Second-Order Asymptotics of Divergence Tests

Second-Order Asymptotics of Divergence Tests

K. V. Harsha,* Jithin Ravi,† and Tobias Koch‡

*Gandhi Institute of Technology and Management, Hyderabad, India
†Indian Institute of Technology Kharagpur, India
‡Universidad Carlos III de Madrid, Leganés, Spain and Gregorio Marañón Health Research Institute, Madrid, Spain.
Emails: hkalluma@gitam.edu, jithin@ece.iitkgp.ac.in, koch@tsc.uc3m.es

***Abstract*—Consider a binary statistical hypothesis testing problem, where n independent and identically distributed random variables Z^n are either distributed according to the null hypothesis P or the alternate hypothesis Q, and only P is known. A well-known test that is suitable for this case is the so-called Hoeffding test, which accepts P if the Kullback-Leibler (KL) divergence between the empirical distribution of Z^n and P is below some threshold. In this work, we characterize the first and second-order terms of the type-II error probability for a fixed type-I error probability for the Hoeffding test as well as for divergence tests, where the KL divergence is replaced by a general divergence. We demonstrate that, irrespective of the divergence, divergence tests achieve the first-order term of the Neyman-Pearson test, which is the optimal test when both P and Q are known. In contrast, the second-order term of divergence tests is strictly worse than that of the Neyman-Pearson test. We further demonstrate that divergence tests with an invariant divergence achieve the same second-order term as the Hoeffding test, but divergence tests with a non-invariant divergence may outperform the Hoeffding test for some alternate hypotheses Q.**

I. Introduction

Consider a binary hypothesis testing problem that decides whether a sequence of independent and identically distributed (i.i.d.) random variables Z^n is either generated from distribution P or from distribution Q. Assume that both distributions are discrete and the hypothesis test has access to P but not to Q. A suitable test for this case is the well-known *Hoeffding test* [1], which accepts P if $D_{\mathrm{KL}}(T_{Z^n}\|P) < c$, for some $c > 0$, and otherwise accepts Q. Here, T_{Z^n} is the type (the empirical distribution) of Z^n and $D_{\mathrm{KL}}(P\|Q)$ is the Kullback-Leibler (KL) divergence between P and Q [2]. In this paper, we analyze the second-order performance of the Hoeffding test as well as of Hoeffding-like tests, referred to as *divergence tests*, where the KL divergence is replaced by other divergences (see Section II for a rigorous definition).

Part of this work was done while K. V. Harsha and J. Ravi were with the Signal Theory and Communications Department, Universidad Carlos III de Madrid, Leganés, Spain. K. V. Harsha, J. Ravi, and T. Koch have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 714161). J. Ravi has further received funding from the Science and Engineering Research Board (SERB) under Start-up Research Grant (SRG) and also from IIT Kharagpur under Faculty Start-up Research Grant (FSRG). T. Koch has further received funding from the Spanish Ministerio de Ciencia e Innovación under Grant PID2020-116683GB-C21 / AEI / 10.13039/501100011033.

We focus on the asymptotic behaviour of the type-II error β_n (the probability of declaring hypothesis P under hypothesis Q) for a fixed type-I error α_n (the probability of declaring hypothesis Q under hypothesis P). When both P and Q are known, the optimal test is the likelihood ratio test, also known as the Neyman-Pearson test. For this test, the smallest type-II error β_n for which $\alpha_n \le \epsilon$ satisfies [3, Prop. 2.3]

$$-\ln \beta_n = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV(P\|Q)}\mathsf{Q}^{-1}(\epsilon) + o(\sqrt{n}) \quad (1)$$

as $n \to \infty$, where

$$V(P\|Q) \triangleq \sum_{i=1}^{k} P_i \left[\left(\ln \frac{P_i}{Q_i} - D_{\mathrm{KL}}(P\|Q)\right)^2\right] \quad (2)$$

denotes the divergence variance; $\mathsf{Q}^{-1}(\cdot)$ denotes the inverse of the tail probability of the standard Normal distribution; P_i and Q_i denote the i-th components of P and Q; and k denotes their dimension. Here and throughout this paper, we write $a_n = o(b_n)$ for two sequences $\{a_n\}$ and $\{b_n\}$ of real numbers if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. We write $a_n = O(b_n)$ if $\overline{\lim}_{n\to\infty} |\frac{a_n}{b_n}| < \infty$. By inspecting the expansion of $-\ln \beta_n$ in (1), one can define the first-order term β' and the second-order term β'' of any hypothesis test $\mathbb{T}$ as

