Universal and Composite Hypothesis Testing via Mismatched Divergence

Jayakrishnan Unnikrishnan*, Dayu Huang,
Sean Meyn, Amit Surana and Venugopal Veeravalli

Abstract

For the *universal* hypothesis testing problem, where the goal is to decide between the known null hypothesis distribution and some other unknown distribution, Hoeffding proposed a universal test in the nineteen sixties. Hoeffding's universal test statistic can be written in terms of Kullback-Leibler (K-L) divergence between the empirical distribution of the observations and the null hypothesis distribution. In this paper a modification of Hoeffding's test is considered based on a relaxation of the K-L divergence test statistic, referred to as the mismatched divergence. The resulting mismatched test is shown to be a generalized likelihood-ratio test for the case where the alternate distribution lies in a parametric family of the distributions characterized by a finite dimensional parameter, i.e., it is a solution to the corresponding *composite* hypothesis testing problem. For certain choices of the alternate distribution, it is shown that both the Hoeffding test and the mismatched test have the same asymptotic performance in terms of error exponents. More importantly, it is established that the mismatched test has a significant advantage over the Hoeffding test in terms of finite sample size performance. This advantage is due to the difference in the asymptotic variances of the two test statistics under the null hypothesis. In particular, the variance of the K-L divergence grows linearly with the alphabet size, making the test impractical for applications involving large alphabet distributions. The variance of the mismatched divergence on the other hand grows linearly with the dimension of the parameter space, and can hence be controlled through a prudent choice of the function class defining the mismatched divergence.

Keywords: Hypothesis testing, entropy, Generalized Likelihood-Ratio Test, Kullback–Leibler information, online detection

* Corresponding author.

Amit Surana is with United Technologies Research Center, 411 Silver Lane, E. Hartford, CT. Email: SuranaA@utrc.utc.com. The remaining authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL. Email: {junnikr2, dhuang8, meyn, vvv}@illinois.edu.

This research partially supported by NSF under grant CCF 07-29031 and by UTRC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or UTRC.

Portions of the results presented here were published in abridged form in [1].

I. INTRODUCTION AND BACKGROUND

This paper is concerned with the following hypothesis testing problem: Suppose that the observations $\boldsymbol{Z} = \{Z_t : t = 1, \dots\}$ form an i.i.d. sequence evolving on a set of cardinality N, denoted by $\mathsf{Z} = \{z_1, z_2, \dots, z_N\}$. Based on observations of this sequence we wish to decide if the marginal distribution is a given distribution π^0, or some other distribution that is either unknown or known only to belong to a certain class of distributions.

A decision rule is characterized by a *sequence* of tests $\boldsymbol{\phi} := \{\phi_n : n \geq 1\}$, where $\phi_n : \mathsf{Z}^n \mapsto \{0, 1\}$. The decision based on the first n elements of the observation sequence is given by $\phi_n(Z_1, Z_2, \dots, Z_n)$, where $\phi_n = 0$ represents a decision in favor of accepting π^0 as the true marginal distribution.

The set of probability measures on Z is denoted $\mathcal{P}(\mathsf{Z})$. The relative entropy (or Kullback-Leibler divergence) between two distributions $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$ is denoted $D(\nu^1 \| \nu^2)$, and for a given $\mu \in \mathcal{P}(\mathsf{Z})$ and $\eta > 0$ the *divergence ball* of *radius* η around μ is defined as,

$$\mathcal{Q}_\eta(\mu) := \{\nu \in \mathcal{P}(\mathsf{Z}) : D(\nu \| \mu) < \eta\}. \tag{1}$$

The empirical distribution or *type* of the finite set of observations $(Z_1, Z_2, \dots, Z_n)$ is a random variable Γ^n taking values in $\mathcal{P}(\mathsf{Z})$:

$$\Gamma^n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i = z\}, \qquad z \in \mathsf{Z} \tag{2}$$

where $\mathbb{I}$ denotes the indicator function.

In the general universal hypothesis testing problem, we do not have any prior information about the alternate distribution. For such a setting Hoeffding [2] proposed a generalized likelihood-ratio test (GLRT) where one assumes that the alternate distribution π^1 could be any arbitrary distribution in $\mathcal{P}(\mathsf{Z})$, the set of probability distributions on Z. The resulting test sequence is given by:

$$\phi_n^{\mathrm{H}}(Z_1, Z_2, \dots, Z_n) = \mathbb{I}\{\sup_{\pi^1 \in \mathcal{P}(\mathsf{Z})} \frac{1}{n} \sum_{i=1}^{n} \log \frac{\pi^1(Z_i)}{\pi^0(Z_i)} \geq \eta\} \tag{3}$$

It is easy to see that the test (3) can be rewritten as follows:

$$\begin{aligned} \phi_n^{\mathrm{H}}(Z_1, Z_2, \dots, Z_n) &= \mathbb{I}\{\frac{1}{n} \sum_{i=1}^{n} \log \frac{\Gamma^n(Z_i)}{\pi^0(Z_i)} \geq \eta\} \\ &= \mathbb{I}\{\sum_{z \in \mathsf{Z}} \Gamma^n(z) \log \frac{\Gamma^n(z)}{\pi^0(z)} \geq \eta\} \\ &= \mathbb{I}\{D(\Gamma^n \| \pi^0) \geq \eta\} \\ &= \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta(\pi^0)\} \end{aligned} \tag{4}$$

We refer to the above test as the Hoeffding test.

If we have some prior information on the alternate distribution, a different version of the GLRT is used. In particular, suppose it is known that the alternate distribution lies in a parametric family of distributions of the following form:

$$\mathcal{E}_{\pi^0} := \{\check{\pi}^r : r \in \mathbb{R}^d\}.$$

where $\check{\pi}^r \in \mathcal{P}(\mathsf{Z})$ are probability distributions on Z parameterized by a parameter $r \in \mathbb{R}^d$. The specific form of $\check{\pi}^r$ is defined later in the paper. In this case, the resulting composite hypothesis testing problem is typically solved using a GLRT (see [3] for results related to the present paper, and [4] for a more recent account) of the following form:

$$\phi_n^{\text{MM}}(Z_1, Z_2, \ldots, Z_n) = \mathbb{I}\{\sup_{\pi^1 \in \mathcal{E}_{\pi^0}} \langle \Gamma^n, \log \frac{\pi^1}{\pi^0} \rangle \geq \eta\}. \tag{5}$$

We show that this test can be interpreted as a relaxation the Hoeffding test of (4). In particular we show that

$$\phi_n^{\text{MM}}(Z_1, Z_2, \ldots, Z_n) = \mathbb{I}\{D^{\text{MM}}(\Gamma^n \| \pi^0) \geq \eta\}. \tag{6}$$

where D^{MM} is a relaxation the K-L divergence. We refer to this quantity as the *mismatched divergence* and the test (6) as the *mismatched test*. The mismatched-divergence is a lower bound based on a relaxation of a variational representation of the K-L divergence. We illustrate various properties of the mismatched divergence later in the paper. It is important to note that although the mismatched test is the GLRT solution to the composite hypothesis testing problem where the alternate distribution belongs to $\mathcal{E}_{\pi^0}$, it can also be applied to the universal hypothesis testing problem where no prior information is available about the alternate distribution. Thus the mismatched test can be considered to be a surrogate to the Hoeffding test for the universal hypothesis testing problem.

The terminology is borrowed from the *mismatched-channel* introduced by Lapidoth in [5]. The mismatched-divergence described here is a generalization of the relaxation introduced in [6], and in the setting of [6] it can be viewed as a particular exponential family model assumption. In this way we embed the analysis of the resulting universal test within the framework of Csiszár and Shields [7]. The mismatched test statistic can also be viewed as a generalization of the robust hypothesis testing statistic introduced in [8], [9].

When the alternate distribution satisfies $\pi^1 \in \mathcal{E}_{\pi^0}$, we show that, under some regularity conditions on $\mathcal{E}_{\pi^0}$, the mismatched test of (6) and Hoeffding's test of (4) have identical asymptotic performance in terms of error exponents. More importantly, we establish that the proposed mismatched test has a significant advantage over the Hoeffding test in terms of finite sample size performance. This advantage is due to the difference in the asymptotic variances of the two test statistics under the null hypothesis. In particular, we show that the variance of the K-L divergence grows linearly with the alphabet size, making the test impractical for applications involving large alphabet distributions. We also show that the variance of the mismatched divergence grows linearly with the dimension d of the parameter space, and can hence be controlled through a prudent choice of the function class defining the mismatched divergence.

The remainder of the paper is organized as follows. We begin in Section II with a description of mismatched divergence and the mismatched test, and describe their relation to other concepts including robust hypothesis testing, composite hypothesis testing, reverse I-projection, and maximum likelihood (ML) estimation. Formulae for the asymptotic mean and variance of the test statistics are presented in Section III. Section III also contains a discussion interpreting these asymptotic results in terms of the performance of the detection rule. Proofs of the main results are provided in the appendix. Conclusions and directions for future research are contained in Section IV.

II. Mismatched Divergence

We adopt the following compact notation in the paper: For any function $f\colon \mathsf{Z} \to \mathbb{R}$ and $\pi \in \mathcal{P}(\mathsf{Z})$ we denote the mean $\sum_{i=1}^N f(z_i)\pi_i$ by $\pi(f)$, or by $\langle \pi, f\rangle$ when we wish to emphasize the convex-analytic setting. At times we will extend these definitions to allow functions f taking values in a vector space.

The logarithmic moment generating function (log-MGF) is denoted

$$\Lambda_\pi(f) = \log(\pi(e^f)).$$

For any two probability measures $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$ the relative entropy is expressed,

$$D(\nu^1 \| \nu^2) = \begin{cases} \langle \nu^1, \log(\nu^1/\nu^2)\rangle & \text{if } \nu^1 \prec \nu^2 \\ \infty & \text{else} \end{cases}$$

where $\nu^1 \prec \nu^2$ denotes absolute continuity. The following proposition recalls a well-known variational representation. This can be obtained, for instance, by specializing the representation in [10] to an i.i.d. setting.

