1/2}$. It can be verified that $DD = D$, $D^T = D$, and the rank of $D$ is equal to $d$ given in (46). The result then follows by Lemma II.2.

**REFERENCES**