$$\beta' \triangleq \lim_{n\to\infty} \frac{-\ln \beta_n(\mathbb{T})}{n} \quad (3)$$

and

$$\beta'' \triangleq \lim_{n\to\infty} \frac{-\ln \beta_n(\mathbb{T}) - n\beta'}{\sqrt{n}} \quad (4)$$

if the limits exist. The first-order term β' is sometimes referred to as the *error exponent*. For the Neyman-Pearson test, we have $\beta' = D_{\mathrm{KL}}(P\|Q)$ and $\beta'' = -\sqrt{V(P\|Q)}\mathsf{Q}^{-1}(\epsilon)$.

It was shown in [1] that the first-order term β' of the Hoeffding test is also $D_{\mathrm{KL}}(P\|Q)$. In other words, the Hoeffding test is first-order optimal. Recently, we have demonstrated [4] that the second-order term of the Hoeffding test is $\beta'' = -\sqrt{V(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)}$, where $\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\cdot)$ denotes the inverse of the tail probability of the chi-square distribution with $k-1$ degrees of freedom. Since $\sqrt{\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} > \mathsf{Q}^{-1}(\epsilon)$, it follows that the second-order performance of the Hoeffding test is worse than that of the Neyman-Pearson test.

In this paper, we analyze the second-order performance of the divergence test $\mathbb{T}^D$, which accepts P if $D(T_{Z^n}\|P) < c$, for some $c > 0$, and otherwise accepts Q. The divergence D of the divergence test $\mathbb{T}^D$ is arbitrary, so $\mathbb{T}^D$ includes the Hoeffding test as a special case when $D = D_{\text{KL}}$. We demonstrate that the divergence test $\mathbb{T}^D$ achieves the same first-order term β' as the Neyman-Pearson test, irrespective of the divergence D. Hence, $\mathbb{T}^D$ is first-order optimal for every divergence D. We further demonstrate that, for the class of *invariant divergences* [5], which includes the Rényi divergence and the f-divergence (and, hence, also the KL divergence), the divergence test $\mathbb{T}^D$ achieves the same second-order term β'' as the Hoeffding test. In contrast, we show that a divergence test $\mathbb{T}^D$ with a non-invariant divergence may achieve a second-order term β'' that is strictly better than that of the Hoeffding test for some Q and ϵ.

A. Related Work

The considered hypothesis testing problem falls under the category of *composite hypothesis testing* [6]. Indeed, in composite hypothesis testing, the test has no access to the distribution P of the null hypothesis and the distribution Q of the alternate hypothesis, but it has the knowledge that P and Q belong to the sets of distributions $\mathcal{P}$ and $\mathcal{Q}$, respectively. Our setting corresponds to the case where $\mathcal{P} = \{P\}$ and $\mathcal{Q} = \mathcal{P}^c$ (where we use the notation $\mathcal{A}^c$ to denote the complement of a set $\mathcal{A}$).

The Hoeffding test is a particular instance of the *generalized likelihood-ratio test (GLRT)* [7], which is arguably the most common test used in composite hypothesis testing. A useful benchmark for the Hoeffding test is the Neyman-Pearson test, which is the optimal test when both P and Q are known. As mentioned before, the Hoeffding test achieves the same first-order term β' as the Neyman-Pearson test, both in *Stein's regime*, where the type-I error satisfies $\alpha_n \leq \epsilon$, as well as in the *doubly-exponential regime*, where $\alpha_n \leq e^{-n\gamma}$, $\gamma > 0$; see, e.g., [1], [8]–[11]. Thus, the first-order term of the Neyman-Pearson test can be achieved without having access to the distribution Q of the alternate hypothesis. However, not having access to Q negatively affects higher-order terms. For example, for a given threshold γ, the type-I error of the Hoeffding test satisfies [11, Eq. (10)]

$$\alpha_n = n^{\frac{k-3}{2}} e^{-n\gamma}(c' + o(1)) \tag{5}$$

whereas for the corresponding Neyman-Pearson test [11, Eq. (9)]