Proposition II.1. *The relative entropy can be expressed as the convex dual of the log moment generating function: For any two probability measures* $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$,

$$D(\nu^1 \| \nu^2) = \sup_f \big(\nu^1(f) - \Lambda_{\nu^2}(f)\big) \tag{7}$$

where the supremum is taken over the space of all real-valued functions on Z*. Furthermore, if* ν^1 *and* ν^2 *have equal supports, then the supremum is achieved by the log likelihood ratio function* $f^* = \log(\nu^1/\nu^2)$.

The representation (7) is the basis of the mismatched divergence. We fix a set of functions denoted $\mathcal{F}$, and obtain a lower bound on the relative entropy by taking the supremum over the smaller set as follows,

$$D^{\text{MM}}(\nu^1 \| \nu^2) := \sup_{f \in \mathcal{F}} \{\nu^1(f) - \Lambda_{\nu^2}(f)\} \tag{8}$$

If ν^1 and ν^2 have full support, and if the function class $\mathcal{F}$ contains the log-likelihood ratio $f^* = \log(\nu^1/\nu^2)$, then it is immediate from Proposition II.1 that the supremum in (8) is achieved by f^*, and in this case $D^{\text{MM}}(\nu^1 \| \nu^2) = D(\nu^1 \| \nu^2)$. Moreover, since the objective function in (8) is invariant to shifts of f, it follows that even if a constant scalar is added to the function f^*, it still achieves the supremum in (8).

In this paper the function class is assumed to be defined through a finite-dimensional parametrization of the form,

$$\mathcal{F} = \{f_r : r \in \mathbb{R}^d\} \tag{9}$$

Further assumptions will be imposed in our main results. In particular, we will assume that $f_r(z)$ is differentiable as a function of r for each z.

A. Basic structure of mismatched divergence

The mismatched test is defined to be a relaxation of the Hoeffding test described in (4). We replace the divergence functional with the mismatched divergence $S(\Gamma^n) := D^{\text{MM}}(\Gamma^n \| \pi^0)$. Thus the mismatched test sequence is given by

$$\begin{aligned} \phi_n^{\text{MM}}(Z_1, Z_2, \dots, Z_n) &= \mathbb{I}\{S(\Gamma^n) \geq \eta\} \\ &= \mathbb{I}\{D^{\text{MM}}(\Gamma^n \| \pi^0) \geq \eta\} = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta^{\text{MM}}(\pi^0)\} \end{aligned} \tag{10}$$

where $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ is the mismatched divergence ball of radius η around π^0 defined analogously to (1):

$$\mathcal{Q}_\eta^{\text{MM}}(\mu) = \{\nu \in \mathcal{P}(\mathsf{Z}) : D^{\text{MM}}(\nu \| \mu) < \eta\}. \tag{11}$$

The next proposition establishes some basic geometry of the mismatched divergence balls. For any function g we define the following hyperplane and half-space:

$$\begin{aligned} \mathcal{H}_g &:= \{\nu : \nu(g) = 0\} \\ \mathcal{H}_g^- &:= \{\nu : \nu(g) < 0\}. \end{aligned} \tag{12}$$

Proposition II.2. *The following hold for any $\nu, \pi \in \mathcal{P}(\mathsf{Z})$, and any collection of functions $\mathcal{F}$:*

(i) *For each $\eta > 0$ we have $\mathcal{Q}_\eta^{\rm MM}(\pi) \subset \bigcap \mathcal{H}_g^-$, where the intersection is over all normalized functions,*

$$g = f - \Lambda_\pi(f) - \eta \tag{13}$$

with $f \in \mathcal{F}$.

(ii) *Suppose that $\eta = D^{\rm MM}(\nu \| \pi)$ is finite and non-zero. Suppose the supremum in* (8) *is achieved by $f^* \in \mathcal{F}$. Then $\mathcal{H}_{g^*}$ is a supporting hyperplane to $\mathcal{Q}_\eta^{\rm MM}(\pi)$, where g^* is given in* (13) *with $f = f^*$.*

Proof: (i) Suppose $\mu \in \mathcal{Q}_\eta^{\rm MM}(\pi)$. Then, for any $f \in \mathcal{F}$,

$$\nu(f) - \Lambda_\pi(f) - \eta \le D^{\rm MM}(\mu \| \pi) - \eta < 0$$

That is, for any $f \in \mathcal{F}$, on defining g by (13) we obtain the desired inclusion $\mathcal{Q}_\eta^{\rm MM}(\pi) \subset \mathcal{H}_g^-$.

(ii) Let $\gamma \in \mathcal{H}_{g^*}$ be arbitrary. Then we have:

$$\begin{aligned} D^{\rm MM}(\gamma \| \pi) &= \sup_r \big(\gamma(f_r) - \Lambda_\pi(f_r)\big) \\ &\ge \gamma(f^*) - \Lambda_\pi(f^*) = \Lambda_\pi(f^*) + \eta - \Lambda_\pi(f^*) = \eta. \end{aligned}$$

Hence it follows that $\mathcal{H}_{g^*}$ supports $\mathcal{Q}_\eta^{\rm MM}(\pi)$ at ν.

■

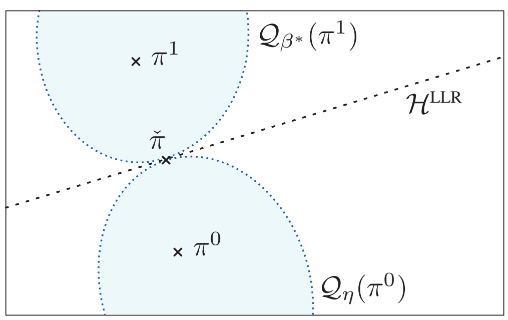

Fig. 1. *Geometric interpretation of the log likelihood ratio test.* The exponent $\beta^* = \beta^*(\eta)$ is the largest constant satisfying $\mathcal{Q}_\eta(\pi^0) \cap \mathcal{Q}_{\beta^*}(\pi^1) = \emptyset$. The hyperplane $\mathcal{H}^{\rm LLR} := \{\nu : \nu(L) = \check\pi(L)\}$ separates the convex sets $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$.

B. Asymptotic optimality of the mismatched test

The asymptotic performance of a sequential binary hypothesis testing problem is typically characterized in terms of error exponents. We adopt the following criterion for performance evaluation, following Hoeffding [2] (and others, notably [11], [12].) Suppose that the observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ form an i.i.d. sequence evolving on Z. For a given π^0, and a given alternate distribution π^1, the type I and type II error exponents are denoted respectively by,

$$\begin{aligned} J^0_\phi &:= \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^0}\{\phi_n(Z_1,\dots,Z_n)=1\}), \\ J^1_\phi &:= \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^1}\{\phi_n(Z_1,\dots,Z_n)=0\}) \end{aligned} \tag{14}$$

where in the first limit the marginal distribution of Z_t is π^0, and in the second it is π^1. The limit J^0_ϕ is also called the false-alarm error exponent, and J^1_ϕ the missed-detection error exponent.

For a given constraint $\eta > 0$ on the false-alarm exponent J^0_ϕ, an optimal test is the solution to the asymptotic Neyman-Pearson hypothesis testing problem,

$$\beta^*(\eta) = \sup\{J^1_\phi : \textit{ subject to } J^0_\phi \ge \eta\} \tag{15}$$

where the supremum is over all allowed test sequences $\boldsymbol{\phi}$. While the exponent $\beta^*(\eta) = \beta^*(\eta, \pi^1)$ depends upon π^1, Hoeffding's test we described in (4) does not require knowledge of π^1, yet achieves the optimal exponent $\beta^*(\eta, \pi^1)$ for any π^1. The optimality of Hoeffding's test established in [2] easily follows from Sanov's theorem.

While the mismatched test described in (6) is not always optimal for (15) for a general choice of π^1, it is optimal for some specific choices of the alternate distributions. The following corollary to Proposition II.2 captures this idea.

Corollary II.1. *Suppose $\pi^0, \pi^1 \in \mathcal{P}(\mathsf{Z})$ have equal supports. If the function class $\mathcal{F}$ contains all functions in the set $\{rL : r > 0\}$, where L is the log likelihood-ratio $L := \log(\pi^1/\pi^0)$, then the mismatched test is optimal in the sense that the constraint $J^0_{\phi^{\text{MM}}} \ge \eta$ is satisfied with equality, and under π^1 the optimal error exponent $J^1_{\phi^{\text{MM}}} = \beta^*(\eta)$ is achieved for all $\eta \in (0, D(\pi^1\|\pi^0))$.*

Proof: Suppose $\mathcal{F}$ contains all functions in the set $\{rL : r > 0\}$. Consider the twisted distribution $\check{\pi} = \kappa(\pi^0)^{1-\varrho}(\pi^1)^{\varrho}$, where κ is a normalizing constant and $\varrho \in (0,1)$ is chosen so as to guarantee $D(\check{\pi}\|\pi^0) = \eta$. It is known that the hyperplane $\mathcal{H}^{\text{LLR}} := \{\nu : \nu(L) = \check{\pi}(L)\}$ separates the divergence balls $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ at $\check{\pi}$. This geometry, which is implicit in [11], is illustrated in Figure 1.

From the form of $\check{\pi}$ it is also clear that

$$\log\frac{\check{\pi}}{\pi^0} = \varrho L - \Lambda_{\pi^0}(\varrho L).$$

Hence it follows that the supremum in the variational representation of $D(\check{\pi}\|\pi^0)$ is achieved by ϱL. Furthermore, since $\varrho L \in \mathcal{F}$

$$D^{\text{MM}}(\check{\pi}\|\pi^0) = D(\check{\pi}\|\pi^0) = \eta = \check{\pi}(\varrho L) - \Lambda_{\pi^0}(\varrho L).$$

This means that $\mathcal{H}^{\text{LLR}} = \{\nu : \nu(\varrho L - \Lambda_{\pi^0}(\varrho L) - \eta) = 0\}$. Hence, by applying the result of Proposition II.2 (ii) it follows that the hyperplane $\mathcal{H}^{\text{LLR}}$ separates $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$. This in particular means that the sets $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ are disjoint. This fact, together with Sanov's theorem proves the corollary. ■

The corollary indicates that while using the mismatched test in practice, the function class might be chosen to include approximations to scaled versions of the log-likelihood ratios of the anticipated alternate distributions $\{\pi^1\}$ with respect to π^0.

The mismatched divergence has several equivalent characterizations. We first relate it to an ML estimate based on a parametric family of distributions.

C. *Mismatched divergence and ML estimation*

We fix a distribution $\pi \in \mathcal{P}(\mathsf{Z})$ and a function class of the form (9). For each $r \in \mathbb{R}^d$ the *twisted distribution* $\check{\pi}^r \in \mathcal{P}(\mathsf{Z})$ is defined as,

$$\check{\pi}^r := \pi \exp(f_r - \Lambda_\pi(f_r)). \tag{16}$$

The collection of all such distributions is denoted

$$\mathcal{E}_\pi := \{\check{\pi}^r : r \in \mathbb{R}^d\}.$$

On interpreting $f_r - \Lambda_\pi(f_r)$ as a log-likelihood ratio we obtain in Proposition II.3 the representation of mismatched divergence,

$$D^{\text{MM}}(\mu \| \pi) = \sup_{r \in \mathbb{R}^d} \big(\mu(f_r) - \Lambda_\pi(f_r)\big) = D(\mu \| \pi) - \inf_{\nu \in \mathcal{E}_\pi} D(\mu \| \nu) \tag{17}$$

The infimum on the RHS of (17) is known as *reverse I-projection* [7]. Proposition II.4 that follows uses this representation to obtain other interpretations of the mismatched test.

Proposition II.3. *The identity* (17) *holds for any function class* $\mathcal{F}$. *The supremum is achieved by some* $r^* \in \mathbb{R}^d$ *if and only if the infimum is attained at* $\nu^* = \check{\pi}^{r^*} \in \mathcal{E}_\pi$. *If a minimizer* ν^* *exists, we obtain the* generalized Pythagorean identity,

$$D(\mu \| \pi) = D^{\text{MM}}(\mu \| \pi) + D(\mu \| \nu^*)$$

Proof: For any r we have $\mu(f_r) - \Lambda_\pi(f_r) = \mu(\log(\check{\pi}^r/\pi))$. Consequently,

$$\begin{aligned} D^{\text{MM}}(\mu\|\pi) &= \sup_r(\mu(f_r) - \Lambda_\pi(f_r)) \\ &= \sup_r \mu\left(\log\left(\frac{\mu}{\pi}\frac{\check{\pi}^r}{\mu}\right)\right) \\ &= \sup_r \{D(\mu\|\pi) - D(\mu\|\check{\pi}^r)\} \end{aligned}$$

This proves the identity (17), and the remaining conclusions follow directly. ■

The representation of Proposition II.3 invites the interpretation of the optimizer in the definition of the mismatched test statistic in terms of an ML estimate. Given the well-known correspondence between maximum-likelihood estimation and the generalized likelihood ratio test (GLRT), Proposition II.4 implies that the mismatched test is a special case of the GLRT analyzed in [3].

Proposition II.4. *Suppose that the observations $\boldsymbol{Z}$ are modeled as an i.i.d. sequence, with marginal in the exponential family $\mathcal{E}_\pi$. Let $\hat{r}^n$ denote the ML estimate of r based on the first n samples,*

$$\hat{r}^n \in \arg\max_{r\in\mathbb{R}^d} \mathsf{P}_{\check{\pi}^r}\{Z_1 = a_1, Z_2 = a_2, \ldots, Z_n = a_n\} = \arg\max_{r\in\mathbb{R}^d} \Pi_{i=1}^n \check{\pi}^r(a_i)$$

where a_i indicates the observed value of the i-th symbol. Assuming the maximum is attained we have the following interpretations:

(i) *The distribution $\check{\pi}^{\hat{r}^n}$ solves the reverse I-projection problem,*

$$\check{\pi}^{\hat{r}^n} \in \arg\min_{\nu\in\mathcal{E}_\pi} D(\Gamma^n\|\nu).$$

(ii) *The function $f^* = f_{\hat{r}^n}$ achieves the supremum that defines the mismatched divergence, $D^{\text{MM}}(\Gamma^n\|\pi) = \Gamma^n(f^*) - \Lambda_\pi(f^*)$.*

Proof: The ML estimate can be expressed $\hat{r}^n = \arg\max_{r\in\mathbb{R}^d}\langle\Gamma^n, \log\check{\pi}^r\rangle$, which obviously gives (i).

Applying (i) we obtain the identity,

$$\arg\min_{\nu\in\mathcal{E}_\pi} D(\Gamma^n\|\nu) = \arg\max_{\nu\in\mathcal{E}_\pi}\langle\Gamma^n, \log\nu\rangle, \qquad \nu\in\mathcal{P}.$$

This combined with Proposition II.3 completes the proof. ■

From conclusions of Proposition II.3 Proposition II.4 we have,

$$\begin{aligned} D^{\text{MM}}(\Gamma^n\|\pi) &= \langle\Gamma^n, \frac{\check{\pi}^{\hat{r}^n}}{\pi}\rangle \\ &= \max_{\nu\in\mathcal{E}_\pi}\langle\Gamma^n, \frac{\nu}{\pi}\rangle \\ &= \max_{\nu\in\mathcal{E}_\pi}\frac{1}{n}\sum_{i=1}^n \log\frac{\nu(Z_i)}{\pi(Z_i)}. \end{aligned}$$

In general when the supremum in the definition of $D^{\text{MM}}(\Gamma^n\|\pi)$ may not be achieved, the maxima in the above equations become replaced with suprema and we have the following identity:

$$D^{\text{MM}}(\Gamma^n\|\pi) = \sup_{\nu\in\mathcal{E}_\pi} \frac{1}{n}\sum_{i=1}^{n} \log\frac{\nu(Z_i)}{\pi(Z_i)}.$$

Thus the test statistic used in the mismatched test of (6) is exactly the generalized likelihood ratio between the family of distributions $\mathcal{E}_{\pi^0}$ and π^0 where

$$\mathcal{E}_{\pi^0} = \{\pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)) : r \in \mathbb{R}^d\}.$$

More structure can be established when the function class is linear.

D. Linear function class and I-projection

The mismatched-divergence introduced in [6] was restricted to a linear function class. Let $\{\psi_i : 1 \le i \le d\}$ denote d functions on Z. Let $\psi = (\psi_1, \dots, \psi_d)^T$ and let $f_r = r^T\psi$ in the definition (9):

$$\mathcal{F} = \Big\{ f_r = \sum_{i=1}^{d} r_i\psi_i : r \in \mathbb{R}^d \Big\}. \tag{18}$$

Proposition II.3 expresses $D^{\text{MM}}(\mu\|\pi)$ as a difference between the ordinary divergence and the value of a reverse I-projection $\inf_{\nu\in\mathcal{E}_\pi} D(\mu\|\nu)$. The next result establishes a characterization in terms of a (forward) I-projection. For a given vector $c \in \mathbb{R}^d$ we let $\mathbb{P}$ denote the *moment class*

$$\mathbb{P} = \{\nu \in \mathcal{P}(\mathsf{Z}) : \nu(\psi) = c\}. \tag{19}$$

Proposition II.5. *Suppose that the supremum in the definition of* $D^{\text{MM}}(\mu\|\pi)$ *is achieved at some* $r^* \in \mathbb{R}^d$. *Then,*

(i) *The distribution* $\nu^* := \check{\pi}^{r^*} \in \mathcal{E}_\pi$ *satisfies,*

$$D^{\text{MM}}(\mu\|\pi) = D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\},$$

where $\mathbb{P}$ *is defined using* $c = \mu(\psi)$.

(ii) $D^{\text{MM}}(\mu\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}$, *where* g^* *is given in* (13) *with* $f = r^{*T}\psi$, *and* $\eta = D^{\text{MM}}(\mu\|\pi)$.

Proof: Since the supremum is achieved, the gradient must vanish by the first order condition for optimality:

$$\nabla\big(\mu(f_r) - \Lambda_\pi(f_r)\big)\Big|_{r=r^*} = 0$$

The gradient is computable, and the identity above can thus be expressed $\mu(\psi) - \check{\pi}^{r^*}(\psi) = 0$. That is, the first order condition for optimality is equivalent to the constraint $\check{\pi}^{r^*} \in \mathbb{P}$. Consequently,

$$\begin{aligned} D(\nu^* \| \pi) &= \langle \check{\pi}^{r^*}, \log \frac{\check{\pi}^{r^*}}{\pi} \rangle \\ &= \check{\pi}^{r^*}(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) \\ &= \mu(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) = D^{\mathrm{MM}}(\mu\|\pi) \end{aligned}$$

Furthermore, by the convexity of $\Lambda_\pi(f_r)$ in r, it follows that the optimal r^* in the definition of $D^{\mathrm{MM}}(\nu\|\pi)$ is the same for all $\nu \in \mathbb{P}$. Hence, it follows by the Pythagorean equality of Proposition II.3 that

$$D(\nu\|\pi) = D(\nu\|\nu^*) + D(\nu^*\|\pi), \text{ for all } \nu \in \mathbb{P}.$$

Minimizing over $\nu \in \mathbb{P}$ it follows that ν^* is the I-projection of π onto $\mathbb{P}$:

$$D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\}$$

which gives (i).

To establish (ii), note first that by (i) and the inclusion $\mathbb{P} \subset \mathcal{H}_{g^*}$ we have,

$$D^{\mathrm{MM}}(\mu\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\} \geq \inf\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}$$

The reverse inequality follows from Proposition II.2 (i), and moreover the infimum is achieved with ν^*. ■

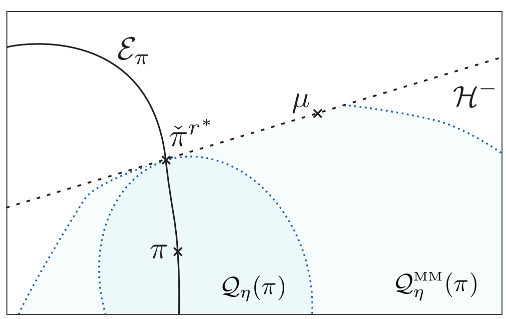

Fig. 2. *Interpretations of the mismatched divergence for a linear function class.* The distribution $\check{\pi}^{r^*}$ is the I-projection of π onto a hyperplane $\mathcal{H}_{g^*}$. It is also the reverse I-projection of μ onto the exponential family $\mathcal{E}_\pi$.