$$\alpha_n = n^{-\frac{1}{2}} e^{-n\gamma}(c + o(1)). \tag{6}$$

Here, c and c' are constants that only depend on P, Q, and γ. Moreover, it was demonstrated in [9] that the variance of the normalized Hoeffding test statistic $nD_{\text{KL}}(T_{Z^n}\|P)$ converges to $\frac{1}{2}(k-1)$ as $n \to \infty$. Both results suggest that, for moderate n, the Hoeffding test scales unfavorably with the cardinality of P and Q, which motivated the authors of [9] to propose their *test via mismatched divergence*. The same observation can be made for Stein's regime. Indeed, as mentioned before, the second-order term of the Hoeffding test is [4]

$$\beta'' = -\sqrt{V(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} \tag{7}$$

whereas the second-order term of the Neyman-Pearson test is [3, Prop. 2.3]

$$\beta'' = -\sqrt{V(P\|Q)}\mathsf{Q}^{-1}(\epsilon). \tag{8}$$

Since $\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)$ is monotonically increasing in k, this again suggests an unfavorable scaling with the cardinality of P and Q.

Our setting where $\mathcal{P} = \{P\}$ and $\mathcal{Q} = \mathcal{P}^c$ was also studied by Watanabe [12], who proposed a test that is second-order optimal in some sense. The related case where only training sequences are available for both P and Q was considered in [13]. The test proposed in [13] was later shown to be second-order optimal [14].

II. Divergence and Divergence Test

A. Divergence

Let us consider a random variable Z that takes value in a discrete set $\mathcal{Z} = \{a_1, \cdots, a_k\}$ with cardinality $|\mathcal{Z}| = k \geq 2$. Let $\overline{\mathcal{P}}(\mathcal{Z})$ denote the set of probability distributions on $\mathcal{Z}$, and let $\mathcal{P}(\mathcal{Z})$ denote the set of probability distributions with strictly positive probabilities. Any probability distribution $R \in \mathcal{P}(\mathcal{Z})$ can be written as a length-k vector $R = (R_1, \cdots, R_k)^\mathsf{T}$, where $R_i \triangleq \Pr\{Z = a_i\}, i = 1, \cdots, k$. Note that this R can also be represented by its first $(k-1)$ components, denoted by the vector $\mathbf{R} = (R_1, \cdots, R_{k-1})^\mathsf{T}$, which takes value in the coordinate space

$$\Xi \triangleq \left\{ (R_1, \cdots, R_{k-1})^\mathsf{T} : R_i > 0, \sum_{i=1}^{k-1} R_i < 1 \right\}. \tag{9}$$

Given any two probability distributions $S, R \in \mathcal{P}(\mathcal{Z})$, one can define a non-negative function $D(S\|R)$, called a *divergence*, which represents a measure of discrepancy between them. A divergence is not necessarily symmetric in its arguments and also need not satisfy the triangle inequality; see [15], [16] for more details. More precisely, a divergence is defined as follows [15]:

Definition 1: Consider two distributions S and R in $\mathcal{P}(\mathcal{Z})$. A *divergence* $D \colon \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty)$ between S and R, denoted by $D(S\|R)$, is a smooth function[1] of $\mathbf{S} \in \Xi$ and $\mathbf{R} \in \Xi$ (we may write $D(S\|R) = D(\mathbf{S}\|\mathbf{R})$) satisfying the following conditions:

1) $D(S\|R) \geq 0$ for every $S, R \in \mathcal{P}(\mathcal{Z})$.
2) $D(S\|R) = 0$ if, and only if, $S = R$.
3) When $\mathbf{S} = \mathbf{R} + \boldsymbol{\varepsilon}$ for some $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_{k-1})^\mathsf{T}$, the Taylor expansion of D satisfies

$$D(\mathbf{R} + \boldsymbol{\varepsilon}\|\mathbf{R}) = \frac{1}{2}\sum_{i,j=1}^{k-1} g_{ij}(\mathbf{R})\varepsilon_i\varepsilon_j + O(\|\boldsymbol{\varepsilon}\|_2^3) \tag{10}$$

as $\|\boldsymbol{\varepsilon}\|_2 \to 0$ for some $(k-1) \times (k-1)$-dimensional positive-definite matrix $G(\mathbf{R}) = [g_{ij}(\mathbf{R})]$ that depends on $\mathbf{R}$. In (10), $\|\boldsymbol{\varepsilon}\|_2$ is the Euclidean norm of $\boldsymbol{\varepsilon}$.

[1] We shall say that a function is *smooth* if it has partial derivatives of all orders.