The geometry underlying mismatched divergence for a linear function class is illustrated in Figure 2. Suppose that the assumptions of Proposition II.5 hold, so that the supremum in (17) is achieved at r^*. Let $\eta = D^{\mathrm{MM}}(\mu\|\pi) = \mu(f_{r^*}) - \Lambda_\pi(f_{r^*})$, and $g^* = f_{r^*} - \big(\eta + \Lambda_\pi(f_{r^*})\big)$. Proposition II.2 implies that $\mathcal{H}_{g^*}$ defines a hyperplane passing through μ, with $\mathcal{Q}_\eta(\pi) \subset \mathcal{Q}_\eta^{\mathrm{MM}}(\pi) \subset \mathcal{H}_{g^*}^-$. This is strengthened in the linear case by Proposition II.5, which states that $\mathcal{H}_{g^*}$ supports $\mathcal{Q}_\eta(\pi)$ at the distribution $\check{\pi}^{r^*}$. Furthermore Proposition II.3 asserts that the distribution $\check{\pi}^{r^*}$ minimizes $D(\mu\|\check{\pi})$ over all $\check{\pi} \in \mathcal{E}_\pi$.

E. Log-linear function class and robust hypothesis testing

In the prior work [8], [9] the following relaxation of entropy is considered,

$$D^{\text{ROB}}(\mu\|\pi) := \inf_{\nu\in\mathbb{P}} D(\mu\|\nu) \tag{20}$$

where the moment class $\mathbb{P}$ is defined in (19) with $c = \pi(\psi)$, for a given collection of functions $\{\psi_i : 1 \le i \le d\}$. The associated universal test solves a min-max robust hypothesis testing problem.

We show here that D^{ROB} coincides with D^{MM} for a particular function class. It is described as (9) in which each function f_r is of the log-linear form,

$$f_r = \log(1 + r^T\psi)$$

subject to the constraint that $1 + r^T\psi(z)$ is strictly positive for each z.

Proposition II.6. *For a given $\pi \in \mathcal{P}(\mathsf{Z})$, suppose that the log-linear function class $\mathcal{F}$ is chosen with functions $\{\psi_i\}$ satisfying $\pi(\psi) = 0$. Suppose that the moment class used in the definition of D^{ROB} is chosen consistently, with $c = 0$. We then have for each $\mu \in \mathcal{P}(\mathsf{Z})$,*

$$D^{\text{MM}}(\mu\|\pi) = D^{\text{ROB}}(\mu\|\pi)$$

Proof: For each $\mu \in \mathcal{P}(\mathsf{Z})$, we obtain the following identity by applying Theorem 1.4 in [9],

$$\inf_{\nu\in\mathbb{P}} D(\mu\|\nu) = \sup\{\mu(\log(1 + r^T\psi)) : 1 + r^T\psi(z) > 0 \text{ for all } z \in \mathsf{Z}\}$$

Moreover, under the assumption that $\pi(\psi) = 0$ we obtain,

$$\Lambda_\pi(\log(1 + r^T\psi)) = \log(\pi(1 + r^T\psi)) = 0$$

Combining these identities gives,

$$\begin{aligned} D^{\text{ROB}}(\mu\|\pi) &:= \inf_{\nu\in\mathbb{P}} D(\mu\|\nu) \\ &= \sup\{\mu(\log(1 + r^T\psi)) - \Lambda_\pi(\log(1 + r^T\psi)) : 1 + r^T\psi(z) > 0 \text{ for all } z \in \mathsf{Z}\} \\ &= \sup_{f\in\mathcal{F}}\{\mu(f) - \Lambda_\pi(f)\} = D^{\text{MM}}(\mu\|\pi) \end{aligned}$$

■

III. Asymptotic Statistics

Theorem III.1 establishes a drawback of the Hoeffding test: the test statistic suffers from large bias and variance when the alphabet size N is large.

Theorem III.1. *Let $\pi^0, \pi^1 \in \mathcal{P}(\mathsf{Z})$ have full supports over Z.*

(i) *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal π^0. Then the normalized Hoeffding test statistic sequence $\{nD(\Gamma^n \| \pi^0) : n \geq 1\}$ has the following asymptotic bias and variance:*

$$\lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{21}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{22}$$

where $N = |\mathsf{Z}|$ denotes the size (cardinality) of Z. Furthermore, the following weak convergence result holds:

$$nD(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \tfrac{1}{2}\chi^2_{N-1} \tag{23}$$

(ii) *Suppose the sequence $\boldsymbol{Z}$ is drawn i.i.d. under $\pi^1 \neq \pi^0$. Then we have,*

$$\lim_{n\to\infty} \mathsf{E}\big[n\big(D(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0)\big)\big] = \tfrac{1}{2}(N-1)$$

□

The bias result of (21) follows from the unpublished report [13] and the weak convergence result of (23) follows from the result of [14]. All the results of the theorem, including (22) also follow from Theorem III.3 — We elaborate on this later in this section.

The weak convergence result of (23) and other similar results established later in this paper can be used to set thresholds for a finite sample test designed for a particular probability of false alarm (see for example, [7, p. 457]). For a large enough n, the distribution of the test statistic can be approximated by its limiting distribution whose tail probabilities can be obtained from tables of the χ^2 distribution.

We see from Theorem III.1 that the bias of the divergence statistic $D(\Gamma^n\|\pi^0)$ decays as $\frac{N-1}{2n}$, irrespective of the true distribution of the observations. One could argue that the problem of high bias in the Hoeffding test statistic can be addressed by setting a higher threshold. However, we also notice that when the observations are drawn under π^0, the variance of the divergence statistic decays as $\frac{N-1}{2n^2}$, which can be significant when N is of the order of n^2. This is a more serious flaw of the Hoeffding test for large alphabet sizes, since it cannot be addressed as easily. The high variance indicates that the

Hoeffding test is not reliable in situations where the alphabet size is of the same order as the square of the sequence length.

We now analyze the asymptotic statistics of the mismatched test. We require further assumptions regarding the function class $\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}$ to establish these results. Note that the second and third assumptions given below involve a distribution $\mu^0 \in \mathcal{P}(\mathsf{Z})$, and a vector $s \in \mathbb{R}^d$. We will make specialized versions of these assumptions in establishing our results, based on specific values of μ^0 and s.

Assumptions

(A1) $f_r(z)$ is C^2 in r for each $z \in \mathsf{Z}$

(A2) There exists an open neighborhood $B \subset \mathcal{P}(\mathsf{Z})$ of μ^0 such that for each $\mu \in B$, the supremum in the definition of $D^{\text{MM}}(\mu \| \mu^0)$ in (8) is achieved at a unique point $r(\mu)$.

(A3) The vectors $\{\psi_0, \dots, \psi_d\}$ are linearly independent over the support of μ^0, where $\psi_0 \equiv 1$, and for each $i \geq 1$

$$\psi_i(z) = \frac{\partial}{\partial r_i} f_r(z)\Big|_{r=s}, \qquad z \in \mathsf{Z}. \tag{24}$$

The linear-independence assumption in (A3) is defined as follows: If there are constants $\{a_0, \dots, a_d\}$ satisfying $\sum_i \psi_i(z) = 0$ a.e. $[\mu^0]$, then $a_i = 0$ for each i. In the case of a linear function class, the functions $\{\psi_i, i \geq 1\}$ defined in (24) are just the basis functions in (18). Lemma III.2 provides an alternate characterization of Assumption (A3).

For any $\mu \in \mathcal{P}(\mathsf{Z})$ define the covariance matrix Σ_μ via,

$$\Sigma_\mu(i,j) = \mu(\psi_i \psi_j) - \mu(\psi_i)\mu(\psi_j), \qquad 1 \leq i,j \leq d. \tag{25}$$

We use $\mathsf{Cov}_\mu(g)$ to denote the covariance of an arbitrary real-valued function g under μ:

$$\mathsf{Cov}_\mu(g) := \mu(g^2) - \mu(g)^2 \tag{26}$$

Lemma III.2. *Assumption (A3) holds if and only if* $\Sigma_{\mu^0} > 0$.

Proof: We evidently have $v^T \Sigma_{\mu^0} v = \mathsf{Cov}_{\mu^0}(v^T \psi) \geq 0$ for any vector $v \in \mathbb{R}^d$. Hence, we have the following equivalence: For any $v \in \mathbb{R}^d$, on denoting $c_v = \mu^0(v^T \psi)$,

$$v^T \Sigma_{\mu^0} v = 0 \quad \Leftrightarrow \quad \sum_{i=1}^d v_i \psi_i(z) = c_v \quad \text{a.e. } [\mu^0]$$

The conclusion of the lemma follows. ■

We now present our main asymptotic results. Theorem III.3 identifies the asymptotic bias and variance of the mismatched test statistic under the null hypothesis, and also under the alternate hypothesis. A key observation is that the asymptotic bias and variance does not depend on N.

Theorem III.3. *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal π. Suppose that there exists r^* satisfying $f_{r^*} = \log(\pi/\pi^0)$. Further, suppose that Assumptions (A1), (A2), (A3) hold with $\mu^0 = \pi$ and $s = r^*$. Then,*

(i) *When $\pi = \pi^0$,*

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d \tag{27}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d \tag{28}$$

$$nD^{\text{MM}}(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \tfrac{1}{2}\chi^2_d$$

(ii) *When $\pi = \pi^1 \neq \pi^0$, we have with $\sigma_1^2 := \mathsf{Cov}_{\pi^1}(f_{r^*})$,*

$$\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0))] = \tfrac{1}{2}d \tag{29}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}}D^{\text{MM}}(\Gamma^n\|\pi^0)] = \sigma_1^2 \tag{30}$$

$$n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0)) \xrightarrow[n\to\infty]{d.} \mathcal{N}(0,\sigma_1^2). \tag{31}$$

In part (ii) of Theorem III.3, the assumption that r^* exists implies that π^1 and π^0 have equal supports. Furthermore, if Assumption (A3) holds in part (ii), then a sufficient condition for Assumption (A2) is that the function $V(r) := (-\pi^1(f_r) + \Lambda_{\pi^0}(f_r))$ be coercive in r. And, under (A3), the function V is strictly convex and coercive in the following settings: (i) If the function class is linear, or (ii) the function class is log-linear, and the two distributions π^1 and π^0 have common support. We will use this fact in Proposition III.5 for the linear function class.

The weak convergence results in Theorem III.3(i) can be derived from Clarke and Barron [13], [15] (see also [7, Theorem 4.2]), following the maximum-likelihood estimation interpretation of the mismatched test obtained in Proposition II.4.

The following lemma will be used to deduce part (ii) of the theorem from part (i).

Lemma III.4. *Let $D^{\text{MM}}_{\mathcal{F}}$ denote the mismatched divergence defined using function class $\mathcal{F}$. Suppose $\pi^1 \prec \pi^0$ and the supremum in the definition of $D^{\text{MM}}_{\mathcal{F}}(\pi^1\|\pi^0)$ is achieved at some $f_{r^*} \in \mathcal{F}$. Let $\check{\pi} = \pi^0 \exp(f_{r^*} - \Lambda_{\pi^0}(f_{r^*}))$ and $\mathcal{G} = \mathcal{F} - f_{r^*} := \{f_r - f_{r^*} : r \in \mathbb{R}^d\}$. Then for and μ satisfying $\mu \prec \pi^0$, we* *have*

$$D_{\mathcal{F}}^{\mathrm{MM}}(\mu\|\pi^0) = D_{\mathcal{F}}^{\mathrm{MM}}(\pi^1\|\pi^0) + D_{\mathcal{G}}^{\mathrm{MM}}(\mu\|\check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle. \tag{32}$$

Proof: In the following chain of identities, the first, third and fifth equalities follow from Proposition II.3.

$$\begin{aligned}
D_{\mathcal{F}}^{\mathrm{MM}}(\mu\|\pi^0) &= D(\mu\|\pi^0) - \inf\{D(\mu\|\nu) : \nu = \pi^0 \exp(f - \Lambda_{\pi^0}(f)), f \in \mathcal{F}\} \\
&= D(\mu\|\check{\pi}) + \langle \mu, \log(\frac{\check{\pi}}{\pi^0})\rangle - \inf\{D(\mu\|\nu) : \nu = \check{\pi} \exp(f - \Lambda_{\check{\pi}}(f)), f \in \mathcal{G}\} \\
&= D_{\mathcal{G}}^{\mathrm{MM}}(\mu\|\check{\pi}) + \langle \mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D_{\mathcal{G}}^{\mathrm{MM}}(\mu\|\check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle + D(\pi^1\|\pi^0) - D(\pi^1\|\check{\pi}) \\
&= D_{\mathcal{G}}^{\mathrm{MM}}(\mu\|\check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle + D_{\mathcal{F}}^{\mathrm{MM}}(\pi^1\|\pi^0)
\end{aligned}$$

■

Now we apply the decomposition result from Lemma III.4 to the type of the observation sequence $\boldsymbol{Z}$, assumed to be drawn i.i.d. with marginal π^1. If there exists r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$, then we have $\check{\pi} = \pi^1$. The decomposition becomes

$$D_{\mathcal{F}}^{\mathrm{MM}}(\Gamma^n\|\pi^0) = D_{\mathcal{F}}^{\mathrm{MM}}(\pi^1\|\pi^0) + D_{\mathcal{G}}^{\mathrm{MM}}(\Gamma^n\|\pi^1) + \langle \Gamma^n - \pi^1, f_{r^*}\rangle. \tag{33}$$

For large n, the second term in the decomposition (33) has a mean of order n^{-1} and variance of order n^{-2}, as shown in part (i) of Theorem III.3. The third term has zero mean and variance of order n^{-1}, since by the Central Limit Theorem,

$$n^{\frac{1}{2}}\langle \Gamma^n - \pi^1, f_{r^*}\rangle \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \mathsf{Cov}_{\pi^1}(f_{r^*})). \tag{34}$$

Thus, the asymptotic variance of $D_{\mathcal{F}}^{\mathrm{MM}}(\Gamma^n\|\pi^0)$ is dominated by that of the third term and the asymptotic bias is dominated by that of the second term.

Since the divergence can be interpreted as a special case of mismatched divergence defined with respect to a linear function class, the results of Theorem III.3 can also be specialized to obtain results on the Hoeffding test statistic. To satisfy the uniqueness condition of Assumption (A2), we require that the function class should not contain any constant functions. Now suppose that the span of the linear function class $\mathcal{F}$ together with the constant function $f^0 \equiv 1$ spans the set of all functions on Z. This together with (A3) would imply that $d = N - 1$, where N is the size of the alphabet Z. It follows from Proposition II.1 that for such a function class the mismatched divergence coincides with the divergence. Thus, an application of Theorem III.3 (i) gives rise to the results stated in Theorem III.1.

The assumption of the existence of r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$ in part (ii) of Theorem III.3 can be relaxed. In the case of a linear function class we have the following extension of part (ii).

Proposition III.5. *Suppose that the observation sequence* $\boldsymbol{Z}$ *is drawn i.i.d. with marginal* π^1 *satisfying* $\pi^1 \prec \pi^0$. *Let* $\mathcal{F}$ *be the linear function class defined in (18). Suppose the supremum in the definition of* $D^{\text{MM}}(\pi^1\|\pi^0)$ *is achieved at some* $r^1 \in \mathbb{R}^d$. *Further, suppose that the functions* $\{\psi_i\}$ *satisfy the linear independence condition of Assumption (A3) with* $\mu^0 = \pi^1$. *Then we have,*

$$\begin{aligned}\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi^1\|\pi^0))] &= \tfrac{1}{2}\mathsf{trace}(\Sigma_{\pi^1}\Sigma_{\check{\pi}}^{-1}) \\ \lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}} D^{\text{MM}}(\Gamma^n\|\pi^0)] &= \sigma_1^2 \\ n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi^1\|\pi^0)) &\xrightarrow[n\to\infty]{d.} \mathcal{N}(0,\sigma_1^2)\end{aligned}$$

where in the first limit $\check{\pi} = \pi^0\exp(f_{r^1} - \Lambda_{\pi^0}(f_{r^1}))$, *and* Σ_{π^1} *and* $\Sigma_{\check{\pi}}$ *are defined as in (25). In the second two limits* $\sigma_1^2 = \mathsf{Cov}_{\pi^1}(f_{r^1})$. ☐

Although we have not explicitly imposed Assumption (A2) in Proposition III.5, the argument we presented following Theorem III.3 ensures that when $\pi^1 \prec \pi^0$, Assumption (A2) is satisfied whenever Assumption (A3) holds. Furthermore, it can be shown that the achievement of the supremum required in Proposition III.5 is guaranteed if π^1 and π^0 have equal supports. We also note that the vector s appearing in eq. (24) of Assumption (A3) is arbitrary when the parametrization of the function class is linear.

To prove Theorem III.3 and Proposition III.5 we need the following lemmas, whose proofs are given in the Appendix.

Lemma III.6. *Let* $\boldsymbol{X} = \{X^i : i = 1,2,\dots\}$ *be an i.i.d. sequence with mean* $\bar{x}$ *taking values in a compact convex set* $\mathsf{X} \subset \mathbb{R}^m$, *containing* $\bar{x}$ *as a relative interior point. Define* $S^n = \frac{1}{n}\sum_{i=1}^n X^i$. *Suppose we are given a function* $h : \mathbb{R}^m \mapsto \mathbb{R}$, *that is continuous over* X *and a compact set* K *containing* $\bar{x}$ *as a relative interior point such that*

1) *The gradient* $\nabla h(x)$ *and the Hessian* $\nabla^2 h(x)$ *are continuous over a neighborhood of* K.
2) $\lim_{n\to\infty} -\dfrac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0$.

Let $M = \nabla^2 h(\bar{x})$ *and* $\Xi = \mathsf{Cov}(X^1)$. *Then,*

(i) *The normalized asymptotic bias of* $\{h(S^n) : n \geq 1\}$ *is obtained via,*

$$\lim_{n\to\infty} n\mathsf{E}[h(S^n) - h(\bar{x})] = \tfrac{1}{2}\mathsf{trace}(M\Xi)$$

(ii) *If in addition to the above conditions, the directional derivative satisfies* $\nabla h(\bar{x})^{\intercal}(X^1 - \bar{x}) = 0$ *almost surely, then the asymptotic variance decays as* n^{-2}*, with*

$$\lim_{n\to\infty} \mathsf{Var}\,[nh(S^n)] = \tfrac{1}{2}\mathsf{trace}(M\Xi M\Xi)$$

□

Lemma III.7. *Suppose that the observation sequence* $\boldsymbol{Z}$ *is drawn i.i.d. with marginal* $\mu \in \mathcal{P}(\mathsf{Z})$*. Let* $h : \mathcal{P}(\mathsf{Z}) \mapsto \mathbb{R}$ *be a continuous real-valued function whose gradient and Hessian are continuous in a neighborhood of* μ*. If the directional derivative satisfies* $\nabla h(\mu)^{\intercal}(\nu - \mu) \equiv 0$ *for all* $\nu \in \mathcal{P}(\mathsf{Z})$*, then*

$$n(h(\Gamma^n) - h(\mu)) \xrightarrow[n\to\infty]{d.} \tfrac{1}{2}W^{\intercal}MW \tag{35}$$

where $M = \nabla^2 h(\mu)$ *and* $W \sim \mathcal{N}(0, \Sigma_W)$ *with* $\Sigma_W = diag(\mu) - \mu\mu^{\intercal}$*.* □

Lemma III.8. *Suppose that* V *is an* m*-dimensional,* $\mathcal{N}(0, I_m)$ *random variable, and* $D\colon \mathbb{R}^m \to \mathbb{R}^m$ *is a projection matrix. Then* $\xi := \|DV\|^2$ *is a chi-squared random variable with* K *degrees of freedom, where* K *denotes the rank of* D*.*

Proof: The assumption that D is a projection matrix implies that $D^2 = D$. Let $\{u^1, \dots, u^m\}$ denote an orthonormal basis, chosen so that the first K vectors span the range space of D. Hence $Du^i = u^i$ for $1 \le i \le K$, and $Du^i = 0$ for all other i.