4) Let $R \in \mathcal{P}(\mathcal{Z})$, and let $\{S_n\}$ be a sequence of distributions in $\mathcal{P}(\mathcal{Z})$ that converges to a distribution S on the boundary of $\mathcal{P}(\mathcal{Z})$. Then,

$$\varliminf_{n\to\infty} D(S_n \| R) > 0. \tag{11}$$

Remark 1: We follow the definition of divergence from the information geometry literature. In particular, according to [15, Def. 1.1], a divergence must satisfy the first three conditions in Definition 1. Often, the behavior of divergence on the boundary of $\mathcal{P}(\mathcal{Z})$ is not specified. In Definition 1, we add the fourth condition to treat the case of sequences of distributions $\{S_n\}$ that lie in $\mathcal{P}(\mathcal{Z})$ but converge to a distribution on the boundary of $\mathcal{P}(\mathcal{Z})$. Note that condition 4) is consistent with conditions 1) and 2).

Given a divergence D and $R \in \mathcal{P}(\mathcal{Z})$, consider the function $D(\cdot\|R)\colon \mathbb{R}^{k-1} \to \mathbb{R}$. By computing the partial derivatives of $D(S\|R)$ with respect to the first variable $\mathbf{S} = (S_1, \cdots, S_{k-1})^{\mathsf{T}}$, it follows from the third condition in Definition 1 that

$$D(S\|R) = (\mathbf{S}-\mathbf{R})^T \boldsymbol{A}_{D,\mathbf{R}}(\mathbf{S}-\mathbf{R}) + O(\|\mathbf{S}-\mathbf{R}\|_2^3) \tag{12}$$

as $\|\mathbf{S}-\mathbf{R}\|_2 \to 0$, where $\boldsymbol{A}_{D,\mathbf{R}}$ is the matrix associated with the divergence D at $\mathbf{R}$, which has components

$$a_{ij}(\mathbf{R}) \triangleq \frac{1}{2}\left.\frac{\partial^2}{\partial S_i \partial S_j} D(S\|R)\right|_{S=R}, \quad i,j = 1,\cdots,k-1. \tag{13}$$

Based on $\boldsymbol{A}_{D,\mathbf{R}}$, we can introduce the notion of an *invariant divergence*.

Definition 2: Let D be a divergence, and let $R \in \mathcal{P}(\mathcal{Z})$. Then, D is said to be an *invariant divergence* on $\mathcal{P}(\mathcal{Z})$ if the matrix associated with the divergence D at $\mathbf{R}$ is of the form $\boldsymbol{A}_{D,\mathbf{R}} = \eta \boldsymbol{\Sigma}_{\mathbf{R}}$ for a constant $\eta > 0$ (possibly depending on $\mathbf{R}$) and a matrix $\boldsymbol{\Sigma}_{\mathbf{R}}$ with components

$$\boldsymbol{\Sigma}_{ij}(\mathbf{R}) = \begin{cases} \frac{1}{R_i} + \frac{1}{R_k}, & i = j \\ \frac{1}{R_k}, & i \neq j. \end{cases} \tag{14}$$

The notion of an invariant divergence is adapted from the notion of invariance of geometric structures in information geometry; see [15], [17] for more details. The matrix $\boldsymbol{\Sigma}_{\mathbf{R}}$ represents the unique invariant Riemannian metric in $\mathcal{P}(\mathcal{Z})$ with respect to the coordinate system Ξ; see [18, Eq. (47)], [5] for more details. However, in the information geometry literature, the constant η is often required to be independent of $\mathbf{R}$. Well-known divergences, such as the KL divergence, the f-divergence, and the Rényi divergence, are invariant [19]. For an invariant divergence, (12) becomes

$$D(S\|R) = \eta(\mathbf{S}-\mathbf{R})^T \boldsymbol{\Sigma}_{\mathbf{R}}(\mathbf{S}-\mathbf{R}) + O(\|\mathbf{S}-\mathbf{R}\|_2^3) \tag{15}$$

as $\|\mathbf{S}-\mathbf{R}\|_2 \to 0$, where η is a positive constant.

There are many divergences that do not satisfy (15). An example is the *squared Mahalanobis distance*, which is of the form

$$D_{\text{SM}}(S\|R) = (\mathbf{S}-\mathbf{R})^{\mathsf{T}} \boldsymbol{W}_{\mathbf{R}}(\mathbf{S}-\mathbf{R}) \tag{16}$$

for some positive-definite matrix $\boldsymbol{W}_{\mathbf{R}}$. This divergence is non-invariant if $\boldsymbol{W}_{\mathbf{R}}$ is not a constant multiple of $\boldsymbol{\Sigma}_{\mathbf{R}}$.

For a detailed list of divergences and their properties, we refer to [19, Ch. 2].

B. General Setting and Divergence Test

We consider a binary hypothesis testing problem with null hypothesis H_0 and alternate hypothesis H_1. We assume that, under hypothesis H_0, the length-n sequence Z^n of observations is i.i.d. according to $P \in \mathcal{P}(\mathcal{Z})$; under hypothesis H_1, the sequence of observations Z^n is i.i.d. according to Q, where $Q \in \mathcal{P}(\mathcal{Z}) \setminus \{P\}$.

We next define the divergence test. To this end, we first introduce the *type distribution*, which for every sequence z^n is defined as

$$T_{z^n}(a_i) \triangleq \frac{1}{n}\sum_{\ell=1}^{n} \mathbf{1}\{z_\ell = a_i\}, \quad i = 1,\ldots,k \tag{17}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

For a divergence D and a threshold $r > 0$, a *divergence test* $\mathbb{T}_n^D(r)$ for testing H_0 against the alternative H_1 is defined as follows:

Observe Z^n: if $D(T_{Z^n}\|P) < r$, then H_0 is accepted; else H_1 is accepted.

When the divergence D is the Kullback-Leibler divergence D_{KL}, the divergence test becomes the Hoeffding test, proposed by Hoeffding in [1].

For $r > 0$, define the acceptance region for H_0 as

$$\mathcal{A}_n^D(r) \triangleq \{z^n \colon D(T_{z^n}\|P) < r\}. \tag{18}$$

Then, the type-I and the type-II errors are given by

$$\alpha_n\big(\mathbb{T}_n^D(r)\big) \triangleq P^n\big(\mathcal{A}_n^D(r)^c\big) \tag{19}$$

$$\beta_n\big(\mathbb{T}_n^D(r)\big) \triangleq Q^n\big(\mathcal{A}_n^D(r)\big). \tag{20}$$

Our goal is to analyze the asymptotic behavior of the type-II error β_n when the type-I error satisfies $\alpha_n \le \epsilon$, $0 < \epsilon < 1$.

III. Main Results

The asymptotic behavior of the divergence test depends on the asymptotic behavior of the random variable $nD(T_{Z^n}\|P)$ in the limit as $n \to \infty$. For certain divergences, the limiting distribution of $nD(T_{Z^n}\|P)$ has been analyzed in the literature. For example, when D is the KL divergence, a well-known result by Wilks [20] states that $2nD_{\text{KL}}(T_{Z^n}\|P)$ converges in distribution to a chi-square random variable with $k-1$ degrees of freedom. This result generalizes to the α-divergence [21, Th. 3.1], [22, Th. 3]. In Lemma 1, we show that, for a general divergence D, $nD(T_{Z^n}\|P)$ converges in distribution to a *generalized chi-square random variable*, defined as follows:

Definition 3: The *generalized chi-square distribution* is the distribution of the random variable

$$\xi = \sum_{i=1}^{m} w_i \Upsilon_i \tag{21}$$

where w_i, $i = 1, \cdots, m$ are deterministic weight parameters and $\Upsilon_i, i = 1, \cdots, m$ are independent chi-square random variables with degree of freedom 1. We shall denote the generalized chi-square distribution with weight vector $\mathbf{w} = (w_1, \cdots, w_m)^\mathsf{T}$ and degrees of freedom m by $\chi^2_{\boldsymbol{w},m}$. If $w_i = 1$ for all i, then the generalized chi-square distribution becomes the chi-square distribution χ^2_m with degrees of freedom m.