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

To complete the proof we demonstrate that $\|D\widetilde{V}\|^2$ has a chi-squared distribution: By construction the vector $\widetilde{Y} = D\widetilde{V}$ has components given by

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

It follows that $\|\widetilde{Y}\|^2 = \|D\widetilde{V}\|^2 = \widetilde{V}_1^2 + \cdots + \widetilde{V}_K^2$ has a chi-squared distribution with K degrees of freedom. ■

Before we proceed to the proofs of Theorem III.3 and Proposition III.5, we recall the optimization problem (17) defining the mismatched divergence:

$$D^{\text{MM}}(\mu\|\pi^0) = \sup_{r\in\mathbb{R}^d} \bigl(\mu(f_r) - \Lambda_{\pi^0}(f_r)\bigr). \tag{36}$$

The first order condition for optimality is given by,

$$g(\mu, r) = 0 \tag{37}$$

where g is the vector valued function that defines the gradient of the objective function in (36):

$$\begin{aligned} g(\mu, r) &:= \nabla_r\big(\mu(f_r) - \Lambda_{\pi^0}(f_r)\big) \\ &= \mu(\nabla_r f_r) - \frac{\pi^0(e^{f_r}\nabla_r f_r)}{\pi^0(e^{f_r})} \end{aligned} \tag{38}$$

The derivative of $g(\mu, r)$ with respect to r is given by

$$\nabla_r g(\mu, r) = \mu(\nabla_r^2 f_r) - \left[\frac{\pi^0(e^{f_r}\nabla_r f_r \nabla_r f_r^{\scriptscriptstyle T}) + \pi^0(e^{f_r}\nabla_r^2 f_r)}{\pi^0(e^{f_r})} - \frac{\pi^0(e^{f_r}\nabla_r f_r)\pi^0(e^{f_r}\nabla_r f_r^{\scriptscriptstyle T})}{(\pi^0(e^{f_r}))^2}\right] \tag{39}$$

In these formulae we have extended the definition of $\mu(M)$ for matrix-valued functions M on Z via $[\mu(M)]_{ij} := \mu(M_{ij}) = \sum_z M_{ij}(z)\mu(z)$. On letting $\psi^r = \nabla_r f_r$ we obtain,

$$g(\mu, r) = \mu(\psi^r) - \check{\pi}^r(\psi^r) \tag{40}$$

$$\nabla_r g(\mu, r) = \mu(\nabla_r^2 f_r) - \check{\pi}^r(\nabla_r^2 f_r) - [\check{\pi}^r(\psi^r \psi^{r{\scriptscriptstyle T}}) - \check{\pi}^r(\psi^r)\check{\pi}^r(\psi^{r{\scriptscriptstyle T}})] \tag{41}$$

where the definition of the twisted distribution is as given in (16):

$$\check{\pi}^r := \pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)).$$

Proof of Theorem III.3: Without loss of generality, we assume that π^0 has full support over Z. Suppose that the observation sequence $\boldsymbol{Z}$ is drawn i.i.d. with marginal distribution $\pi \in \mathcal{P}(\mathsf{Z})$. We have $D^{\text{MM}}(\Gamma^n \| \pi^0) \xrightarrow[n\to\infty]{a.s.} D^{\text{MM}}(\pi \| \pi^0)$ by the law of large numbers.

1) *Proof of part (i):* We first prove the results concerning the bias and variance of the mismatched test statistic. We apply Lemma III.6 to the function $h(\mu) := D^{\text{MM}}(\mu \| \pi^0)$. The other terms appearing in the lemma are taken to be: $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \ldots, \mathbb{I}_{z_N}(Z_i))^{\scriptscriptstyle T}$, $\mathsf{X} = \mathcal{P}(\mathsf{Z})$, $\bar{x} = \pi^0$, and $S^n = \Gamma^n$. Let $\Xi = \mathsf{Cov}(X^1)$. It follows that $\Xi = \text{diag}(\pi^0) - \pi^0 \pi^{0{\scriptscriptstyle T}}$ and $\Sigma_{\pi^0} = \Psi \Xi \Psi^{\scriptscriptstyle T}$, where Σ_{π^0} is defined in (25), and Ψ is a $d \times N$ matrix defined by,

$$\Psi(i, j) = \psi_i(z_j). \tag{42}$$

This can be expressed as the concatenation of column vectors via $\Psi = [\psi(z_1), \psi(z_2), \ldots, \psi(z_N)]$.

We first demonstrate that

$$M = \nabla^2 h(\pi_0) = \Psi^{\scriptscriptstyle T}(\Sigma_{\pi^0})^{-1}\Psi, \tag{43}$$

and then check to make sure that the other requirements of Lemma III.6 are satisfied. The first two conclusions of Theorem III.3 (i) will then follow from Lemma III.6, since

$$\mathsf{trace}(M\Xi) = \mathsf{trace}((\Sigma_{\pi^0})^{-1}\Psi\Xi\Psi^{\scriptscriptstyle T}) = \mathsf{trace}(I_d) = d,$$

and similarly $\mathsf{trace}(M\Xi M\Xi) = \mathsf{trace}(I_d) = d$.

We first prove that under the assumptions of Theorem III.3 (i), there is a function $r : \mathcal{P}(\mathsf{Z}) \mapsto \mathbb{R}$ that is C^1 in a neighborhood of π^0 such that $r(\mu)$ solves (36) for μ in this neighborhood. Under the uniqueness assumption (Assumption (A2)) the function $r(\mu)$ coincides with the function given in (A2).

By the assumptions, we know that when $\mu = \pi^0$, (37) is satisfied by r^* with $f_{r^*} \equiv 0$. It follows that $\pi^0 = \check{\pi}^{r^*}$. Substituting this into (41), we obtain $\nabla_r g(\mu, r)\Big|_{\substack{\mu=\pi^0 \\ r=r^*}} = -\Sigma_{\pi^0}$, which is negative-definite by Assumption (A3) and Lemma III.2. Therefore, by the Implicit Function Theorem, there is an open neighborhood U around $\mu = \pi^0$, an open neighborhood V of r^*, and a continuously differentiable function $r : U \to V$ that satisfies $g(\mu, r(\mu)) = 0$, for $\mu \in U$. This fact together with Assumptions (A2) and (A3) ensure that when $\mu \in U \cap B$, the vector $r(\mu)$ uniquely achieves the supremum in (36).

Taking the total derivative of (37) with respect to $\mu(z)$ we get,

$$\frac{\partial r(\mu)}{\partial \mu(z)} = -\Big[\nabla_r g(\mu, r(\mu))\Big]^{-1} \frac{\partial g(\mu, r(\mu))}{\partial \mu(z)}. \tag{44}$$

Consequently, when $\mu = \pi^0$,

$$\frac{\partial r(\mu)}{\partial \mu(z)}\bigg|_{\mu=\pi^0} = \Sigma_{\pi^0}^{-1} \psi(z). \tag{45}$$

These results enable us to identify the first and second order derivative of $h(\mu) = D^{\text{MM}}(\mu \| \pi^0)$. Applying $g(\mu, r(\mu)) = 0$, we obtain the derivatives of h as follows,

$$\frac{\partial}{\partial \mu(z)} h(\mu) = f_{r(\mu)}(z). \tag{46}$$

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\mu) = (\nabla_r f_{r(\mu)}(z))^T \frac{\partial r(\mu)}{\partial \mu(\bar{z})}. \tag{47}$$

When $\mu = \pi^0$, substituting (45) in (47), we obtain (43).

We now verify the remaining conditions required for applying Lemma III.6:

(a) It is straightforward to see that $h(\pi^0) = 0$.

(b) The function h is uniformly bounded since $h(\mu) = D^{\text{MM}}(\mu \| \pi^0) \le D(\mu \| \pi^0) \le \max_z \log(\frac{1}{\pi^0(z)})$ and π^0 has full support.

(c) Since $f_{r(\mu)} = 0$ when $\mu = \pi^0$, it follows by (46) that $\frac{\partial}{\partial \mu(z)} h(\mu)\Big|_{\mu=\pi^0} = 0$.

(d) Pick a compact $K \subset U \cap B$ so that K contains π^0 as an interior point, and $K \subset \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \frac{1}{2} \min_u |\pi^0(u)|\}$. This choice of K ensures that $\lim_{n\to\infty} -\frac{1}{n} \log \mathsf{P}\{S^n \notin K\} > 0$. Note that since $r(\mu)$ is continuously differentiable on $U \cap B$, it follows by (46) and (47) that h is C^2 on K.

Thus the results on convergence of the bias and variance follow from Lemma III.6.

The weak convergence result is proved using Lemma III.7. We observe that the covariance matrix of the Gaussian vector W given in Lemma III.7 is $\Sigma_W = \Xi = \text{diag}(\pi^0) - \pi^0\pi^{0T}$. This does not have full rank since $\Xi\mathbf{1} = 0$, where $\mathbf{1}$ is the $N \times 1$ vector of ones. Hence we can write,

$$\Xi = GG^T$$

where G is an $N \times k$ matrix for some $k < N$. In fact, since the support of π^0 is full, we have $k = N-1$ (see Lemma III.2). Based on this representation we can write $W = GV$, where $V \sim \mathcal{N}(0, I_k)$.

Now, by the lemma, the limiting random variable is given by $U := \frac{1}{2}W^T MW = \frac{1}{2}V^T G^T MGV$, where $M = \nabla^2_\mu D^{\text{MM}}(\mu\|\pi^0)\Big|_{\pi^0} = \Psi^T(\Psi\Xi\Psi^T)^{-1}\Psi$. We observe that the matrix $G^T MG$ satisfies $(G^T MG)^2 = G^T MG$. Applying Lemma III.8 with $D = G^T MG$ and noting that Ψ has rank d under Assumption (A3), we conclude that $U \sim \frac{1}{2}\chi^2_d$.

2) *Proof of part (ii):*The conclusion of part (ii) is derived using part (i) and the decomposition in (33). We will study the bias, variance, and limiting distribution of each term in the decomposition.

For the second term, note that the dimensionality of the function class $\mathcal{G}$ is also d. Applying part (i) of this theorem to $D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)$, we conclude that its asymptotic bias and variance are given by

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)] = \tfrac{1}{2}d, \tag{48}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)] = \tfrac{1}{2}d. \tag{49}$$

For the third term, since $\boldsymbol{Z}$ is i.i.d. with marginal π^1, we have

$$\mathsf{E}[\langle\Gamma^n - \pi^1, f_{r^*}\rangle] = 0, \tag{50}$$

$$\mathsf{Var}\,[n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, f_{r^*}\rangle] = \mathsf{Cov}_{\pi^1}(f_{r^*}). \tag{51}$$

The bias result (29) follows by combining (48), (50) and using the decomposition (33). To prove the variance result (30), we again apply the decomposition (33) to obtain,

$$\begin{aligned}\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}}D^{\text{MM}}_{\mathcal{F}}(\Gamma^n\|\pi^0)] &= \lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}}D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)] + \mathsf{Var}\,[n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, f_{r^*}\rangle] \\ &\quad + \lim_{n\to\infty} 2E\big[n^{\frac{1}{2}}\big(D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1) - \mathsf{E}[D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)]\big)n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, f_{r^*}\rangle\big]\end{aligned} \tag{52}$$

From (49) it follows that the first term on the right hand side of (52) is 0. The third term is also 0 by applying the Cauchy–Schwarz-Bunyakovsky inequality. Thus, (52) together with (51) gives (30).

Finally, we prove the weak convergence result (31) by again applying the decomposition (33). By (48) and (49), we conclude that the second term $n^{\frac{1}{2}}D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)$ converges in mean square to 0 as $n \to \infty$. The weak convergence of the third term is given in (34). Applying Slutsky's theorem, we obtain (31). ■

Proof of Proposition III.5: The proof of this result is very similar to that of Theorem III.3 (ii) except that we use the decomposition in (32) with $\mu = \Gamma^n$.

We first prove the following generalization of (48) and (49):

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\breve{\pi})] = \tfrac{1}{2}\mathsf{trace}(\Sigma_{\pi^1}(\Sigma_{\breve{\pi}})^{-1}) \tag{53}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\breve{\pi})] = \tfrac{1}{2}\mathsf{trace}(\Sigma_{\pi^1}(\Sigma_{\breve{\pi}})^{-1}\Sigma_{\pi^1}(\Sigma_{\breve{\pi}})^{-1}) \tag{54}$$

where $\mathcal{G} = \mathcal{F} - f_{r^1}$, and $\breve{\pi}$ is defined in the statement of the proposition. The argument is similar to that of Theorem III.3 (i): We denote $\tilde{f}_r := f_r - f_{r^1}$, and define $h(\mu) := D^{\text{MM}}_{\mathcal{G}}(\mu\|\breve{\pi}) = \sup_{r\in\mathbb{R}^d}(\mu(\tilde{f}_r) - \Lambda_{\breve{\pi}}(\tilde{f}_r))$. To apply Lemma III.6, we prove the following

$$h(\pi^1) = 0, \tag{55}$$

$$\nabla_\mu h(\pi^1) = 0, \tag{56}$$

$$\textit{and} \qquad M = \nabla^2_\mu h(\pi^1) = \Psi^T(\Sigma_{\breve{\pi}})^{-1}\Psi. \tag{57}$$

The last two inequalities (56) and (57) are analogous to (46) and (47). We can also verify that the rest of the conditions of Lemma III.6 hold. This establishes (53) and (54).

To prove (55), first note that the supremum in the optimization problem defining $D^{\text{MM}}(\pi^1\|\breve{\pi})$ is achieved by $\tilde{f}_{r^1}$, and we know by definition that $\tilde{f}_{r^1} = 0$. Together with the definition $D^{\text{MM}}(\pi^1\|\breve{\pi}) = \pi^1(\tilde{f}_{r^1}) - \Lambda_{\breve{\pi}}(\tilde{f}_r)$, we obtain (55).

Redefine $g(\mu, r) := \nabla_r(\mu(\tilde{f}_r) - \Lambda_{\breve{\pi}}(\tilde{f}_r))$. The first order optimality condition of the optimization problem defining $D^{\text{MM}}(\mu\|\breve{\pi})$ gives $g(\mu, r) = 0$. The assumption that $\mathcal{F}$ is a linear function class implies that $\tilde{f}_r$ is linear in r. Consequently $\nabla^2_r\tilde{f}_r = 0$. By the same argument that leads to (39), we can show that

$$\nabla_r g(\mu, r) = -\left[\frac{\breve{\pi}(e^{\tilde{f}_r}\nabla_r\tilde{f}_r\nabla_r\tilde{f}_r^T)}{\breve{\pi}(e^{\tilde{f}_r})} - \frac{\breve{\pi}(e^{\tilde{f}_r}\nabla_r\tilde{f}_r)\breve{\pi}(e^{\tilde{f}_r}\nabla_r\tilde{f}_r^T)}{(\breve{\pi}(e^{\tilde{f}_r}))^2}\right] \tag{58}$$

Together with the fact that $\tilde{f}_{r^1} = 0$ and $\nabla_r\tilde{f}_r = \nabla_r f_r$, we obtain

$$\nabla_r g(\mu, r)\Big|_{\substack{\mu=\pi^1\\ r=r^1}} = -\Sigma_{\breve{\pi}}. \tag{59}$$

Proceeding as in the proof of Theorem III.3 (i), we obtain (56) and (57).

Now using similar steps as in the proof of Theorem III.3 (ii), and noticing that $\log(\frac{\breve{\pi}}{\pi^0}) = f_{r^1}$, we can establish the following results on the third term of (32):

$$\mathsf{E}[\langle\Gamma^n - \pi^1, \log(\frac{\breve{\pi}}{\pi^0})\rangle] = 0$$

$$\mathsf{Var}\,[n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, \log(\frac{\breve{\pi}}{\pi^0})\rangle] = \mathsf{Cov}_{\pi^1}(f_{r^1})$$

$$n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, \log(\frac{\breve{\pi}}{\pi^0})\rangle \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \mathsf{Cov}_{\pi^1}(f_{r^1})).$$

Continuing the same arguments as before, we obtain the result of Proposition III.5. ■

Interpretation of the asymptotic results The asymptotic results established above can be used to study the finite sample performance of the mismatched test and Hoeffding's test. Suppose the log-likelihood ratio $\log(\pi^1/\pi^0)$ lies in the function class $\mathcal{F}$. In this case, the results of Theorem III.3 and Lemma III.4 are informally summarized in the following approximations: With Γ^n denoting the empirical distributions of the i.i.d. process $\boldsymbol{Z}$,

$$D^{\text{MM}}(\Gamma^n\|\pi^0) \approx \begin{cases} D(\pi^0\|\pi^0) + \frac{1}{2}\frac{1}{n}\sum_{k=1}^d W_k^2\,, & Z_i \sim \pi^0 \\ D(\pi^1\|\pi^0) + \frac{1}{2}\frac{1}{n}\sum_{k=1}^d W_k^2 + \frac{1}{\sqrt{n}}\sigma_1 U\,, & Z_i \sim \pi^1 \end{cases} \tag{60}$$

where $\{W_k\}$ is i.i.d., $N(0,1)$, and U is also $N(0,1)$ but not independent of the W_k's. The standard deviation σ_1 is given in Theorem III.3. These distributional approximations are valid for large n, and are subject to assumptions on the function class used in the theorem. We observe from (60) that for large enough n the mismatched divergence is well approximated by a Gaussian random variable with fixed variance when the observations are drawn under π^1, and by one-half times a chi-squared random variable with d degrees of freedom when the observations are drawn under π^0. Thus we expect to see a better receiver operating characteristic (ROC) for a lower value of d. Since the mismatched test can be interpreted as a GLRT, these results capture the rate of degradation of the finite sample performance of a GLRT as the dimensionality of the parameterized family of alternate hypotheses increases.

As we saw earlier, the Hoeffding test can be interpreted as a special case of the mismatched test for a specific choice of a function class with $d = N - 1$. Thus the arguments presented above also suggest that the Hoeffding test is not expected to perform well when the alphabet size N is large relative to n^2.

To summarize, the above results suggest that although the Hoeffding test is asymptotically optimal, it is disadvantageous in terms of finite sample performance to blindly use the Hoeffding test if it is known a priori that the alternate distribution belongs to some parameterized family of distributions. Furthermore, even if no prior information about the alternate distribution is available, the lower variance of the mismatched test might make it the preferred choice in practical applications.

IV. CONCLUSIONS

The mismatched test provides a solution to the universal hypothesis testing problem that can incorporate prior knowledge in order to reduce variance. The main results of Section III show that the variance reduction over Hoeffding's optimal test is substantial when the state space is large.

The dimensionality of the function class can be chosen by the designer to ensure that the the bias and variance are within tolerable limits. It is in this phase of design that prior knowledge is required to ensure that the error-exponent remains sufficiently large under the alternate hypothesis (see e.g. Corollary II.1). In this way the designer can make effective tradeoffs between the power of the test and the variance of the test statistic.

The mismatched divergence provides a unification of several approaches to robust and universal hypothesis testing. Although performance analysis is restricted to an i.i.d. setting, generalization to stationary processes satisfying the Central Limit Theorem is straightforward. Moreover, the resulting tests are applicable in very general settings.

There are many directions for future research. Topics of current research include,

(i) Algorithms for basis synthesis and basis adaptation.

(ii) Extensions to Markovian models.