Lemma 1: Let Z^n be a sequence of i.i.d. random variables distributed according to the distribution P of the null hypothesis, and let D be a divergence. Further let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_{k-1})^\mathsf{T}$ be a vector that contains the eigenvalues of the matrix $\boldsymbol{\Sigma}_\mathbf{P}^{-1/2} \boldsymbol{A}_{D,\mathbf{P}} \boldsymbol{\Sigma}_\mathbf{P}^{-1/2}$, where $\boldsymbol{A}_{D,\mathbf{P}}$ is the matrix associated with the divergence D at $\mathbf{P}$ and the matrix $\boldsymbol{\Sigma}_\mathbf{P}$ is defined in (14). Then, the tail probability of the random variable $nD(T_{Z^n}\|P)$ satisfies

$$P^n(nD(T_{Z^n}\|P) \geq c) = \mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(c) + O(\delta_n), \quad c \geq 0 \quad (22)$$

for some positive sequence $\{\delta_n\}$ that is independent of c and satisfies $\lim_{n\to\infty} \delta_n = 0$. Here, $\mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}(c) \triangleq \Pr(\xi \geq c)$ is the tail probability of the generalized chi-square random variable ξ with weight vector $\boldsymbol{\lambda}$ and degrees of freedom $k-1$.

Proof: Omitted due to space limitations. ■

We are now ready to present the main result of this paper:

Theorem 1: Let D be a divergence as defined in Definition 1, and let $0 < \epsilon < 1$. Further let $P, Q \in \mathcal{P}(\mathcal{Z})$ and $P \neq Q$. Recall that the cardinality of $\mathcal{Z}$ is $k \geq 2$. Then, for all sequences of thresholds $\{r_n\}$ satisfying

$$\alpha_n(\mathbb{T}_n^D(r_n)) \leq \epsilon \quad (23)$$

the divergence test $\mathbb{T}_n^D$ introduced in Section II-B satisfies

$$\begin{aligned} &\sup_{r_n:\, \alpha_n(\mathbb{T}_n^D(r_n)) \leq \epsilon} -\ln \beta_n\big(\mathbb{T}_n^D(r_n)\big) \\ &\quad = nD_{\mathrm{KL}}(P\|Q) - \sqrt{n}\sqrt{\mathbf{c}^\mathsf{T} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}} \sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} \\ &\quad\quad + O(\max\{\delta_n\sqrt{n}, \ln n\}). \end{aligned} \quad (24)$$

Here, $\boldsymbol{A}_{D,\mathbf{P}}$ is the matrix associated with the divergence D at $\mathbf{P}$; the sequence $\{\delta_n\}$ was defined in (22); $\mathbf{c} = (c_1, \cdots, c_{k-1})^\mathsf{T}$ is a vector with components

$$c_i \triangleq \ln\left(\frac{P_i}{Q_i}\right) - \ln\left(\frac{P_k}{Q_k}\right), \quad i = 1, \cdots, k-1; \quad (25)$$

and $\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}$ is the inverse of the tail probability $\mathsf{Q}_{\chi^2_{\boldsymbol{\lambda},k-1}}$ introduced in Lemma 1.

Proof: Omitted due to space limitations. ■

Remark 2: Since the sequence $\{\delta_n\}$ tends to zero as $n \to \infty$, we have that $O(\max\{\delta_n\sqrt{n}, \ln n\}) = o(\sqrt{n})$.

Corollary 1: For the class of invariant divergences, (24) in Theorem 1 becomes

$$\begin{aligned} &\sup_{r_n:\, \alpha_n(\mathbb{T}_n^D(r_n)) \leq \epsilon} -\ln \beta_n\big(\mathbb{T}_n^D(r_n)\big) \\ &\quad = nD_{\mathrm{KL}}(P\|Q) - \sqrt{nV(P\|Q)\mathsf{Q}^{-1}_{\chi^2_{k-1}}(\epsilon)} + o(\sqrt{n}). \end{aligned} \quad (26)$$

Since the KL divergence belongs to the class of invariant divergences, it follows that (26) also characterizes the second-order performance of the Hoeffding test.

We observe from Theorem 1 that the divergence test $\mathbb{T}_n^D$ achieves the same first-order term β' as the Neyman-Pearson test, irrespective of D. In contrast, it can be shown that

$$-\sqrt{\mathbf{c}^\mathsf{T} \boldsymbol{A}_{D,\mathbf{P}}^{-1} \mathbf{c}} \sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},k-1}}(\epsilon)} < -\sqrt{V(P\|Q)}\mathsf{Q}_\mathcal{N}^{-1}(\epsilon). \quad (27)$$

Thus, the second-order term β'' of the divergence test $\mathbb{T}^D$ is strictly smaller than the second-order term of the Neyman-Pearson test.

In the next section, we show that there are divergences for which the divergence test outperforms the Hoeffding test for certain distributions Q of the alternate hypothesis.