(iii) Extensions to change detection.

We are also actively pursuing applications to problems surrounding building energy and surveillance. Some initial progress is reported in [16].

Acknowledgements

We thank Prof. Imre Csiszár for his insightful comments.

Appendix

A. Proof of Lemma III.6

The following simple lemma will be used in several places in the proofs that follow.

Lemma A.1. *If a sequence of random variables* $\{A^n\}$ *satisfies* $\mathsf{E}[A^n] \xrightarrow[n\to\infty]{} a$ *and* $\{\mathsf{E}[(A^n)^2]\}$ *is a bounded sequence, and another sequence of random variables* $\{B^n\}$ *satisfies* $B^n \xrightarrow[n\to\infty]{m.s.} b$*, then* $\mathsf{E}[A^n B^n] \xrightarrow[n\to\infty]{} ab$*.* □

Proof of Lemma III.6: Without loss of generality, we can assume that the mean $\bar{x}$ is the origin in $\mathbb{R}^m$ and that $h(\bar{x}) = 0$.

Since the Hessian is continuous over the set K, we have by Taylor's theorem:

$$n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n\in K\}} = n[h(\bar{x}) + \tfrac{1}{2} S^{nT}\nabla^2 h(\widetilde{S}^n) S^n]\mathbb{I}_{\{S^n\in K\}} \tag{61}$$

$$= \frac{n}{2} S^{nT}\nabla^2 h(\widetilde{S}^n) S^n \mathbb{I}_{\{S^n\in K\}} \tag{62}$$

where $\widetilde{S}^n = \gamma S^n$ for some $\gamma = \gamma(n) \in [0,1]$. By the strong law of large numbers we have $S^n \xrightarrow[n\to\infty]{a.s.} \bar{x}$. Hence $\widetilde{S}^n \xrightarrow[n\to\infty]{a.s.} \bar{x}$ and $\nabla^2 h(\widetilde{S}^n) \xrightarrow[n\to\infty]{a.s.} \nabla^2 h(\bar{x}) = M$ since $\nabla^2 h$ is continuous at $\bar{x}$. Now by the boundedness of the second derivative over K and the fact that

$$\mathbb{I}_{\{S^n \in K\}} \xrightarrow[n\to\infty]{a.s.} 1$$

we have $(\nabla^2 h(\widetilde{S}^n))_{i,j}\mathbb{I}_{\{S^n\in K\}} \xrightarrow[n\to\infty]{m.s.} M_{i,j}$.

Under the assumption that $\boldsymbol{X}$ is i.i.d. on the compact set X, we have

$$\mathsf{E}[nS_i^n S_j^n] = \Sigma_{i,j} \text{ for all } n,$$

and $\mathsf{E}[(nS_i^n S_j^n)^2]$ converges to a finite quantity as $n \to \infty$. Hence the results of Lemma A.1 are applicable with $A^n = nS_i^n S_j^n$ and $B^n = \nabla^2 h(\widetilde{S}^n)_{i,j}\mathbb{I}_{\{S^n\in K\}}$, which gives:

$$\mathsf{E}[nS_i^n S_j^n \nabla^2 h(\widetilde{S}^n)_{i,j}\mathbb{I}_{\{S^n\in K\}}] \xrightarrow[n\to\infty]{} \Sigma_{i,j}M_{i,j} \tag{63}$$

Thus we have:

$$\mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n\in K\}}] = \mathsf{E}[\frac{n}{2} S^{nT}\nabla^2 h(\widetilde{S}^n)S^n \mathbb{I}_{\{S^n\in K\}}] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi) \tag{64}$$

Since X is compact, h is continuous, and h is differentiable at $\bar{x}$, it follows that there are scalars $\overline{h}$ and $\overline{x}$ such that $\sup_{x\in\mathsf{X}} |h(x)| \le \overline{h}$ and $|\nabla h(\bar{x})^T S^n| < \overline{x}$. Hence,

$$|\mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n\notin K\}}]| \le n(\overline{h} + \overline{x})\mathsf{P}\{S^n \notin K\} \xrightarrow[n\to\infty]{} 0 \tag{65}$$

where we use the assumption that the $\mathsf{P}\{S^n \notin K\}$ decays exponentially in n. Combining (64) and (65) and using the fact that S^n has zero mean, we have

$$\mathsf{E}[nh(S^n)] = \mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi)$$

This establishes the result of (i).

Under the condition that the directional derivative is zero, (62) can be written as

$$nf(S^n)\mathbb{I}_{\{S^n\in K\}} = \frac{n}{2} S^{nT}\nabla^2 h(\widetilde{S}^n)S^n \mathbb{I}_{\{S^n\in K\}} \tag{66}$$

Now by squaring (66), we have

$$(nh(S^n)\mathbb{I}_{\{S^n\in K\}})^2 = \frac{n^2}{4}\sum_{i,j,k,\ell} S_i^n(\nabla^2 h(\widetilde{S}^n))_{i,j}S_j^n S_k^n(\nabla^2 h(\widetilde{S}^n))_{k,\ell}S_\ell^n \mathbb{I}_{\{S^n\in K\}} \tag{67}$$

As before, by the boundedness of the Hessian we have:

$$(\nabla^2 h(\widetilde{S}^n))_{i,j}(\nabla^2 h(\widetilde{S}^n))_{k,\ell}\mathbb{I}_{\{S^n\in K\}} \xrightarrow[n\to\infty]{m.s.} M_{i,j}M_{k,\ell}$$

It can also be shown that

$$\mathsf{E}[n^2 S_i^n S_j^n S_k^n S_\ell^n] = \frac{F_{i,j,k,l}}{n} + \Sigma_{i,j}\Sigma_{k,\ell} + \Sigma_{j,k}\Sigma_{i,\ell} + \Sigma_{i,k}\Sigma_{j,\ell} \text{ for all } n$$

where $F_{i,j,k,l} = \mathsf{E}[X_i^1 X_j^1 X_k^1 X_\ell^1]$. Moreover, $\mathsf{E}[(n^2 S_i^n S_j^n S_k^n S_\ell^n)^2]$ is finite for each n and converges to a finite quantity as $n \to \infty$ since the moments of X^i are finite. Thus we can again apply Lemma A.1 to see that

$$\mathsf{E}[n^2 S_i^n \nabla^2 h(\widetilde{S}^n)_{i,j} S_j^n S_k^n \nabla^2 h(\widetilde{S}^n)_{k,\ell} S_\ell^n \mathbb{I}_{\{S^n \in K\}}] \xrightarrow[n\to\infty]{} (\Sigma_{i,j}\Sigma_{k,\ell} + \Sigma_{j,k}\Sigma_{i,\ell} + \Sigma_{i,k}\Sigma_{j,\ell}) M_{i,j} M_{k,\ell} \quad (68)$$

Putting together terms and using (66) we obtain:

$$\mathsf{E}[(nh(S^n))^2 \mathbb{I}_{\{S^n \in K\}}] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi M\Xi) + \tfrac{1}{4}(\text{trace}(M\Xi))^2$$

Now similar to (65) we have:

$$|\mathsf{E}[(nh(S^n))^2 \mathbb{I}_{\{S^n \notin K\}}]| \leq n^2 \overline{h}^2 \mathsf{P}\{S^n \notin K\} \xrightarrow[n\to\infty]{} 0 \quad (69)$$

Consequently

$$\mathsf{E}[(nh(S^n))^2] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi M\Xi) + \tfrac{1}{4}(\text{trace}(M\Xi))^2$$

which gives (ii).

■

B. Proof of Lemma III.7

We know from (2) that Γ^n can be written as an empirical average of i.i.d. vectors. Hence, it satisfies the central limit theorem which says that,

$$n^{\frac{1}{2}}(\Gamma^n - \mu) \xrightarrow[n\to\infty]{d.} W \quad (70)$$

where the distribution of W is defined below (35).

Considering a second-order Taylor's expansion and using the condition on the directional derivative, we have,

$$n(h(\Gamma^n) - h(\mu)) = \tfrac{1}{2} n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu))$$

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

$$\nabla^2 h(\tilde{\Gamma}^n) \xrightarrow[n\to\infty]{a.s.} \nabla^2 h(\mu). \quad (71)$$

By applying the vector-version of Slutsky's theorem [17], together with (70) and (71), we conclude

$$n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu)) \xrightarrow[n\to\infty]{d.} \frac{1}{2} W^T \nabla^2 h(\mu) W,$$

thus establishing the lemma.

References

[1] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, "Statistical SVMs for robust detection, supervised learning, and universal classification," in *Proceedings of the Information Theory Workshop on Networking and Information Theory, Volos, Greece.*, 2009.

[2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.

[3] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, 1992.

[4] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2215–2229, 2002.

[5] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep 1996.

[6] E. Abbe, M. Medard, S. Meyn, and L. Zheng, "Finding the best mismatched detector for channel coding and hypothesis testing," *Information Theory and Applications Workshop, 2007*, pp. 284–288, 29 2007-Feb. 2 2007.

[7] 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.

[8] C. Pandit, S. Meyn, and V. Veeravalli, "Asymptotic robust Neyman-Pearson hypothesis testing based on moment classes," *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, pp. 220–, June-2 July 2004.

[9] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stoch. Proc. Applns.*, vol. 116, no. 5, pp. 724–756, May 2006.

[10] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. I. II," *Comm. Pure Appl. Math.*, vol. 28, pp. 1–47; ibid. **28** (1975), 279–301, 1975.

[11] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 285–290, 1991.

[12] ——, "Correction to: "On universal hypotheses testing via large deviations"," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, part 1, p. 698, 1991.

[13] B. Clarke and A. R. Barron, "Information theoretic asymptotics of bayes' methods," Univ. of Illinois, Department of Statistics, Tech. Rep. 26, July 1989.

[14] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statistics*, vol. 9, pp. 60–62, 1938.

[15] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[16] S. Meyn, A. Surana, Y. Lin, and S. Narayanan, "Anomaly detection using projective Markov models in a distributed sensor network," December 16-18 2009, Accepted for inclusion in the 48th IEEE Conference on Decision and Control.

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