IV. Second-Order Performance Comparison

In order to contrast the performances of different divergence tests, we numerically evaluate the second-order performances of $\mathbb{T}^{D_{\mathrm{KL}}}$ and $\mathbb{T}^{D_{\mathrm{SM}}}$, where D_{KL} is the KL divergence and D_{SM} is the squared Mahalanobis distance. Recall that the KL divergence is an invariant divergence. For the squared Mahalanobis distance, we shall consider (16) with $\boldsymbol{W}_\mathbf{P}$ having components

$$\boldsymbol{W}_{ij}(\mathbf{P}) = \begin{cases} \frac{1}{2P_i^2} + \frac{1}{2P_k^2}, & i = j \\ \frac{1}{2P_k^2}, & i \neq j \end{cases} \quad (28)$$

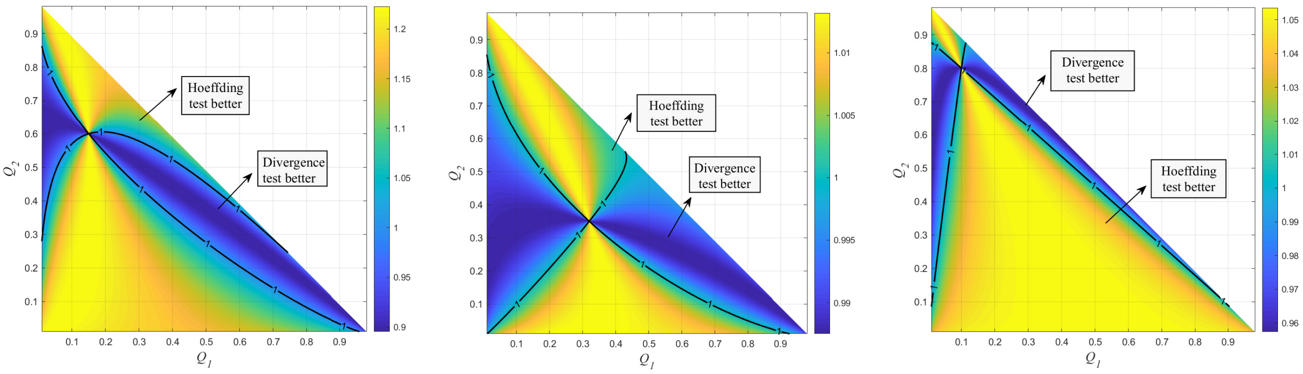
which is a non-invariant divergence. To better visualize the second-order performances, we focus on distributions with dimension $k = 3$ and represent them by the two-dimensional vectors $\mathbf{P} = (P_1, P_2)^\mathsf{T}$ and $\mathbf{Q} = (Q_1, Q_2)^\mathsf{T}$ in the coordinate space Ξ.

Since the first-order term β' of the divergence test $\mathbb{T}^D$ is not affected by the choice of D, we shall compare the second-order performances of $\mathbb{T}^{D_{\mathrm{KL}}}$ and $\mathbb{T}^{D_{\mathrm{SM}}}$ by considering the ratio of the second-order terms β'' as a function of P, Q, and ϵ:

$$\rho(P, Q, \epsilon) \triangleq \frac{\sqrt{\mathbf{c}^\mathsf{T} (\boldsymbol{W}_\mathbf{P})^{-1} \mathbf{c}} \sqrt{\mathsf{Q}^{-1}_{\chi^2_{\boldsymbol{\lambda},2}}(\epsilon)}}{\sqrt{V(P\|Q)} \sqrt{\mathsf{Q}^{-1}_{\chi^2_2}(\epsilon)}}. \quad (29)$$

If $\rho(P, Q, \epsilon) > 1$, then the second-order term of the divergence test is strictly smaller than the second-order term of the Hoeffding test, hence the Hoeffding test has a better second-order performance. In contrast, if $\rho(P, Q, \epsilon) < 1$, then the divergence test has a better second-order performance.

In Fig. 1, we plot the contour lines of the ratio $\rho(P, Q, \epsilon)$ as a function of $\mathbf{Q} \in \Xi$ for $\epsilon = 0.02$ and the three different null hypotheses $\mathbf{P} = (0.15, 0.6)$, $\mathbf{P} = (0.32, 0.35)$, and $\mathbf{P} = (0.1, 0.8)$. In the figure, the coordinate space Ξ is divided into two regions: one region is labeled as "Hoeffding test better" and includes the points $\mathbf{Q} \in \Xi$ for which $\rho(P, Q, \epsilon) > 1$; the other region is labeled as "Divergence test better" and includes the points $\mathbf{Q} \in \Xi$ for which $\rho(P, Q, \epsilon) < 1$. The solid contour line drawn in all three sub-figures shows all the points $\mathbf{Q} \in \Xi$ for which the Hoeffding test and the divergence



(a) $P = (0.15, 0.6, 0.25)$ (b) $P = (0.32, 0.35, 0.33)$ (c) $P = (0.1, 0.8, 0.1)$

Fig. 1: Second-order performance comparison between the Hoeffding test $\mathbb{T}^{D_{\mathrm{KL}}}$ and the divergence test $\mathbb{T}^{D_{\mathrm{SM}}}$ for the three different null hypotheses $P = (0.15, 0.6, 0.25)$, $P = (0.32, 0.35, 0.33)$, and $P = (0.1, 0.8, 0.1)$ and $\epsilon = 0.02$.

test have the same second-order performance. For each subfigure, the color bar on the right indicates the values of the ratio $\rho(P, Q, \epsilon)$.

Observe that there are distributions Q of the alternate hypothesis for which the Hoeffding test has a better second-order performance than the divergence test, and there are distributions Q for which the opposite is true. The set of distributions Q for which one test outperforms the other typically depends on the distribution P of the null hypothesis and on ϵ. Potentially, this behavior could be exploited in a composite hypothesis testing problem by tailoring the divergence D of the divergence test $\mathbb{T}^{D}$ to the set $\mathcal{Q}$ of possible alternate distributions.

Acknowledgment

The authors would like to thank Shun Watanabe for his helpful comments during the initial phase of this work and for pointing out relevant literature.

References

[1] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” *The Annals of Mathematical Statistics*, pp. 369–401, Apr. 1965.
[2] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2006.
[3] V. Y. Tan, “Asymptotic estimates in information theory with non-vanishing error probabilities,” in *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.
[4] K. V. Harsha, J. Ravi, and T. Koch, “Second-order asymptotics of Hoeffding-like hypothesis tests,” in *Proc. 2022 IEEE Information Theory Workshop (ITW)*, Mumbai, India, Nov. 2022, pp. 654–659.
[5] N. N. Cencov, “Statistical decision rules and optimal inference," in *Translations of Mathematical Monographs*, Vol. 53, 1982.
[6] M. Feder and N. Merhav, “Universal composite hypothesis testing: A competitive minimax approach,” *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1504–1517, Jun. 2002.
[7] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* John Wiley & Sons, 1968.
[8] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.
[9] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, “Universal and composite hypothesis testing via mismatched divergence,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1587–1603, Mar. 2011.
[10] P. Boroumand and A. Guillén i Fàbregas, “Mismatched binary hypothesis testing: Error exponent sensitivity," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6738–6761, Oct. 2022.
[11] ——, “Composite Neyman-Pearson hypothesis testing with a known hypothesis," in *Proc. 2022 IEEE Information Theory Workshop (ITW)*, Mumbai, India, Nov. 2022, pp. 131–136.
[12] S. Watanabe, “Second-order optimal test in composite hypothesis testing," in *Proc. 2018 International Symposium on Information Theory and Its Applications (ISITA)*, Singapore, Oct. 2018, pp. 722–726.
[13] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.
[14] Y. Li and V. Y. F. Tan, “Second-order asymptotics of sequential hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 7222–7230, Nov. 2020.
[15] S.-I. Amari, *Information geometry and its applications*. Springer, 2016, vol. 194.
[16] S. Eguchi, “A differential geometric approach to statistical inference on the basis of contrast functionals,” *Hiroshima Mathematical Journal*, vol. 15, no. 2, pp. 341–391, 1985.
[17] L. L. Campbell, “An extended Cencov characterization of the information metric,” *Proceedings of the American Mathematical Society*, vol. 98, no. 1, pp. 135–141, Sep. 1986.
[18] S.-I. Amari and A. Cichocki, “Information geometry of divergence functions,” *Bulletin of the Polish Academy of Sciences. Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.
[19] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.
[20] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, Mar. 1938.
[21] T. R. Read, “Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics,” *Annals of the Institute of Statistical Mathematics*, vol. 36, pp. 59–69, Dec. 1984.
[22] J. K. Yarnold, “Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set,” *The Annals of Mathematical Statistics*, pp. 1566–1580, Oct. 1972.